Examining the Validity of a State Policy-Directed Framework for Evaluating Teacher Instructional Quality: Informing Policy, Impacting Practice

by

Edward F Sloat

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Education

Approved April 2015 by the
Graduate Supervisory Committee:

Keith Wetzel, Co-Chair
Audrey Amrein-Beardsley, Co-Chair
Ann Ewbank
Lori Shough

ARIZONA STATE UNIVERSITY

May 2015

ABSTRACT

This study examines validity evidence of a state policy-directed teacher evaluation system implemented in Arizona during school year 2012-2013. The purpose was to evaluate the warrant for making high stakes, consequential judgments of teacher competence based on value-added (VAM) estimates of instructional impact and observations of professional practice (PP). The research also explores educator influence (voice) in evaluation design and the role information brokers have in local decision making. Findings are situated in an evidentiary and policy context at both the LEA and state policy levels.

The study employs a single-phase, concurrent, mixed-methods research design triangulating multiple sources of qualitative and quantitative evidence onto a single (unified) validation construct: Teacher Instructional Quality. It focuses on assessing the characteristics of metrics used to construct quantitative ratings of instructional competence and the alignment of stakeholder perspectives to facets implicit in the evaluation framework. Validity examinations include assembly of criterion, content, reliability, consequential and construct articulation evidences. Perceptual perspectives were obtained from teachers, principals, district leadership, and state policy decision makers. Data for this study came from a large suburban public school district in metropolitan Phoenix, Arizona.

Study findings suggest that the evaluation framework is insufficient for supporting high stakes, consequential inferences of teacher instructional quality. This is based, in part on the following: (1) Weak associations between VAM and PP metrics; (2) Unstable VAM measures across time and between tested content areas; (3) Less than

adequate scale reliabilities; (4) Lack of coherence between theorized and empirical PP factor structures; (5) Omission/underrepresentation of important instructional attributes/effects; (6) Stakeholder concerns over rater consistency, bias, and the inability of test scores to adequately represent instructional competence; (7) Negative sentiments regarding the system's ability to improve instructional competence and/or student learning; (8) Concerns regarding unintended consequences including increased stress, lower morale, harm to professional identity, and restricted learning opportunities; and (9) The general lack of empowerment and educator exclusion from the decision making process. Study findings also highlight the value of information brokers in policy decision making and the importance of having access to unbiased empirical information during the design and implementation phases of important change initiatives.

DEDICATION

To my wife, Adrianne, who believed in me, encouraged me, and sacrificed so much so that I might chase a dream and make it a reality. Thank you for this gift. I love you more each day we are together.

To my daughter Paige: My pride in you is beyond measure. Your strength gives me courage; your life-journey, inspiration; your character, foundation; your kindness, purpose; and your smile and laughter unbridled joy.

To my parents, Kenneth and Muriel Sloat, who brought me into this world, endured my adolescence, supported me, and raised me to be the man I am. Mom, Dad, I am as surprised as you are. I miss you very much.

ACKNOWLEDGEMENTS

Attending graduate school, completing a dissertation, and obtaining a doctoral degree is a team sport. Recognizing this, I would like to thank the following individuals whose support, encouragement, guidance, understanding, and patience made this moment in my life possible:

My Dissertation Committee: Dr. Audrey Amrein-Beardsley, Dr. Ann Ewbank, and Dr. Lori Shough, and Dr. Keith Wetzel. You allowed me to be myself, required me to be a scholar, and believed that this moment was possible. For this, I am profoundly grateful.

The instructional faculty at the Mary Lou Fulton Teachers College, Arizona State University who bestowed upon me an outstanding graduate education, challenged me to question everything, and never allowed me to do anything less than my best.

My colleagues and friends at the school district in which I am employed. This dissertation would not have been possible without your support, encouragement, patience, and participation. You are all co-authors, co-researchers, and co-conspirators in this journey to critically evaluate our roles as professional educators, contribute to our profession's collective knowledge, improve ourselves as practitioners, and impact the lives of the students we serve.

A special acknowledgment to Patti: my compatriot, colleague, friend, and confidant. You made this journey possible, interesting, exciting, meaningful, and important. I am better at what I do because of your professionalism, honesty, integrity, dedication, compassion, and tolerance. We were on this road together and I never felt alone. Thank you for that.

A special acknowledgement to the team in the district's Research Office: Alejandra and Noelle. You are the best at what you do – bar none. Above everything else you had on your plates, you were always willing to discuss, question, challenge, contribute, and reflect on the research presented herein. You read my early drafts, you listened to my endless ramblings, and you kept me grounded. For this, and for many other things, I am forever in your debt.

I cannot leave out the eateries and cafes within which much of this documented was conceived and written over the course of two plus years: The Coffee Bean and Tea Leaf on Bell Road, Peoria Arizona; Paradise Bakery and Café on Bell Road, Surprise, Arizona; and Paradise Bakery and Café on McDowell Road, Avondale, Arizona.

TABLE OF CONTENTS

Page

LIST OF TABLES xxiv

LIST OF FIGURES xxviii

CHAPTER

1 INTRODUCTION AND CONTEXT 1

Introduction 1

Problem Statement 2

Background: Teacher Evaluation and Accountability 4

Arizona's Teacher Evaluation System 7

Arizona Task Force on Teacher and Principal Evaluation 9

Focus of Research 13

Research Questions 15

Theoretical and Conceptual Frameworks 15

Theoretical Framework #1: World Views 16

Theoretical Framework #2: Context of Teacher Evaluation 19

Conceptual Framework for Studying Validity Issues in Teacher Evaluation. 22

Validity and Validity Evidence 23

The Question of Consequence 26

Situational Context: Local Contextual Setting 28

District Demographics 28

District Academic Performance 29

The Researcher 29

CHAPTER Page

Evaluation Context: The District's Teacher Evaluation System 30

Construction and Design Process..... 30

Stated Purpose of the Teacher Evaluation System 32

Constructing a Composite Measure of Teacher Instructional Quality (TIQ) . 33

Danielson Framework for Teaching (FFT)..... 35

FFT Domains 36

Evaluating Professional Practice..... 37

The Professional Practice Evaluation Scale..... 38

Measures of Academic Progress (Test Scores)..... 39

Final Composite Evaluation Scores 40

Prior Action Research Cycles 40

Protocol Field Testing..... 42

Action Research Cycle Reflections 43

Evaluator Observations..... 43

Stakeholder Interviews..... 44

Early Research Journal Reflections 46

Chapter 1 Conclusion..... 46

Preface on Literature Review..... 47

2 LITERATURE REVIEW 50

Schools and Society 50

Validity 64

Historical Context of Validity..... 64

CHAPTER Page

Consequential Validity 74

Full Circle – Validity Components 81

Validity Within Teacher Evaluation Systems 84

Measuring Teacher Effectiveness: Value-Added Models 116

Organizational Change Theory 125

The Concerns-Based Adoption Model (CBAM) 126

Everett Roger's Diffusion of Innovation 128

Community and Change Theory 130

Competency and Empowerment 132

Comment 134

3 METHODS 135

Mixed-Method Designs 136

Primary and Supporting Research Questions 138

Localized Setting 139

Policy Criteria and Impact on Selecting Teacher Sample 141

Impact on Teacher Sample 142

Measures and Data Analysis Plan 143

Research Question #1 143

RQ1A: Criterion Evidence 144

Supporting Research Question RQ1A(a) 144

Supporting Research Question RQ1A(b) 145

Supporting Research Question RQ1A(c) 146

CHAPTER Page

Supporting Research Question RQ1A(d) 147

Supporting Research Question RQ1A(e) 148

RQ1B: Content Evidence 150

Supporting Research Question RQ1B(a) 151

Supporting Research Question RQ1B(b) 152

Supporting Research Question RQ1B(c) 154

Supporting Research Question RQ1B(d) 156

RQ1C: Consequential Evidence 157

Supporting Research Question RQ1C(a) 158

Supporting Research Question RQ1C(b) 158

RQ1D: Reliability Evidence. 159

Supporting Research Question RQ1D(a) 159

RQ1-E: Theoretical Construct Articulation 161

Supporting Research Question RQ1E(a) 161

Supporting Research Question RQ1E(b) 162

Supporting Research Question RQ1E(c) 163

Research Question #2 (RQ2) 163

Supporting Research Question RQ12(a) 164

Supporting Research Question RQ12(b) 164

Research Question #3 (RQ3) 165

Supporting Research Question RQ3(a) 167

Research Question #4 (RQ4) 167

CHAPTER Page

Supporting Research Question RQ4(a) 168

Supporting Research Question RQ4(b) 168

Analytic Methods Supporting Research Questions 169

Multi-Level Value-Added Models of Academic Growth 169

Value-Added Multi-Level Linear Regression Models (VAMLRM) 174

Value-Added Multi-Level Linear Regression Model (VAMLRM) Specification 176

Computing Value-Added Measures of Instructional Effectiveness 181

Measures of Value-Added Model Adequacy 182

Factor Analytic Techniques of Content Validation 185

A Priori Construct Articulation 185

Factor Analysis – Context 187

Content Validity Index 194

Analytic Framework for Qualitative Investigations 201

Qualitative Methods 204

Stakeholder Interview Protocol 204

Observational Protocol 205

Journaling Protocol 205

Qualitative Data Analysis Plan 206

Approach of Qualitative Data Collection and Analysis 206

Data Analysis Methods for Qualitative Information 208

Data Quality Methods 209

CHAPTER Page

Sample Selection of Interview Participants 211

Principal Sample Selection 214

Policy Maker Sample Selection 214

Instructional Growth Coach Sample Selection 214

Evaluation Committee Member Selection 214

4 DATA ANALYSIS 215

Part 1: Descriptive Summary of Quantitative Data Collections 217

Background 217

Quantitative Information: Data Reduction and Selection 219

Teacher Status 220

School Locations 223

Evaluation Component Scores 225

Professional Practice Element Rating Scores 227

Construct Validity Index Questionnaire 229

Part 2: Descriptive Summary of Qualitative Data Collections 230

Introduction 230

Data Verification and Reliability Checks 236

Member Checking 236

Interview Data Summary 238

Research Journal 239

Part 3: Research Questions – Analysis and Findings 241

Introduction 241

CHAPTER Page

Primary Research Question 1 242

Research Question 1A: Criterion Evidence (RQ1A). 243

RQ1A (a) 244

RQ1A (b) 246

RQ1A (c) 247

RQ1A (d) 252

RQ1A (e) 255

Research Question 1B: Content Evidence (RQ1B) 261

The Danielson PP Scale 263

RQ1B (a) 263

Confirmatory Factor Analysis 264

Element Correlations 267

Uncorrelated, Four Factor, CFA Model 267

Correlated, Four-Factor, CFA Model 273

Exploratory Factor Analysis 280

Factor Extraction 281

Summary of Extraction Routines 298

Pattern Matrix Loadings for the Two-Factor EFA Model 300

Pattern Matrix Loadings for the Four-Factor EFA Model 302

Factor Correlations for the Two- and Four-Factor EFA Models 303

RQ1B (b) 304

CHAPTER Page

Summary of Model Fit Information – Instructional Experience 305
Factor Pattern Matrix for Probationary (Three Years or Less Experience) Teachers. 307
Factor Pattern Matrix for Continuing (Four or More Years of Experience) Teachers. 309
Comparative Factor Correlations. 311
RQ1B (c) 313
Construct Validity Index (CVR) 314
Stakeholder Reflections 322
Construct Overview 323
Summary – RQ1B (c) Stakeholder Narratives 346
Petite Assertions – RQ1B (c) Stakeholder Narratives 347
Petite Assertion #1 347
Petite Assertion #2 347
Petite Assertion #3 348
Petite Assertion #4 348
RQ1B (d) 348
Introduction…. 348
Construct Overview 349
Component 1 - General Efficacy (of the Evaluation System) 351

CHAPTER Page

Component 2 - Time, Frequency, and Observation 353

Component 3 - Test Scores/Measurement 354

Stakeholder Narratives 359

Component 1 – General Efficacy 359

Component 2 - Time, Frequency, and Observation (Lack of) 379

Component 3 - Test Scores/Measurement 394

Overall Construct Summary 440

Petite Assertions - RQ1b(d) 441

General Efficacy 441

Time, Frequency, and Observation 442

Test Scores/Measurement (Adequacy and Metric Preference) 443

Research Question 1C: Consequential Evidence (RQ1C) 445

RQ1C (a) and (b) 445

Process 446

Positive Impacts 447

Construct Overview – Positive Impacts 447

Stakeholder Narrative – Positive Impacts 448

Teacher (Positive) 448

Principal (Positive) 454

District (Positive) 466

State (Positive) 470

Summary – Positive Impacts 471

Negative Impacts 473

Construct Overview – Negative Impacts 473

Stakeholder Narrative – Negative Impacts 474

Teachers (Negative) 474

Principals (Negative) 479

District (Negative) 485

State (Negative) 490

Summary – Negative Impacts 495

Petite Assertions: RQ1C (a) 496

Petite Assertion #1 497

Petite Assertion #2 497

Petite Assertion #3 497

Petite Assertion #4 497

Petite Assertions: RQ1C (b) 498

Petite Assertion #1 498

Petite Assertion #2 498

Research Question 1D: Reliability Evidence (RQ1D) 499

RQ1D 499

Reliability Measures of PP Ratings 502

PP Reliability under Conditions of Data Ordinality 502

PP Reliability Indices – Ordinal Alpha 506

PP Reliability Indices – Coefficient Theta 508

PP Reliability Indices – Standard Error of Measure 509

VAM Reliability Indices 511

VAM Reliability Indices – Proportion of Variance Reduction 511

VAM Reliability Indices – Intraclass Correlation Coefficient (ICC) 514

VAM Reliability Indices –Standard Errors of Estimate (SEE) 517

Research Question 1E: Theoretical Construct Articulation (RQ1E) 520

RQ1E (a) and RQ1E (b) 520

Introduction 520

Construct Overview 521

Stakeholder Narrative and Analysis 524

Teachers (What Good/Effective Teachers Do) 524

Principals (What Good/Effective Teachers Do) 529

District (What Good/Effective Teachers Do) 534

State (What Good/Effective Teachers Do) 537

Summary of Findings RQ1E (a) 542

Petite Assertions for RQ1E (a) and (b) 544

RQ1E (c) 545

Introduction 545

Construct Overview 547

Stakeholder Narrative & Analysis 551
Teacher Narrative 551
Principal Narrative 570
District Narrative 583
State Narrative 590
Construct Summary RQ1E (c) 599
Petite Assertions for RQ1E(c) 601
Primary Research Question 2: 602
RQ2 (a) and RQ2 (b) 602
Introduction 602
Construct Overview 604
Stakeholder Narratives 607
Value of Information and Researcher as Information Broker: RQ2 (a) . 607
Value of Information 607
Summary - Value of Information 619
Researcher as Information Broker 619
Summary - Researcher as Information Broker 624
Influence on Decision Making: RQ2(b) 624
Changes in Training Focus 625
Interrater Reliability 627
Changing Perspective, Instilling Confidence and Clarity 628
Post Conference Emphasis 630

SPA Initiative 631

Setting Performance Cut Scores (Ineffective) 632

Correlation between Achievement and Professional Practice Scores 635

Devaluation of Test Scores 639

Summary of RQ2 (b) 642

Overall Summary of RQ2 644

Petite Assertions: RQ2 646

Primary Research Question 3 647

RQ3(a)..... 647

Introduction..... 647

Construct Overview 648

Stakeholder Narratives..... 652

Teachers (Voice) 652

Teacher (Summary) 657

Principal (Voice). 658

Principal (Summary) 666

District (Voice) 666

District (Summary) 670

State (Voice) 671

State (Summary) 673

Voice Summary 673

CHAPTER Page

Petite Assertions for RQ3 (a) 674

Chapter 4 Concluding Comments 674

5 FINDINGS 676

Introduction 676

Summary of Findings 677

Research Question 1 (RQ1) 677

RQ1: Criterion Evidence 678

Criterion Evidence Summary 681

RQ1A: Criterion Assertions 682

RQ1B: Content Evidence 682

CVR 684

Factor Analysis 685

Stakeholder Feedback 685

RQ1B: Content Assertions 686

RQ1C: Consequential Evidence 687

RQ1C: Consequential Assertions 690

RQ1D: Reliability Evidence 692

Professional Practice (PP) Scale Reliabilities 693

VAM Model Precision ..... .695

Qualitative Reliability Measures ..... .696

RQ1D: Reliability Assertions 698

RQ1E: Construct Articulation 698

CHAPTER Page

RQ1E: Construct Articulation Assertions 705

Concluding Remarks Regarding Research Question 1 706

RQ1 Closing Assertions 708

Research Question 2 (RQ2) 710

Summary: Value of Information 710

Summary: Researcher as Information Broker 711

Summary: Influence of Information 712

RQ2 Closing Assertions 713

Research Question 3 (RQ3) 714

Educator Voice 715

RQ3 Summary 717

RQ3 Closing Assertions 718

6 DISCUSSION 720

Introduction 720

Strength and Significance of the Study 721

Limitations of the Study 725

Positionality 725

Qualitative: Limited Stakeholder Subgroup Representation 726

Restricted Teacher Generalization 728

Single District Representation 729

Nested FFT Rating Data 730

Unknown Interrater Reliability 730

CHAPTER Page

Comment on Generalizability 732
Recommendations for Extending the Organization's Current Evaluation System 736
Recommendations for Designing and Implementing Teacher Evaluation Systems 741
Recommendations for Future Evaluation Research 744
Research Question 4: Personal Reflections 748
Introduction 748
Construct Overview 749
Researcher Narrative 751
TEval: Technical Issues/Problems – 2011 Course Schedule 751
TEval: Technical Issues/Problems – Student-Teacher Assignments 754
TEval: Technical Issues/Problems – Problematic Teacher Assignments 755
TEval: Communication and Complexity 756
TEval: Power and Influence/Ineffective Classifications 763
Researcher: Positionality 767
Researcher: Information Broker 770
Researcher: Complexity of Study 774
Summary RQ4(a) 776
Reflection on Analytic Skill Sets 777

CHAPTER Page

Reflections on Qualitative Analysis 778

Printed Documents/Touching the Data 778

Low-Tech 781

Cycles of Analysis 781

Reflections on the Writing Process 783

Transition to Scholarly Writing 785

Difficulty of Transitioning to Scholarly Writing 786

Personal Significance 787

Next Steps 789

Summary RQ4 (b) 790

REFERENCES 791

APPENDIX

A DANIELSON FRAMEWORK FOR TEACHING 810

B OUTLINE OF STUDY RESEARCH QUESTIONS 812

C CONTENT VALIDITY ASSESSMENT QUESTIONNAIRE 816

D SUMMARY OF DATA COLLECTION PROTOCOLS 820

E PARTICIPANT INFORMED CONSENT FORM 824

F INFORMATION LETTER-INTERVIEWS 827

G ALIGNMENT BETWEEN SELECTED RESEARCH QUESTIONS AND QUALITATIVE DATA SOURCES: INTERVIEW PROMPTS, RESEARCH JOURNAL, OBSERVATIONS 829

H DESCRIPTIVE SUMMARY OF QUALITATIVE COMPONENTS 832

APPENDIX Page

I POLYCHORIC CORRELATIONS: FFT ELEMENTS 834

J STANDARD ERRORS OF POLYCHORIC CORRELATIONS: FFT ELEMENTS 836

K Z-VALUE FOR POLYCHORIC CORRELATIONS: FFT ELEMENTS 838

L UNCORRELATED CFA MODEL RESIDUAL COVARIANCE MATRIX 840

M CORRELATED CFA MODEL RESIDUAL COVARIANCE MATRIX 842

N MPLUS CFA SPECIFICATION CODE 844

O PERMISSION TO CONDUCT ORGANIZATIONAL RESEARCH 846

P INSTITUTIONAL REVIEW BOARD APPROVAL LETTER 849

Q ANALYTIC SKILL SETS LEVERAGED FOR STUDY 851

LIST OF TABLES

Table		Page
1	Selected District Demographics	140
2	Teacher Participant Count by Grade Level	143
3	Summary of Models Estimated by Grade Level and Subject.	179
4	Lawshe CVI Minimum Values	199
5	Grade Level Distribution of Teachers	220
6	Distribution of Teachers by Employment Status	221
7	Teacher: Years in District – Descriptive Statistics	221
8	Teacher: Years in District – Frequency Distribution	222
9	Distribution of Group A Teachers Across Elementary Locations	224
10	Descriptive Statistics – Total Evaluation Scores	225
11	Distribution of Assigned Professional Practice Ratings	228
12	Response Rates for Qualitative Data Collections	231
13	Restructured Interview Database – Global Code Group Designations by Stakeholder Group	234
14	Narrative Transcript Re-Coding (Co-Researchers)	237
15	RQ#1A: Criterion Evidence: Five Supporting Research Questions	244
16	Descriptive Statistics – Professional Practice and VAM Scores	245
17	Correlation Between PP and VAM Scores	246
18	Descriptive Statistics for Component VAM Measures	249
19	Kolmogorov–Smirnov Test for Normality	250
20	VAM Correlations at the Classroom Level for Group A Teachers	251

Table Page

21 VAM Correlations Across Years 252

22 Descriptive Statistics – PP Domains 253

23 Kolmogorov–Smirnov Test for Normality: FFT Behavioral Domains 254

24 Spearman's rho Correlations of PP Sub-Domains with VAM 255

25 Teachers Assigned to VAM Percentile Group 258

26 Descriptive Statistics by Growth-Performance Group 259

27 ANOVA Tests of Mean Differences 260

28 Tukey Multiple Comparisons Test of Mean Differences 261

29 Supporting Questions for Research Question 1B 262

30 CFA Model Fit Statistics for Uncorrelated Model 268

31 Standardized Factor Loadings for the Uncorrelated Four-Factor Model 271

32 Uncorrelated Four Factor Model - Correlation of Factor Scores 273

33 CFA Model Fit Statistics for Correlated Model 273

34 Standardized Factor Loadings for the Correlated Four-Factor Model 275

35 Comparative Summary Residual Correlation Statistics 276

36 Correlated Model Factor Correlations 279

37 Factor Extraction Criteria 282

38 EFA Tests of Sample Adequacy 285

39 Component Variance (Eigenvalues) for Full Factor Extraction 286

40 Velicer's Minimum Average Partial (MAP) Test 293

41 Comparative Eigenvalues, PAF Versus Parallel Analysis 295

42 Summary of Chi-Square Model Fit Information 297

Table Page

43 Summary of Extraction Criterion 298

44 EFA Two-Factor OBLIMIN Rotated Pattern Matrix 301

45 EFA Four-Factor OBLIMIN Rotated Pattern Matrix 302

46 Oblimin Factor Correlations for the Two- and Four-Factor EFA Models 303

47 Summary of Model Fit Information by Instructional Experience 305

48 Eigenvalues for Probationary and Continuing Instructional Group EFA 306

49 Pattern Matrix for a Four- and Two-Factor Model for Probationary Teachers 308

50 Pattern Matrix for a Four-, Three-, and Two-Factor Model for Continuing Teachers 310

51 Factor Correlations for the Probationary Two- and Four-Factor Models 312

52 Factor Correlations for the Continuing Four-, Three-, and Two-Factor Models 313

53 Descriptive Sampling Information for the LCVR 316

54 LCVR Values for SMEs Rating of Danielson FFT Elements 319

55 Lawshe CVI by Category 321

56 Ordinal Alpha Reliability Estimates for FFT Evaluation Domains 507

57 Ordinal Theta Reliability Estimates for FFT Evaluation Domains 509

58 FFT Scale Descriptive Statistics 510

59 Ordinal Alpha Reliability Estimates for FFT Evaluation Domains 510

60 PVE VAM-Model Indices by Grade by Subject 513

Table Page

61 ICC Variance Components for Unconditional and Conditional VAM Models..... 516

62 Conditional Model Level-1 95% Confidence Interval..... 519

63 Summary of Criteria Evidence..... 678

64 Summary of Content Evidences..... 683

65 Summary of Reliability Evidences 693

66 Outline of Discussion Topics Addressed Under RQ4 (a)..... 750

67 Outline of Discussion Topics Addressed Under RQ4 (b)..... 751

68 Stakeholder Comfort in Explaining Evaluation Components..... 762

LIST OF FIGURES

Figure Page

1 Components of a Pragmatic Approach for Examining Teacher Evaluation Systems 19

2 Conceptual Framework Placing Teacher Evaluation in a Broader Social Context 20

3 Sources of Validity Evidence 23

4 Example Teacher Instructional Quality Scale 35

5 Danielson's Four Primary Domains and Component Elements 36

6 Total Point Structure for Danielson's Four Primary Domains 39

7 Components of the Study's Theoretical Framework 47

8 Organizational Structure of the Literature Review 49

9 Messick's Facets of Validity 72

10 Comparison of Validity Construct From Selected Sources and Authors. 76

11 Example of Concurrent Mixed Methods Approach to Construct Examination .. 137

12 An Example of a Regression Line Representing the Best Fit Between One Covariate and the Dependent Variable 171

13 Depiction of Predicted Versus Actual Scores in a Value-Added Context 174

14 A Depiction of a Two-Level Data Structure of Students Nested Within Schools 177

15 Covariates Incorporated Into Each Level of the Model 178

16 Example Math Test Item Correlations 188

17 Hypothetical Factor Extraction From a Six Item Math Test 189

Figure Page

18 Rating Element Alignments for the Danielson Framework for Teachers..... 190

19 CFA Structural Representation of the Danielson FFT Framework 193

20 Carmines and Zeller's Ordered Sequence of Test Content Validation..... 197

21 Qualitative [GT] Data Analysis Stream: Protocol Adaptation and Knowledge Banking 207

22 Generalized Coding Framework for Analyzing Qualitative Information..... 209

23 Primary Questions and Components..... 215

24 Study Components 216

25 Histogram of Teachers' Experience (Years in District) 223

26 Distribution - Evaluation Scores..... 226

27 Distribution - Professional Practice Scores..... 227

28 Distribution - VAM Scores..... 227

29 Theoretical Structure of the Danielson Framework for Teaching (Uncorrelated Model)..... 265

30 Theoretical Structure of the Danielson Framework for Teaching (Correlated Model)..... 266

31 Parameter Estimates for the Correlated CFA Model 280

32 SPSS/Mplus Factor Scree Plots 289

33 Codes and Identities Delineating Missing/Incomplete Attributes 324

34 Codes and Identities Related to Adequacy of Existing FFT Components..... 325

35 Codes and Identities Delineating Instructional Complexity. 326

36 Component Structure for Research Question RQ1B (d)..... 350

Figure Page

37 Codes and Identities Extracted From the Stakeholder Narratives Delineating General Efficacy (of the Evaluation System) 352

38 Codes and Identities Extracted From the Stakeholder Narratives Delineating Time, Frequency, and Observation 354

39 Two Sub-Concepts of Test Scores/Measurement 356

40 Codes and Identities Extracted From the Narrative Information Related to the Test Score Adequacy 357

41 Codes/Identities for PP = TS (Aligned) and PP > TS (Not Aligned). 358

42 Codes/Identities for Test Score > Professional Practice. 359

43 Codes and Identities Related to Clarity, Focus, Structure. 448

44 Codes and Identities Related to Reflection, Communication, Dialog. 448

45 Codes and Identities Related to Fear/Stress 473

46 Codes and Identities Related to Conformity/Reductionism 474

47 RQ1E Theoretical Construct Articulation ..... .520

48 Codes and Identities Related to Providing a Safe/Nurturing Environment ..... .522

49 Codes and Identities Related to Mentor, Support, and Motivate 523

50 Components for Global Construct - Purpose of Education. 548

51 Codes and Identities Evaluation-to-Improve-Practice 550

52 Codes and Identities Evaluation-as-Accountability 550

53 Teacher Concept Components 569

54 Codes and Identities Delineating Value. 606

55 Codes and Identities Delineating Influence. 606

Figure Page

56 Codes and Identities Related to Lack of Voice. 650

57 Codes and Identities Related to Input/Positive 651

58 Codes and Identities Related to Feedback vs. Decision Making. 652

59 Summary of Consequential Reflections by Stakeholder Attribution. 688

60 Attributes of the Teacher Instructional Quality Construct ..... .700

Chapter 1: Introduction and Context

Introduction

Public education agencies throughout the country utilize varied approaches for evaluating and reviewing the instructional practices of classroom teachers (McClellan,2012; Learning Sciences Marzano Center, 2012a; Marzano, 2011; Strong, 2012E; MET Project, 2010). The purpose of these systems has primarily focused on providing reflective feedback to teachers, improving instructional practice, and increasing efficacy on student learning (Milanowski, 2011; Stumbo & McWalters, 2010; Kimball & Milanowski, 2009). However, these evaluation approaches have not traditionally been directly integrated into state or national-directed accountability systems. That is, direct measures of teacher efficacy at the individual classroom level have not traditionally been a primary component of policy-directed education accountability or reform (Erpenbach, 2009, 2011).

In 2009, the role of teacher evaluation within high stakes accountability environments changed substantively under the federal Race to the Top (RttT) program (USDOE, 2009). Under this program, states were required to implement systemic teacher evaluation systems in order to qualify for competitive education grants. Additionally, these systems required states to incorporate quantitative measures of student learning as a primary component of the evaluation process. For states such as Arizona, this necessitated legislative adjustments to existing education policy such that all district evaluation systems incorporate (at minimum) measures associated with classroom professional practice and student academic progress (Ariz. Rev. Stat. §15-203A.38, 2010).

In Arizona, School Year 2012-13 (SY2012-13) represents a pilot year for districts to design and implement these systems. High stakes reporting and accountability rules do not begin until SY2013-14.

Problem Statement

As reflected in the latest edition of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), "… validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). In this regard, the *Standards* position such evidence as the most fundamental consideration when justifying inferences and decisions made from scores. The *Standards* go on to state that "… the process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations" (p. 9).

Arguably, the act of implementing score-based instructional evaluation and accountability systems from which judgments of professional competency and practice are based require examination of inferential soundness. It seems reasonable to suggest that the efficacy of such systems is impacted by policy decisions made at the local level. It follows that examination of localized validity evidence is critical for informing decision making and ensuring efficacy of such systems (Danielson, 2010). However, for local education agencies (LEAs) designing and implementing their own systems, little such evidence exists. To this end, Danielson (2010) argues that,

> … credibility in an evaluation system is essential. A principal or a superintendent must be able to say to the school board and the public, "… everyone who teaches here is good and here's how I know." (p. 36)

Similarly, Fast and Hebbler (2004) argue that in any type of accountability system "… validation evidence is necessary to support the accountability claims made about individuals and agencies and the accompanying imposition of stakes" (p. 2).

In this context, this study examines the validity evidence associated with implementing Arizona's policy-directed teacher evaluation framework within a large suburban school district in Phoenix, Arizona. Using a concurrent mixed methods approach (Gelo, Braakmann, & Benetka, 2008; Creswell, 2009; Plano-Clark & Creswell, 2010; Greene, 2007), it situates the research in both an evidentiary and larger policy context and explores the impact such evidence have on the system's implementation as a high stakes consequential decision system. In addition, a teacher advocacy perspective is adopted, exploring stakeholder attitudes regarding the evaluation process (Creswell, 2009; Stringer, 2007).

The research is framed by a hierarchy of theoretical constructs which positions teacher evaluation within a local LEA policy context, a larger state/national policy context of systemic evaluation and school accountability, and an overarching context of public education in society. Each contributes to the validity examination by providing contextual meaning for implementing teacher evaluation systems and provides a basis to evaluate evidence. In addition, this study utilizes a unified conceptualization of validity advanced by Messick (1989a) as stipulated in the *Standards* (AERA et al., 1999, 2014). Under this conceptualization, teacher instructional quality is measured from a variety of perspectives using multiple sources of evidence.

Finally, the research examines how the process of developing and examining validity evidence impacts the organizational policy context related to the system's implementation.

In Chapter 1, background information on accountability and teacher evaluation is discussed followed by an introduction to Arizona's efforts to implement a new systemic policy-directed evaluation system. The focus of research is then introduced along with the global research questions. Theoretical and conceptual frameworks are then thoroughly discussed in order to position the work in a larger context and to provide the basis of the research designs detailed in Chapter 3. The contextual (situational) setting of the research is reviewed in order to provide the reader an understanding of locality. Finally, Chapter 1 ends by revisiting the research questions as a preface to reviewing the pertinent literature.

Chapter 2 examines the literature relevant to this study. It is organized according to four main constructs: schools and society, validity theory, teacher evaluation systems, and value-added models of instructional effectiveness. Chapter 3 presents the study's research design and methods while Chapter 4 provides an analysis of the relevant data. Study findings are more thoroughly discussed in Chapter 5 followed by pertinent conclusions and discussions in Chapter 6. This chapter also provides reflections on the research and dissertation writing process, a review of study significance and limitations, and suggestions for further research.

Background: Teacher Evaluation and Accountability

Over the past two decades, an increasing number of states and selected LEAs have begun to utilize structured evaluation systems in their attempt to quantify the

impacts of classroom instruction and evaluate the efficacy of instructional practice (Blank, 2010). A popular approach for doing so involves the use of value-added models, a statistical technique intended to measure changes in achievement attributable to instructional intervention (McCaffrey, Lockwood, Koretz, & Hamilton, 2003, 2007; Hill, 2009; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; Braun, 2005). These models compare students' predicted performance to actual achievement after adjusting for non-instructional background factors and prior learning history. The difference is interpreted to be the instructional effect ascribed to the classroom teacher: the greater the deviation, the greater the instructional effect.

For example, the state of Tennessee has been utilizing value-added models to isolate and measure instructional effects since 1992 (Kupermintz, 2003). Many other states have undertaken similar activities (Blank, 2010). Large districts such as Austin ISD, Los Angeles, New York City, Houston, Cincinnati, and Chicago public schools have utilized similar statistical techniques to measure classroom instructional effects based on standardized student achievement scores (Watanabe, 2011; Pearson, & Yan, 2012; Amrein-Beardsley & Collins, 2012; Baum, 2010; Corcoran, 2010; Weisberg, Sexton, Mulhern, & Keeling, 2009; Austin Independent School District, 2012).

Indeed, LEAs such as New York City and Los Angeles have been the focus of news headlines for publically releasing teacher evaluation and effectiveness ratings to the general public (Watanabe, 2011; Baum, 2010; Butrymowicz & Garland, 2012). And in spring 2012, teachers in Chicago Public Schools made headlines when contract negotiations stalled due to disputes based, in part, on the relative importance given to the achievement component of the LEA's proposed teacher evaluation system, arguing that

test scores should constitute less influence in the overall assessment of teaching quality (Pearson & Yan, 2012).

The presence of highly competitive RttT grants worth millions of dollars, coupled with declining education funding and easily accessible statistical procedures, have contributed to an evolving new norm in America's policies toward teacher evaluation. Under an "audit culture" (Apple, 2007), this new norm is the expectation that education accountability systems at the LEA and school level must rely on quantitative measures of student learning and professional practice as the basis for evaluating the instructional quality of individual classroom teachers. Here, Apple argues that contemporary school reform efforts based on a managerial (accounting) perspective of governance has not resulted in substantive improvements. The logical facility, Apple (2007) argues, is a perspective of that "… only that which is measurable is important …" and that this approach is causing "… some of the most creative and critical [education] practices that have been developed through concerted efforts in some of the most difficult settings to be threatened" (p. 4). Similarly, referring to the United States (U.S.) Department of Education's Teacher Incentive Fund, Kimball and Milanowski (2009) reflect that the "… premise of this law is that school leaders can identify more effective teachers through performance evaluations" (p. 35). Henry et al. (2012) note that this view extends beyond public K-12 agencies and into college/university teacher preparation programs:

> A new era of accountability for teacher preparation programs is being ushered in across the country by recent federal and state policies … which require that these programs be held accountable for producing effective teachers. New accountability reforms define effective teachers, at least in part, as those who produce higher student test score gains. (p. 335)

As of March 2015, 19 states received federal RttT grant monies. Each of these states are required to implement education policy and reform efforts that prescribe formal systemic teacher evaluation systems (USDOE, 2015). Arizona's application was awarded in December 2011 under Phase 3 of the federal program. A description of the state's teacher evaluation policy reforms is provided in the following section. An understanding of the policy perspective in which Arizona's activity is situated is important in formulating the conceptual context for examining validity evidence and for positioning the research activity into a larger social context.

Arizona's Teacher Evaluation System

In spring 2011, the Arizona State Legislature passed Senate Bill 1040 (Ariz. Rev. Stat. §15-203A.38, 2010). Subsequently signed into law by the Governor, this legislation requires all public school districts and charter schools in the state to evaluate teacher effectiveness using (at minimum) quantitative measures of student academic progress and measures of professional practice beginning in SY2013-14. Under this legislation, between 33 and 50% of an individual classroom teacher's evaluation rating must be based on the academic progress of students in their classrooms. In addition, between 50 and 67% of a teacher's evaluation rating must be based on "best practices" measures of classroom instructional quality. As will be discussed, the final mix of weights is determined by each individual district. In the study district, the decision was made to initially weight test scores at 33% (the minimum allowed) and classroom practice at 67% (the maximum permitted). This was due to the uncertainty with developing reliable and valid measures of academic progress. Policy leaders felt that once the system had been

implemented and more empirical information obtained, revisions to the weighting mix could be considered.

Importantly, Senate Bill 1040 did not mandate a specific methodology or set of measures for evaluating teacher effectiveness. Rather, it directed the State Board of Education (SBE) to develop a general framework under which districts would be free to construct and implement their own evaluation systems and related metrics of instructional quality. However, the legislation did specify that all measures used to evaluate teachers and principals must meet "… the data requirements established by the State Board of Education …" (Task Force on Teacher and Principal Evaluations, 2011). Subsequently the SBE directed that all measures used in the evaluation of teacher instructional quality (TIQ) must be both reliable and valid, aligned to Arizona's academic standards, and appropriate to a teacher's content area (Arizona Department of Education, 2012a). In cases where achievement measures were unavailable, school level data from state-wide standardized assessments must be utilized.

At the same time that this legislation was being proposed, the Arizona Department of Education and the Governor's Office were applying for the RttT competitive grant monies. As mentioned above, to successfully compete for these monies, states were required to document the implementation, or planned implementation, of systemic accountability systems that specifically evaluate individual teacher's performance. Accordingly, without passage of SB1040 and the ability to document progress toward use of these types of systems in LEAs throughout the state, Arizona would not have qualified for federal funding.

Arizona Task Force on Teacher and Principal Evaluations

To both interpret and implement the new legislation, the Arizona State Board of Education convened an advisory committee, the Task Force on Teacher and Principal Evaluations (Task Force). This committee was composed of business, community, educator, and policy members (Butterfield & Amator, 2012). Their charge was to develop a framework under which districts could construct and implement structured teacher evaluation systems (Task Force on Teacher and Principal Evaluations, 2011). Meetings were held between October 2010 and January 2011 to review various aspects of the activity. Framework recommendations were drafted in January 2011 and subsequently adopted by the SBE in April 2011 (Butterfield & Amator, 2012).

The Task Force positioned its work by interpreting the legislative intent stating:

> Outstanding teachers and principals make a difference. Great classroom teaching and principal leadership are the strongest predictors of student development and achievement. Based on this reality, in 2010 Arizona legislators passed a law intended to change the culture of education in Arizona, and improve how many LEAs evaluate their teachers and principals. (Task Force on Teacher and Principal Evaluations, 2011, p. 3)

Arguably, based on the language used in the final report, the Task Force saw its work in the following public policy context. First, classroom and school leadership is declared to be the strongest predictor of student achievement. If achievement is low or high, then classroom/school leadership is causally responsible. Second, current evaluation systems used by Arizona LEAs are inadequate, requiring a "…change in culture…", and this change in culture must be accomplished via legislative action and state policy directive. Third, the resulting state policy directive constitutes an improvement over current evaluation practices used by the state's school systems. The Task Force further clarified its position by stating:

> The Task Force on Teacher and Principal Evaluations conducted its work in service of the students in Arizona's public schools. The Task Force members hold that the goal of both teacher and principal evaluations is to enhance performance so that students receive a higher quality education. Further, the work here submitted reflects the belief that evaluations are most effective as one part of a systemic approach to improving educator performance and student achievement. (Task Force on Teacher and Principal Evaluations, 2011, p. 1)

Again, the perspective is one of helping students achieve higher levels of learning by improving teacher and administrator performance. The perspective presumes that this is best accomplished from a systemic (i.e. state policy) approach. In defining their vision statement, the Task Force report offered the following:

> To improve student achievement, Arizona supports effective teachers and principals by developing a model framework that can be incorporated into all Arizona LEA evaluation instruments and ensures that student academic progress is a significant component in the teacher and principal evaluation process. (Task Force on Teacher and Principal Evaluations, 2011, p. 1)

This vision statement also reinforces the systemic statewide perspective and positions the concept of academic progress as a direct measure of instructional quality. Thus, the emerging evaluation context situates (1) the necessity to evaluate teacher/principal quality because of its causal impact on student learning, and (2) the necessity of incorporating student learning into the evaluation to obtain a better measure of quality. Thus, the following logic structure emerges from the policy context:

1. Evaluating teachers improves the quality of instruction.
2. Because the quality of instruction is a causal predictor of achievement, evaluation will therefore improve levels of student learning.
3. Thus, it follows that observed levels of student learning directly informs on the quality of instruction.

This logic forms the conceptual context of the policy-directed evaluation system in Arizona. It is premised on the causal alignment between instructional quality and learning and that the act of evaluating improves quality.

Finally, the Task Force advanced specific goals for its work in developing a statewide policy-directed evaluation framework. The goals were stated as follows:

- To enhance and improve student learning;
- To use the evaluation process and achievement data to drive professional development to enhance teaching, leadership, and student performance;
- To increase data-informed decision making for students and teacher and principal evaluations fostering school cultures where student learning and progress is a continual part of redefining goals for all;
- To use the evaluation process and data to improve teacher and principal performance;
- To incorporate multiple measurements of achievement;
- To communicate clearly defined expectations;
- To allow LEAs to use local instruments to fulfill the requirements of the framework;
- To reflect fairness, flexibility, and a research-based approach;
- To create a culture where data drives instructional decisions (Task Force on Teacher and Principal Evaluations, 2011, p. 1).

Arguably, these goals are multifaceted and convey policy makers' perspectives on what is required in Arizona's school system: that is, the need to improve student learning by increasing the efficacy of teachers and principals, focused professional training of

practitioners, increased reliance on data to drive decision making, and a need to convey "clear" expectations of practitioner's performance. To operationalize these goals, the SBE/ADE presented the following guidelines to LEAs throughout the state (Butterfield & Amator, 2012):

- When available, data from statewide assessments shall be used to inform the evaluation process.
- All assessment data used in educator evaluations shall be aligned with Arizona State Standards.
- LEAs shall include student achievement data for reading and/or math as appropriate; however, student achievement data should not be strictly limited to these content areas.
- Evaluation instruments should integrate student academic progress data with data derived through classroom observations - neither should stand alone.
- All evaluators should receive professional development in the form of Qualified Evaluator Training.
- LEAs should provide for the development of classroom-level achievement data for teachers in those content areas where these data are limited or do not currently exist so that all teachers use the Group A framework. [Reader's Note: *Group A* refers to classroom teachers instructing state-adopted curriculum standards that are also assessed using reliable and valid tests. Group B teachers are those instructing in content areas lacking reliable and valid assessments]

- LEAs should develop and provide professional development on the evaluation process and in those areas articulated in Arizona's Professional Teaching and Administrative Standards, as approved by the State Board of Education

These, then, form the basic tenets for designing and implementing teacher evaluation systems across all public and charter schools in the state of Arizona.

Focus of Research

As mentioned, this paper examines the validity evidence associated with implementing Arizona's policy-directed teacher evaluation framework within a large suburban school district in Phoenix, Arizona. Utilizing a mixed methods approach, validity is seen as a unifying concept composed of multiple perspectives (AERA et al., 1999, 2014; Messick, 1989a; Kane, 2001; Gorin, 2007; Brualdi, 1999). Each perspective contributes evidence on the inferential suitability of the evaluation system to its intended purpose. In addition, the research is situated within an applied, high stakes, consequential, environment with the potential to impact the long-term professional and personal lives of teachers. As such, quantitative examinations of scale reliability and criterion associations are extended to include examination of stakeholder perspectives, attitudes, and their concurrence to system intentions. Messick (1989a) comments on the need for this type of inclusive examination by stating:

> To validate an action inference requires validation not only of score meaning but also of value implications and action outcomes, especially appraisals of the relevance and utility of the test scores for particular applied purposes and of the social consequences of using the scores for applied decision making. (p. 13)

Finally, this research is directly positioned to inform policy makers both within the hosting LEA and the larger state policy context. It is posited that decision makers at the LEA policy level benefit from access to evidence which inform on the system's

suitability and consequential impact. In addition, by documenting and advancing stakeholder perspectives, it is believed that policy makers are able to implement a system better aligned with intended outcomes. Arguably, implementing a new evaluation system signifies significant change in the organization and its membership. In this regard, Fullan (2009) recognizes the importance of validity examinations to organizations implementing this type of substantive change. His Theory of Action for System Change (TASC) specifies six core components of which component five concerns continuous evaluation and inquiry. Of this he states

> … a theory, in essence, is a set of tested hypothesis about reality. It is never assumed to be valid once and for all. Rather … a theory must always be subject to assessment … [and] constant evaluation and inquiry must be built into the mindset of the reform and the actions of the center and those in the field as the reform progresses. (p. 288)

This is consistent with many theories of change where the act of continuous critical evaluation of implementation fidelity is central to organizational growth and sustainability (Rogers, 2003; Kotter, 1996; Hargreaves & Shirley, 2009).

In a larger context, this researcher hopes that the work herein will contribute to and inform on policy perspectives taking shape at both state and national levels with regard to evaluation, school accountability, and education reform. Indeed, it is because of these external policy decisions that this localized evaluation study must be cast within a broader framework. To do otherwise would be to ignore the influencing social constructions driving its implementation (Gergen, 2009). For this reason, findings from the study are intended to add to the broader body of work related to teacher evaluation systems and the socio-policy context under which they are developed.

Research Questions

The primary research questions posed by this study include the following:

1. To what degree do validity evidences (scale reliabilities, criterion associations, content coverage, and consequential effects) generated by the LEA's state policy-directed teacher evaluation system support inferences on Teacher Instructional Quality (TIQ)?
2. How does providing an evolving dialog on validity evidence influence organizational policy decisions regarding implementation of the LEA's teacher evaluation system?
3. To what degree have the perspectives and concerns of stakeholders been incorporated into, and influence, the system's implementation?
4. How did the process of engaging in this research study impact the investigator as a scholarly researcher and organizational leader?

Examination of these research questions employs both quantitative and qualitative analytic approaches. Based on the theoretical and conceptual frameworks discussed below, it is argued that these collections of evidence inform on the latent construct of interest, Teacher Instructional Quality, and the degree to which the evaluation system provides a suitable foundation for interpreting and acting upon teacher's professional practice.

Theoretical and Conceptual Frameworks

The design of any research investigation is dependent upon the theoretical constructs defining the phenomenon of interest. Creswell (2009) comments:

> Research designs are plans and procedures for research that span the decisions from broad assumptions to detailed methods. The overall decision involves which design should be used to study a topic. Informing this decision should be the world view assumptions the researcher brings to the study … (p. 4)

In this regard, the following discussion outlines the theoretical and conceptual constructs within which the research is both situated and guided. Theoretical frameworks from a world view perspective are first discussed to provide grounding of the proposed methodological approach. Here, a pragmatic analytic perspective is advanced. This theoretical framework is then extended by positioning teacher evaluation within a broader school and societal context, establishing the basis for discussing specific approaches to gathering validity evidence relevant to teacher evaluation. Finally, a conceptual model for examining teacher evaluation is constructed based on these theoretical foundations. Taken together, the composite conceptual frame provides purpose and direction for each of the individual design activities and methodological approaches required to explore the research questions established for this study.

Theoretical Framework #1 - World Views

Creswell (2009) argues that every research activity is influenced by the researcher's particular world view. He states that by making these perspectives explicit, it informs the reader of the context under which particular research designs and methods are adopted. In addition, it allows the researcher to acknowledge, explore, and (hopefully) attend to issues that may impact or influence the nature of the investigation undertaken.

To this end, Creswell (2009) outlines four generalized world views: postpositivist, constructivism, advocacy/participatory, and pragmatism. He positions a postpositivist view as being associated with the scientific method and empiricism where a researcher searches for evidence of phenomena. However, claims of certainty can never be made

due to the presence of error in all observation, measurement, and examination. This world view traditionally aligns with quantitative approaches to research questions.

In contrast, the constructivist view is more aligned to qualitative approaches and holds that individuals develop subjective meanings of the world around them. Constructivists seek to understand phenomena within the context of personal relationships, meanings, and the complexities of effects impacting individuals (Creswell, 2009). Gergen (2009) expands on this view by stating that knowledge and meaning are constructed within the context of social relationships and that meaning and interpretation cannot exist outside of collective individuals. That is, meaning is *socially constructed* and therefore research must be contextualized within this framework. In this regard, research conducted under a constructivist world view attempts to identify and represent the perspectives of the individuals involved or impacted.

An advocacy/participatory world view positions the research activity within a political-power context that acknowledges marginalized individuals in society (Creswell, 2009). This world view attaches an action/reform agenda to the inquiry with the intent to improve the quality of the lives of those impacted. Finally, the pragmatic world view is concerned with finding solutions to issues or problems without regard to any particular methodological approach or tradition (Creswell, 2009). Examination of phenomena encompasses all methods the researcher finds appropriate for providing insight and knowledge on the issue.

For this researcher, the issue of examining validity evidence associated with implementing high stakes teacher evaluation systems within a public school district setting involves aspects of all four of these world views. Using quantitative metrics as

representations of TIQ necessitate exploration of the underlying measurement assumptions, reliabilities, and associations between the component scales. At the same time, TIQ is inherently a socially constructed construct that has meaning only when positioned within a commonly defined (social) context. From this perspective, validity examinations need to explore these socially-based constructs inclusive of all stakeholders and evaluate them against the system's stated purpose and intent (Messick, 1989a, 1998; AERA et al., 1999, 2014; Shepard, 1993, 1997).

Similarly, implementing a policy-directed evaluation system situates this activity as a condition of employment and professional membership. That is, it is being imposed upon teachers (and principals) regardless of whether or not they subscribe to its conditions. Here, issues of power and marginalization may play a role in stakeholder acceptance, system implementation, and its efficacy to change professional practice and improve student learning. In this way, an advocacy perspective is warranted in order to better understand, and give stature to those upon which the activity is directed (Creswell, 2009; Stringer, 2007). To the extent that many contemporary theories of organizational change integrate stakeholder communication, involvement, and empowerment as vital components of successful innovation, excluding stakeholder perspectives would ignore an important aspect of the validation process.

From a pragmatic perspective, the process of investigating the validity evidence of any system containing multiple facets, influences, stakeholders, and consequences suggests that a mixed methods approach is warranted. Indeed, in the context of this study, the pragmatic world view acts as an umbrella methodological construct to bring together

various research traditions to explore the validity evidence of teacher evaluation systems. Figure 1 provides a schematic view of these perspectives.

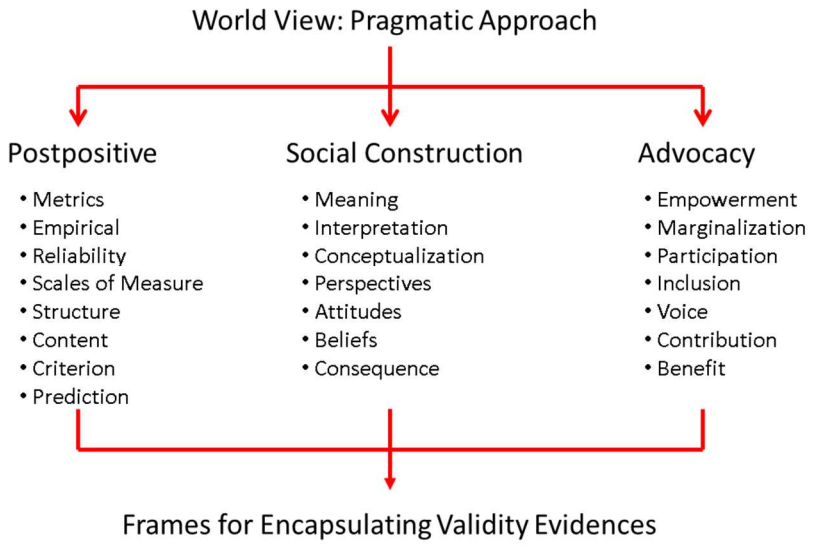


Figure 1. Components of a pragmatic approach for examining teacher evaluation systems.

As depicted, each world view assimilates different aspects of the construct being investigated, implying different methodological approaches (or frames) to examine different perspectives. For evaluating teacher evaluation systems, this implies using multifaceted or mixed methods appropriate to the construct being investigated.

Theoretical Framework #2 – Context of Teacher Evaluation

Arguably, examining validity issues related to teacher evaluation recognizes that the activity is positioned in a larger conceptual context and that this context informs on appropriate methodological approaches, data collections, and inferential analysis. Indeed, by design, teacher evaluation systems are embedded within a context related to the purpose and function of public schooling in society (Gergen, 2009; Dewey, 1900; Good,

1999; McGee-Banks & Banks, 1997). That is, the desire to develop measures of effective teaching is premised on the need for teachers to engage in high quality instruction that facilitates student learning. However, what is taught (curriculum) is seen as a social agreement aligned to both the role of teachers in schools and schools in society (Dewey, 1900). In this manner, teachers are positioned as the main agents of a larger educational process. In turn, the impact that teachers affect is conceptualized as teacher (instructional) quality and the study of these impacts is generalized as the activity of teacher evaluation. This conceptual framework is depicted in the Figure 2.

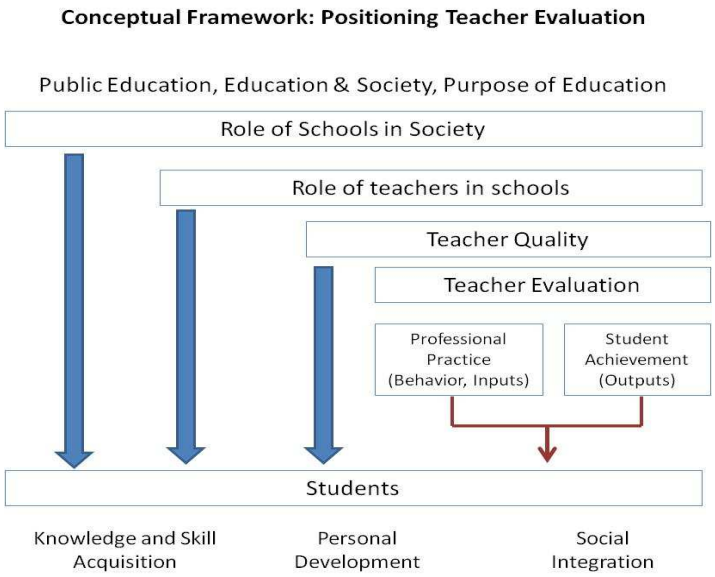


Figure 2. Conceptual framework placing teacher evaluation in a broader social context.

Conceptualizing the positioning of the evaluation activity in this way provides additional theoretical context on the dimensionality of the validation task. As discussed above, the positioning suggests that to fully examine an evaluation system, one must

extend beyond the immediate metrics used to quantify teachers' professional behaviors and/or achievement of students. Rather, because the activity of teaching and learning is positioned within a social (and arguably political) context, an accounting of such systems must incorporate the prominent interconnections binding these components together. The pragmatic approach permits utilization of a variety of methods, data collections, and inferential approaches to understanding these interconnections within a social context.

Together, both a constructivist and pragmatic approach necessitate inclusion of the social constructions that serve to define the evaluation construct (Creswell, 2009; Gergen, 2009). These social constructions include:

- The meaning of effective/ineffective instruction (to assess the alignment of evaluation);
- Acceptable/unacceptable representations of student learning (to assess the representativeness of the learning measures);
- Proper/improper approaches to measuring TIQ (to assess appropriateness of methodological approach); and
- An understanding of the consequential policy decisions emanating from the evaluation activity (to assess alignment of judgments to the established purpose and intent teacher evaluation).

Not to do so limits the validity examinations of such systems (Shepard, 1993, 1997; Kane, 1992b; Messick, 1989a, 1989b, 1998; AERA et al., 1999, 2014). Indeed, failure to acknowledge these dimensions risks leaving essential perspectives unexamined, potentially biasing the inferential judgments or conclusions.

Conceptual Framework for Studying Validity Issues in Teacher Evaluation

The above discussion permits development of a conceptual framework for examining facets of validity applicable to teacher evaluation systems. As will be described in greater detail below, this study adopts a single unifying view of the validity construct defined by interconnected, socially constructed, elements impacting the evaluation process (Messick, 1989a, 1998; Gergen, 2009; Kane 2001; AERA et al., 1999, 2014). These constructions include the following but are not limited to:

- The purpose of schools and teaching
- The role of teachers in schooling
- The purpose of teacher evaluation
- The definition(s) of quality instruction
- The frameworks defining instructional 'best' practices/behavior
- Defined student learning objectives (curriculum)
- Scales of student learning (achievement, outcomes)
- Scales of instructional quality (attributes, behaviors)

These elements may be further categorized to help inform the research activity. Defining the purpose of schooling and the teacher's role in education provides a context for comparing an evaluation system's design and utilization within a larger social context and purpose. Similarly, defining attributes associated with quality instruction, best practices, curriculum, and learning objectives, permits closer examination of system design and its consistency/correspondence. Finally, by articulating scales of learning and instructional quality, measures may be derived for the purpose of interpreting the

magnitude, ordering, and direction of the construct (AERA et al., 1999, 2014). Figure 3 depicts separation of these elements and their relationships.

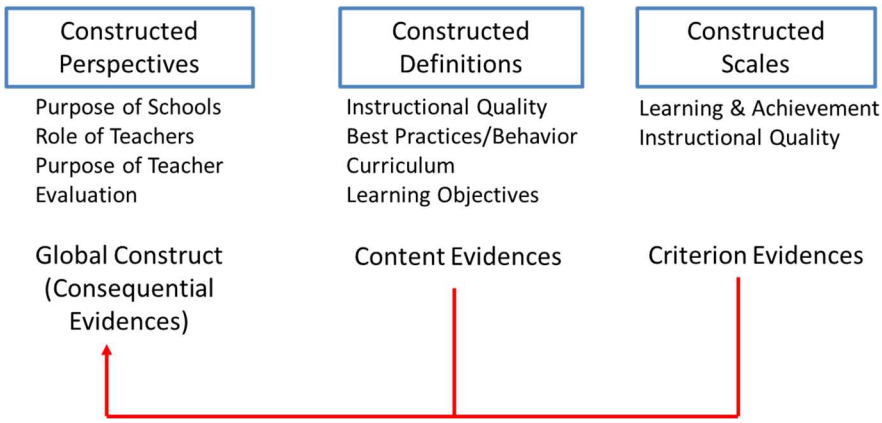


Figure 3. Sources of validity evidence.

Validity and validity evidence. As referenced in the *Standards for Educational and Psychological Testing* (AERA et al., 1999), validity "… refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). This discussion of validity is the first chapter in the *Standards* publication and serves as the conceptual foundation after which all other discussions of testing and measurement are positioned. Its placement signifies the foundational role that validity evidence plays in any type of measurement system including those related to teacher evaluation systems. Messick (1989a) begins his influential chapter in the third edition of *Educational Measurement* by stating:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment. … Broadly speaking, then, validity is an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use. Hence, what is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators. (p. 13)

As adopted in the most current edition of *Standards*, construct validity is seen as a single unifying construct under which many different types of evidence contribute (AERA et al., 1999, 2014) However, this was not always the case. In the 1950s, construct validity was initially characterized by Cronbach and Meehl (1955) as only one of a number of different types of validity (Westen & Rosenthal, 2005). It was not fully accepted by researchers as a unifying construct until the end of the 1970s (Kane, 2001) and subsequently articulated by Messick in the 1989 edition of *Educational Measurement* (Messick 1989a; Kane 2001).

Contemporary perspectives outlined in the *Standards* (AERA, APA, & NCME, 2014) refer to (construct) validity as the following:

1. "Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests" (p. 11).
2. "Validation logically begins with an explicit statement of the proposed interpretation of test scores, along with a rational for the relevance of the interpretation to the proposed use. The proposed interpretation includes specifying the construct the test is intended to measure." (p. 11).
3. "Validation can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use." (p.11)

Each of these representations position validity as a collection of evidence that support the relationship between measures and the underlying latent construct they purport to represent. This perspective does not differentiate specific types of validation evidence. Rather, collections of evidence are gathered from a wide variety of perspectives

for the purpose of constructing a single validation argument. Applied to teacher evaluation, this perspective is consistent with the theoretical constructs discussed above, integrating a pragmatic world view with embedded social constructs. In this regard, the conduct of validity studies requires multiple perspectives and methodological approaches.

Kane (2001) points out that advancing a single validation construct differs from historical utilizations of validity in which specific types of evidence were purposefully collected to support specific, but independent, validation arguments. In his chapter appearing in Wainer and Braun's 1988 text titled *Test Validity*, Messick (1988) writes "…the heart of the unified view of [construct] validity is that appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable and that the unifying force is empirically grounded construct interpretation" (p. 35).

In the context of this study, the latent (theoretical) construct is Teacher Instructional Quality (TIQ). Conceptualization of TIQ is derived from social constructions defining the purpose of schooling, the role of teachers, the pre-defined determinations of what constitutes instructional quality and best practices, and the learning outcomes identified for students. In keeping with the language of the *Standards* (1999, 2014), the 'test' is the evaluation system (AERA et al., 1999, 2014). The components parts of the system are expressed in terms of independent scales of measure which, when combined, allow an overall TIQ scale to be constructed. Teacher scores are placed along this scale for the purpose of making inferences about instructional quality. Thus, examination of validity involves evaluating the system's component structures aligned to their latent constructs, the scales as suitable representations of those constructs,

the measures employed as evidencing placement along those scales, and the consequences associated with interpreting and acting upon scores.

The question of consequence. An additional consideration in the examination of the validity evidence associated with teacher evaluation systems concerns the consequential aspects of the system's implementation and the impact it has on participants (Kane, 2001; Kimball & Milanowski, 2009; AERA et al., 1999, 2014; Amrein-Beardsley, 2008, 2009). By design, teacher evaluation systems are intended to inform on the quality and practices of teachers and teaching in the context of schooling and learning. As such, evaluation systems reflect the professional activities of individuals. By extension, systems that impact individuals may have both intended and unintended consequences. It seems reasonable, then, that validity examinations incorporate consequential dimensions of the activity.

However, this view is not without debate. Messick (1989a, 1989b, 1998), Shepard (1997), Linn (2008) and many other theorists argue strongly that examination of consequences must be a component of any validation study. Linn (2008) states:

> Validity is the most fundamental consideration in the evaluation of the appropriateness of claims about uses, and interpretations of educational assessment results. Although casual discussions often refer to the validity of an assessment, it should be understood that it is the uses, interpretations, and claims about assessment results that are validated. Evidence may support the conclusion that a particular use of assessment results has good validity. That same assessment may produce results, however, that have little or no validity when interpreted or used in a different way. (p. 1)

In contrast Popham (1997) and Mehrens (1997) argue that examining the consequential applications of test (evaluation) scores is important, but that it is not a central characteristic of construct validity. Popham argues that inferences placed on a test score can be highly aligned to the construct for which the originating test instrument was

constructed to measure without having to account for someone's proclivity to misinterpret or misapply the score after the fact. Mehrens (1997) extends this argument by contending inclusion of consequential effects as part of construct validity only acts to further obfuscate the fidelity of the score-construct relationship. He argues for an even greater distinction between various types of validity (criterion, content, predictive, concurrent, etc.) in order to clarify the methods and evidence needed in a well-conceived validation study. Consequential concerns, Mehrens argues, should be a separate and distinct form of inquiry. However, both theorists agree that examination of consequential applications of test scores is an important attribute that should be conducted when investigating systems that utilize test scores to make policy decisions.

This author feels that exploration of the consequential aspects of teacher evaluations system is a critical component of the validity examination process. Recalling the preceding discussion regarding theoretical frameworks, teacher evaluation as a generic process is embedded in a hierarchy of social constructs and intent: the role of teachers in the learning process, the intended outcomes of classroom instruction, articulation of student outcomes, and the purpose of school in society. Each of these (and others) provides the foundation upon which the efficacy of the evaluation process is determined. Arguable, examining intended and unintended outcomes forms a key perspective on the process.

With this in mind, examining consequential validity involves comparing the intended purpose of the system to the intended and unintended impact it has on stakeholders (Kimball & Milanowski, 2009; Kane, 2001; Messick, 1998; AERA et al., 1999, 2014; Amrein-Beardsley & Collins, 2012; Collins, 2012). As such, consequential

investigations focus both on the correspondence between stated and intended goals/objectives and on the social-emotional (interpersonal) aspects of the systems impact. That is, a system intending to measure the efficacy of individuals necessarily produces judgments on those individuals which may directly impact their personal and professional well-being. Conceivably, these impacts may be either positive or negative and may extend beyond the individual with respect to collegial relationships and organizational functioning. In this way, consequential evidence becomes part of the process of examining the latent TIQ construct and is a primary component of construct validity.

Situational Context: Local Contextual Setting

District Demographics

The information used in this study comes from a moderately large public school district situated within a middle income residential suburb in the greater Phoenix metropolitan area. The district enrolls approximately 24,500 K-12 students across 20 elementary and 4 high schools and employs approximately 1,230 classroom teachers. Approximately 5% of students are classified as English Language Learners (ELLs), 12% receive some type of special education services, and 52% participate in the Federal National School Lunch Program (NSLP).[1] Eleven of the district's elementary schools qualify for Title I funding from the U.S. Department of Education based on local community poverty rates.

[1] NSLP is a means-tested program based on family income that provides low-cost or free lunch to qualifying students.

District Academic Performance

Academic achievement and school/district accountability measures are based primarily on the state's standardized achievement test, the Arizona Instrument to Measure Standards (AIMS). Given annually in April, the AIMS test measures students' mastery of specific learning objectives on the state's adopted curriculum[2]. In 2012, AIMS school-level passing rates within the district ranged between 66% and 93% for reading (Mean = 80%) and 46% to 86% in mathematics (Mean = 67%).

Arizona also uses a statewide accountability system to assign overall achievement ratings to each school. These ratings are based on current and longitudinal AIMS measures. The system assigns performance ratings using report card-like A-F descriptive grades. In 2012, four district schools (17%) attained an A, 14 schools (58%) received a B, and six schools (25%) were rated as C. The district did not receive any D or F ratings for 2011-12.

The Researcher

The researcher is currently employed within the district as the Director of Research and Accountability and is charged with recommending and developing methodological approaches to measuring teacher instructional effectiveness for use within the organization's teacher evaluation framework. This position is responsible for ensuring that the methodological approach meets technical and inferential quality requirements and serves the purpose of informing professional growth and accountability. In addition, this position provides technical reference to central office decision makers

[2] High school students are provided the opportunity to take (and pass) the AIMS test in the fall and spring of each academic year in order to qualify for graduation.

and contributes to policy discussions surrounding use and interpretation of the system's information. Finally, the researcher is responsible for designing the information systems supporting data reporting and analysis of the evaluation system's information sets.

Evaluation Context: The District's Teacher Evaluation System

Construction and Design Process

During the fall 2011, the LEA began the initial steps for conceptualizing, designing, and implementing a new teacher evaluation system that fell within the policy-directed guidelines established by the State Board of Education. LEA central office policy leaders directed formation of a Teacher Evaluation Committee to begin discussions concerning the system's design and implementation. The committee was composed of 12 individuals: five teachers (one each from high, middle, and elementary levels, special education, and an instructional growth coach), two principals (one elementary and one high school), four central office administrators, and one representative from the district's teacher education association

To implement the requirements outlined in state statute, committee activities began by comparing the district's current approach for evaluating teachers to the new legislative requirements. This identified adjustments needed to bring the district's system into compliance. Two primary areas of inquiry concerned updates/alternatives to the district's current use of the Danielson Framework for Teaching (FFT) and the availability and quality of quantitative measures of student achievement. Outreach to other LEAs in Arizona also provided additional information from which to discuss possible approaches and adjustments. Early in the deliberations, the committee also established an intranet site for the organization and dissemination of information, reports, and announcements. The

site was accessible to all district staff and contained minutes of committee meetings, resources, and related documents.

During winter/spring 2011-2012, committee members conducted faculty workshops at each of the district's 24 schools. The purpose was to communicate requirements of the new legislative policy, update stakeholders on committee deliberations, and present preliminary thinking on possible adjustments to the evaluation process. At the same time, district committee members attended numerous trainings and workshops offered by the Arizona Department of Education and participated in professional conferences at the county and state level discussing issues related to the new system.

In the fall 2012, additional information workshops were held at each of the district's four high schools. Faculty from each location's feeder elementary schools were invited to attend. During these meetings, additional technical details were presented regarding approaches for computing TIQ measures. Specifics regarding the methodological approaches for computing teacher effectiveness scores based on student achievement were provided as were computational approaches related for measuring professional practice based on classroom observations.

During the first set of site-based workshops held in winter 2011 and spring 2012, the committee conducted a faculty survey focused on gathering questions, comments, and concerns. A web-based feedback form consisting of a single open-ended invitation to submit feedback was placed on the district's intranet portal site. During site-based faculty workshops, committee members introduced the survey portal and encouraged all teachers to provide feedback. A total of 121 questions were subsequently submitted. In late spring

2012, committee members individually and collectively reviewed the submitted comments and questions and prepared responses. These responses were then published to the web portal site and notification was sent out to all staff.

Stated Purpose of the Teacher Evaluation System

The district's teacher evaluation system is premised on the stated purposes espoused by the State Board of Education's Task Force on Teacher Evaluation. That is, the LEA evaluation system is intended to enhance student learning by improving the professional practices of classroom teachers. This is accomplished by empowering observers (principals) to provide structured feedback using the 2011 version of Danielson's standards-based Framework for Teaching (FFT). This framework is intended to identify areas of instructional strength and deficiency. Mentoring, training, and capacity building are then targeted to these areas. In this way, instructional efficacy is improved resulting in higher levels of student learning.

Within this context, accountability measures are situated as a driving force. The accountability framework combines measures from observations of professional practice (PP) with measures of student achievement. As mentioned, for the PP component, evaluators observe classroom activities and assign rubric-based numerical ratings across a variety of constructs specified under the FFT framework. Resulting composite scores are interpreted to reflect overall levels of instructional quality along a continuous scale. Sub-group element scores are meant to inform on specific instructional capacity, behaviors, and skills. Comparative accountability measures manifest, in part, from the ability to compare, order, and rank teachers along the PP scale.

The student achievement accountability component is derived from direct measures of student academic growth based on value-added (VAM) residual gain score models. As described in greater detail in Chapter 3, these models estimate student achievement on the latest state achievement tests. The predicted values are compared against actual results and the difference (residual) is interpreted to be associated with instructional effects of classroom teaching. Individual student residuals are aggregated to the classroom level and assigned to the presiding teacher. As with the PP scores, achievement measures manifest as accountability measures from their ability to compare, order, and rank teachers along a continuum.

Training for both principals (evaluators) and teachers (recipients) was embedded as part of the evaluation process. In this regard, teachers were provided a series of on-site workshops in the FFT framework during spring/fall 2012. These workshops were conducted by site principals and central office staff from the Human Resources department. In addition, principals received ongoing in-district training during monthly scheduled administrator meetings. At each meeting, approximately one hour was devoted to evaluation issues and practices. Finally, all site administrators were required to complete evaluator training using the Danielson Group's (Teachscape) online, web-based, training tools. In order to qualify, each administrator had to successfully pass the training's summative evaluation assessment.

Constructing a Composite Measure of Teacher Instructional Quality (TIQ)

The policy-directed teacher evaluation system requires that (at minimum) two measures of teacher instructional quality (TIQ) be utilized. The first relates to a measure of professional practice (PP) while the second reflects measures of student academic

progress based on test scores (TS; Ariz. Rev. Stat. §15-203A.38, 2010). Under this policy, a single measure of TIQ is constructed by combining the PP and TS components. The evaluation structure may be summarized as TIQ = b_1 (TS) + b_2 (PP), where b_1 & b_2 represent relative weights associated with each component. The weighted values are policy derived, not statistically estimated. As mentioned, legislative mandate requires that the PP component be weighted between 50% and 67% of a teacher's composite evaluation score. Similarly, the TS component must represent between 33% and 50% of the composite score. However, Arizona's evaluation framework did not specify the exact weights, leaving it as a policy decision for each of the state's public school districts. Herein, district policy leaders decided that the PP component would be weighted at 67% and the TS component at 33%. As mentioned, this decision was based on the initial lack of familiarity with developing statistical estimates of academic gain. It was felt that once the system was developed and implemented, adjustments to the weighting factors could be considered. Computational details of the scale constructions and associated composite scores are provided in Chapter 3.

As an outcome of the evaluation process, state policy requires that teachers be assigned an overall descriptive rating based on the combination of the PP and TS measures (Ariz. Rev. Stat. §15-203A.38, 2010). Specifically, every classroom teacher must be designated by one of the following performance labels: Ineffective, Developing, Effective, or Highly Effective. In addition, legislative policy requires school districts to annually report teacher evaluation ratings to the Arizona Department of Education. However, currently these reports are restricted to aggregate counts of teachers by performance category by school. For this district, performance designations are based on

each teacher's relative placement along the composite TIQ scale. This is represented in Figure 4.

Figure 4. Example Teacher Instructional Quality Scale.

Danielson Framework for Teaching (FFT)

For SY2012-13, the district adopted the 2011 version of the Danielson Framework for Teaching (FFT) as its primary measure of teacher professional practice (Danielson, 2011). As stated on the Danielson Group website

> The Framework for Teaching is a research-based set of components of instruction … grounded in a constructivist view of learning and teaching … The Framework may be used as the foundation of a school or district's mentoring, coaching, professional development, and teacher evaluation process, thus linking all those activities together and helping teachers become more thoughtful practitioners. (Danielson Group, 2013)

In previous years the district utilized earlier versions of the Danielson FFT. Changes for SY2012-13 included: (1) adoption of the latest series of FFT domain elements, (2) adoption of updated FFT rubric definitions and supporting resources, and (3) the requirement that all school evaluators complete and pass the Danielson Group's online (Teachscape) certification training and qualifying assessment.

The 2011 Danielson FFT incorporated revised language and resources aligned to the rubrics used in evaluating aspects of teacher professional practice. This included an

expanded set of critical attributes, explanations, and performance examples. These resources were added to assist evaluators in distinguishing levels of instructional performance (Danielson, 2011).

FFT domains. The FFT is a standards-based system defined by four primary domains of teaching responsibility: (Domain 1) Planning and Preparation; (Domain 2) Classroom Environment; (Domain 3) Instruction; and (Domain 4) Professional Responsibilities. Across these four domains, the FFT specifies 22 sub-elements. Figure 5 provides a listing of these elements.

Domain 1: Planning and Preparation	**Domain 2: Classroom Environment**
1a - Demonstrating Knowledge of Content and Pedagogy 1b - Demonstrating Knowledge of Students 1c - Setting Instructional Outcomes 1d - Demonstrating Knowledge of Resources 1e - Designing Coherent Instruction 1f - Designing Student Assessments	2a - Creating an Environment of Respect and Rapport 2b - Establishing a Culture for Learning 2c - Managing Classroom Procedures 2d - Managing Student Behavior 2e - Organizing Physical Space
Domain 3: Instruction	**Domain 4: Professional Responsibilities**
3a - Communicating With Students 3b - Using Questioning and Discussion Techniques 3c - Engaging Students in Learning 3d - Using Assessment in Instruction 3e - Demonstrating Flexibility and Responsiveness	4a - Reflecting on Teaching 4b - Maintaining Accurate Records 4c - Communicating with Families 4d - Participating in a Professional Community 4e - Growing and Developing Professionally 4f - Showing Professionalism

Figure 5. Danielson's four primary domains and component elements.

During the evaluation process, evaluators assign performance ratings to each sub-element within the system. Four performances categories are defined for each element. These are specified as *Unsatisfactory, Basic, Proficient,* and *Distinguished.* Under each category, supporting documentation provides evaluators with performance level descriptors, examples, and descriptions of critical attributes. An example of these descriptors is provided in Appendix A for Domain 1 (Planning and Preparation), sub-element 1a (Demonstrating Knowledge of Content and Pedagogy).

Evaluating professional practice. In the district, principals and assistant principals are the primary evaluators of classroom teachers. Classroom teachers are divided into two groups: continuing and non-continuing. Continuing teachers are teachers that have been employed in the district for more than three years. Non-continuing teachers have been employed for three years or less. District evaluation policy directs that all non-continuing teachers receive formal evaluations at least twice per school year (fall/spring). Continuing teachers receive formal evaluations at least once per school year.

Formal evaluations require assessment on all 22 FFT elements and include evaluator observation of a lesson lasting approximately 45 minutes to an hour. In contrast, informal observations occur throughout the year, are shorter in length, and may reflect selected subcomponents of the FFT elements. Information obtained from informal observations may be factored into determinations of the final year-end element ratings.

During formal evaluations, evaluators observe teachers conducting a full lesson lasting between 45 minutes to one hour. During this time, evaluators script all events and activities taking place. Scripting provides the raw observational data used to evaluate elements contained in FFT Domains 2 and 3, Classroom Environment and Instruction,

respectively. Data related to Domains 1 and 4 are gathered from artifacts, materials, and other representations obtained from the teacher. The process of assigning element ratings involves evaluating the quality of the gathered evidence based on each element's rubric criteria and descriptions.

The Professional Practice Evaluation Scale. As described above, teachers are evaluated on 22 performance elements across four domains. For each element, evaluators assign one of four performance ratings (*Unsatisfactory, Basic, Proficient,* and *Distinguished*). In order to utilize the PP ratings within the policy-directed evaluation framework, these ratings must be expressed in terms of a quantitative scale in order to: (1) allow for aggregation into an overall PP composite score, and (2) permit combination with quantitative measures of student academic progress (test scores).

To facilitate this, the district adopted a four point numerical scale to represent each element's performance rating levels. For each element, the lowest score is zero and the highest is three: *Unsatisfactory = 0, Basic = 1*, *Proficient = 2*, and *Distinguished = 3*. Since there are 22 elements, possible values for the overall composite PP score range from 0 to 66. Possible subscale scores depend on the number of individual elements within each domain. A summary of the PP composite scores by domain is presented in Figure 6.

Domain	Number of Elements	Possible Points Per Element	Possible Points
1: Planning and Preparation	6	3	18
2: Classroom Environment	5	3	15
3: Instruction	5	3	15
4: Professional Responsibilities	6	3	18
Total PP Points:	22		66

Figure 6. Total point structure for Danielson's four primary domains.

Since the number of elements differs across domains, the composite PP score does not reflect an equal weighting of content. However, as part of the district's implementation policy, the Teacher Evaluation Committee felt each domain should contribute equally to a teacher's overall composite performance score. In this regard, computational adjustments are made to reflect equal representation of each sub-domain.[3]

Measures of Academic Progress (Test Scores)

As discussed, the test score component of the evaluation system is based on value-added residual gain models of achievement. Under this framework, statistical achievement models are estimated to obtain predicted values on the latest state standardized test. Residual estimates are computed for each student and median residuals are aggregated to the classroom level. Since the achievement measures are represented as scale scores generated from the use of item response theory during the scoring process, the residual values are also expressed in terms of scale scores.

[3] Subscale scores for Domains 1 & 4 are weighted by a factor of -0.083 while scores for domains 2 & 3 are weighted by +1.00.

In this context, residual scores represent deviations from predicted achievement and are assumed to result from instructional effects: the greater the deviation, the greater the instructional effect. Positive values indicate achievement above expected levels (instructional gain) while negative deviations indicate below expected performance (instructional loss).

Final Composite Evaluation Scores

The primary computational task of computing the final TIQ score is to combine the measures of PP and TS into a single composite measure. However, the PP and TS measures are represented by different scales that cannot be combined without some type of transformation. To facilitate this, each teacher's final PP score is expressed as a percentage of total possible PP points. Similarly, each teacher's median classroom achievement residual value is expressed in terms of its percentile location across the range of all classroom residual scores. The two transformed measures are combined using the 33% and 67% weighting factors to arrive at a final TIQ score. Because the final TIQ represents a weighted average of the two measures, the TIQ score range is bounded approximately by 0 and 100.

Prior Action Research Cycles

In preparation for this study, a number of field activities were undertaken to help inform and shape the research design presented herein. During fall 2012 and early spring 2013, efforts were undertaken to design and test a variety of data collection methods including the following: development of stakeholder interview and observation protocols; specification of (statistical) hierarchical value-added models of student academic growth;

approaches to exploratory and confirmatory factor analytic designs related to evaluator ratings of teachers; construction of supporting databases aligned to achievement and teacher professional practice; and writing/testing of associated computer code governing data processing and analytic procedures. Throughout the process, members of the Teacher Evaluation Committee and selected LEA policy makers were consulted and informed.

In addition, the design and field testing of the qualitative interview and observation protocols was conducted with selected stakeholders including teachers, evaluators, and policy makers. For the formal study, these protocols served as data collection instruments to explore content representations and consequential aspects of the evaluation system as well as to document stakeholder input (voice) into policy level discussions. Limited scale digital (photo) ethnography was also undertaken during a number of the observation/interview sessions as a possible additional form of artifact data collection in the research design.

Testing of value-added model specifications was accomplished using SY2011-12 and prior year achievement information. Results of this activity lead to adoption of the statistical procedures for generating academic growth measures used in constructing a teacher's overall TIQ score. Similarly, factor analytic methods were applied to early evaluator rating data sets to help define approaches for exploring content and scale-reliability evidence. Additional statistical methods (correlation, ANOVA) were applied to the available measures to clarify approaches for evaluating specific research questions and alternative dimensions of program (criterion) validity.

Finally, beginning in August 2012, this researcher began maintaining a reflective journal of ongoing activities, discussions, and experiences related to designing and implementing the LEA's teacher evaluation system. This included reflections on meeting discussions, communications, workshops, interviews, and field observation, as well as personal reflections related to the technical details, data issues, and events happening in the larger policy setting at the state and national levels. Additional details regarding the fall 2012 Action Research (AR) Cycle are provided below followed by a summary of reflections and findings that informed this study's final research design.

Protocol Field Testing

During the fall 2012 AR cycle, interviews were conducted with the following stakeholder groups: central office (3), principals (3), and classroom teachers (4). In addition, observational protocols were field tested during three formal classroom evaluation sessions. During two observation sessions, digital (photo) ethnographic methods were explored as a potential additional approach to data collection. Group interviews/discussions were also conducted at four school sites using a small group question/answer format relating to technical aspects of the LEAs evaluation system. Journaling activities were utilized during various evaluation committee meetings, stakeholder discussions, conference sessions, and evaluator training sessions each related to the teacher evaluation system. Finally, a stakeholder questionnaire was constructed for use at the end of SY2012-13 to obtain feedback on the content representation of the rated PP categories and the evaluation process in general. All of the qualitative information were analyzed using procedures outlined by Corbin and Strauss (2008) and Saldana

(2009), progressing from in-vivo to axial and thematic coding. Throughout the process, data were reviewed leading to protocol adjustments in subsequent events.

Action Research Cycle Reflections

In addition to working through analytic approaches, a primary focus of the fall 2012 AR Cycle was to construct and field test qualitative data collection protocols used to support the various research questions posed for this study. Below is a brief discussion of findings related to classroom observation of evaluator activities, stakeholder interviews, and journal reflections.

Evaluator observations. The time required for an evaluator (principal) to complete one full formal classroom evaluation ranged between 45 minutes and 1 hour. During this time, evaluators are expected to script real-time lesson activities as well as collect/note document artifacts and other evidences of instructional practice as defined within the Danielson FFT framework. Observation of multiple classroom-based evaluation sessions highlighted a number of potential threats to validity involving evaluator activities. These factors were coded as Stamina, Skill, and Method. Analysis suggested that evaluators get tired and the level of stamina required to record classroom observations varied across persons. This brings into question the integrity/quality of the evaluator's scripting activity as the observation period progresses. In addition, the skills with which evaluators were able to capture large amounts of scripting data and other evidence varied by person. Some evaluators were able to script (type) and observe at the same time while others type slowly looking directly at the keyboard. Again, this raises questions of data quality, integrity, depth, and coverage. Variability on these factors might negatively impact rater reliability and inferential validity. Finally, the methods that

evaluators use to capture classroom observational data varied considerably. Some evaluators used video cameras in addition to digital scripting. Others utilized additional resources such as rubric sheets and artifact lists. Still others brought memo pads to write down thoughts, questions, and notes outside of the formal scripting activity. Engaging in these early observational activities helped define some of the protocol categories used during the formal research process.

Stakeholder interviews. Analysis of AR-cycle interview data suggested substantive differences between stakeholders. Some policy-level (central office) persons tended to see the teacher evaluation system as necessary for improving classroom instruction and trusted its validity more than teachers and principals. These individuals also seemed to regard the quantitative ratings as legitimate (valid) measures to distinguish levels of instructional quality. Interestingly, these early findings were not fully consistent with follow-up interviews conducted during the formal validation study discussed herein.

In contrast, teachers tended to view the process more in accountability terms and were more critical of the system's ability to facilitate improved instructional practice. Teachers also seemed to question evaluator's ability to capture instructional quality due to the limited number of observational sessions. While policy-level persons believed that evaluator training and certification increased interrater reliability, teachers were less accepting of this. Both teachers and principals noted that it was difficult to be completely objective in the evaluation process. Principals reflected on the burden of conducting one or more formal hour-long evaluations for every teacher in the building. This was consistent with findings in the subsequent formal validation study.

In addition, policy-level persons tended to view test scores as a legitimate component of the evaluation system. In contrast, principals and teachers tended to question whether reading and math scores on standardized tests captured the essence of quality instruction. Teachers viewed test scores as being less valid than classroom observations for an indicator of instructional quality. In addition, they believed that highly effective teachers impact students in many more dimensions than test scores: creativity, curiosity, appreciation, social skills, etc.

Learnings obtained from the field interviews suggested the need to incorporate two major components into the validity examination. The first concerned the need to clearly articulate the underlying construct of Teacher Instructional Quality (TIQ) for each stakeholder group. If each stakeholder group internalizes different meanings, implementing a "one size fits all" evaluation approach would raise substantive validity concerns. Arguably, this needed to be a component of the formal research study. That is, obtaining and documenting how various stakeholders conceptualized quality teaching should be contrasted with the system's explicit measure of instructional efficacy.

Second, if a policy-driven approach to measuring instructional quality mandates test scores as a primary component, teachers who minimize the importance of this measure may become marginalized. Again, arguably, this could manifest in reduced levels of trust in the system and limit the ability of the system to affect changes in professional practice, a core tenet of the policy-directed evaluation construct. Regardless, reflections from the field test activities suggested that this facet be incorporated into the validity examination.

Early research journal reflections. Early journal notes and reflections suggested teachers lacked a clear understanding of the test score component use by the LEA evaluation system. Despite numerous communications, teachers attending various workshops and meetings seemed unable to clearly describe the methods used to compute value-added achievement measures of instructional effectiveness. This suggested that LEA outreach efforts may not have been sufficient and require additional attention. In addition, there seemed to be considerable variance in the skills and abilities of evaluators despite completion of the Danielson online certification exam. The quality of evaluator's observational techniques seemed to vary considerably. In addition, some principals expressed their willingness to adjust instructional ratings based on considerations other than those defined by the Danielson FFT rubric and/or the collected evidence. For example, one principal reflected a reluctance to assign a large number of low element (domain) scores out of concern that it might negatively impact teacher morale and harm future mentoring opportunities. For the evaluation study, this suggested that interview and other data collection protocols should include questions related to scoring fidelity.

Chapter 1 Conclusion

The discussion presented in Chapter 1 provides the following perspectives on the task of implementing and examining the efficacy of policy-directed teacher evaluation systems. First, it is argued that implementation of high stakes evaluation systems require organizations to gather validity evidences to support consequential decision making. Second, the conduct of such systems takes place in a broader social-political context that affects aspects of system efficacy. As such, these contexts need to be accommodated in the research design. Third, a variety of theoretical frameworks inform on both the design

and conduct of the validity examination activity. Fourth, validity theory directs use of multiple sources and types of evidence as well as multiple methodological perspectives. Finally, methodological perspectives (what Creswell, 2009, calls world views) guide the mix of analytic approaches used to collect and interpret data aligned to the research questions of interest. Figure 7 provides a visual representation of these components.

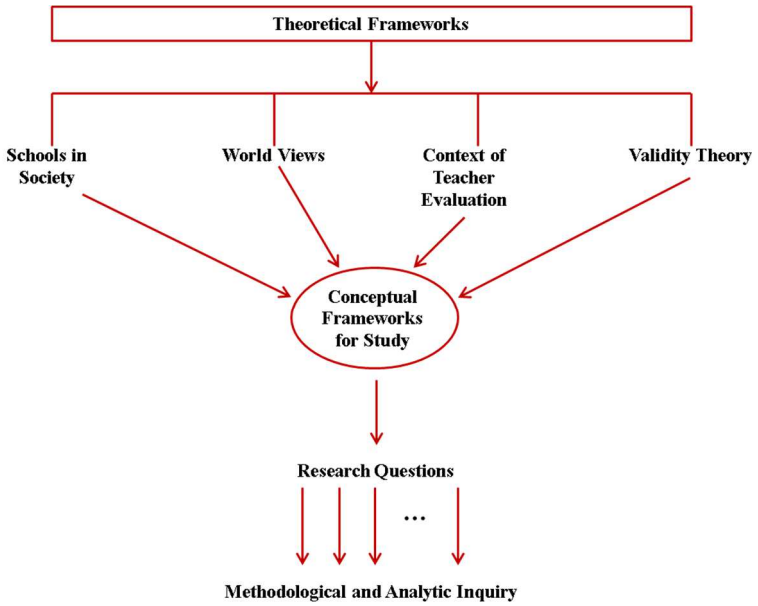


Figure 7. Components of the study's theoretical framework.

Preface on Literature Review

Arguably, the perspectives discussed above inform and guide the structure and scope of a literature review appropriate for this study. That is, contributions and understandings from the published literature may be grouped into the following broad topics: schools and teaching in a societal context, validity theory, validity related to

teacher evaluation systems, and measures of teacher effectiveness. The ordering of these topics is purposeful, reinforcing the hierarchal arrangement of the study beginning with the social context under which the entire process is situated and, arguably, warranted. This is followed by the theoretical context of validity and validation, providing the reader a bridge between the larger social context and the conduct of an empirically-based validation study. Then a more detailed examination of published validity studies related to teacher evaluation systems is presented to provide context and comparison to similar efforts. Since the study conducted herein involves use of value-added models (VAM) of academic gain, a section is devoted to this topic. Here, some of the technical and methodological issues associated with VAM are explored, including their validity implications on teacher evaluation. Finally, it seems reasonable to place the evaluation process within the context of organizational change theory. This section emanates from debates on the role consequence and empowerment play in the validation construct and how they impact the functioning and efficacy of the evaluation process.

Organizing the literature in this way attempts to incorporate both a social constructionist's view of validity as well as a post positivist perspective of the study's technical aspects. Hopefully, this honors the interconnectedness of the evaluation system's implementation with intent, and places necessary emphasis on the theoretical constructs driving teacher evaluation in education. The overall structure of the literature review is presented in Figure 8.

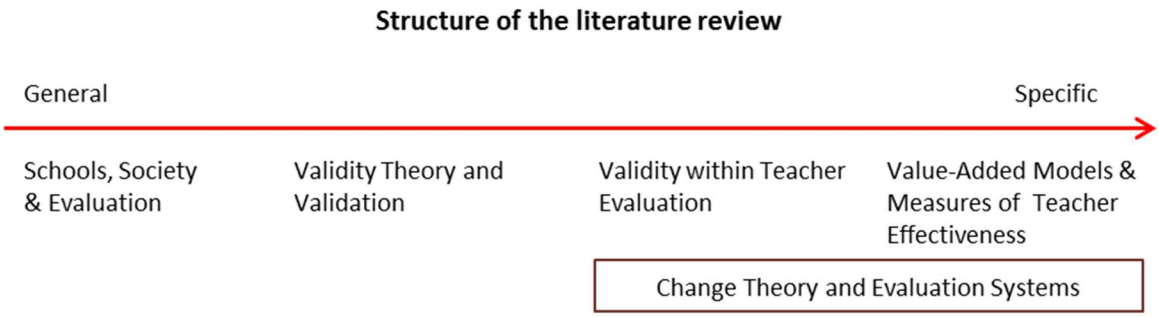


Figure 8. Organizational structure of the literature review.

Chapter 2: Literature Review

This literature review is organized along the theoretical constructs proposed for this paper: Schooling/Teaching in Society, Validity Theory, Validity of Teacher Evaluation Systems, Value-Added Models/Measures of Teacher Effectiveness, and Change Theory.

Schools and Society

It is argued herein that to evaluate the validity of a policy-directed teacher evaluation system, it must be recognized that such systems are social constructions embedded in a larger context. Indeed, the basic purpose of evaluation is grounded in a presumptive social policy context that evaluation is a needed and necessary activity to achieve a larger goal (Task Force on Teacher and Principal Evaluations, 2011; Berliner & Biddle, 1995; Berliner, 2005, 2009; Darling-Hammond, 1997; McGee-Banks & Banks, 1997). Darling-Hammond (1997) attributes the current socio-political connection between school quality and the broader national-social context to the 1983 publication of *A Nation at Risk* (National Commission on Excellence in Education, 1983). She also writes that current educational reform efforts "… are based in large part on the realization that education's importance to individuals and to society is increasing dramatically as we move from a manufacturing economy to a technologically based information economy" (p. 152). Interestingly, this observation is similar to that of Dewey's observation made at the turn of the twentieth century that schools and teaching were in need of fundamental reform due to the transition of American society from rural agriculture to an urban manufacturing society (Dewey, 1900). The difference is that the contemporary education reform is routed in an overt political agenda.

Regarding this socio-political connection, Berliner and Biddle (1995) are not so kind, writing "…*A Nation at Risk* charged that American students never excelled in international comparisons of student achievement and that this failure reflected systematic weakness in our school programs and lack of talent and motivation among American educators" (p. 3). A charge that Berliner and Biddle strongly contest as lacking compelling evidence but concede has been driving political reform efforts ever since.

Within education circles, the perspective of *why do we evaluate teachers* might be viewed by some as being easily articulated. For example, Charlotte Danielson is well-known as the developer of the Danielson Framework for Teaching (FFT), used in over 200 school districts around the country including the district represented herein (Milanowski, 2011). In an article appearing in *Educational Leadership,* Danielson asks "Why do we evaluate teachers?" (Danielson, 2010, p. 36). Her response is that "… public schools are public institutions; they take public money, and the public has a right to expect high-quality teaching" (p. 36). She follows this by stating that educational leaders must be able to tell the district Governing Board "everyone who teaches here is good and here is how I know" (p. 36). That is, her perspective of evaluation is centered in ethically-based accountability and professional expectation. A second reason she gives is for the professional development of teachers (i.e., capacity building). She remarks that "… teaching is so hard that we can always improve" (p. 37). However, neither of these explanations situates teacher evaluation in the broader social context nor examines the purpose of schooling and its links to society.

Berliner (2005) calls the current movement toward high stakes teacher evaluation systems political spectacle, implicating the effort as a misguided political agenda (p.

205). Indeed, Berliner argues that most of America's teachers are already highly qualified as evidenced from a variety of sources such scores on the National Assessment of Educational Progress (NAEP) assessments, international assessments of reading, mathematics, and science, teacher's attainment of college level degrees, teaching certification requirements, and specialized university training. Berliner's argument situates high stakes teacher evaluation as a political response to the perceived inadequacy of America's educational system that, left unattended, would continue to degrade American society and economic competiveness. Berliner argues that what is needed for America's schools is not a renewed focus on teacher evaluation but a solution to societal problems of poverty, inequitable wealth distribution, inadequate community services, and school systems in which educational quality is determined by community income (Berliner, 2009). In this way, the *why* of high stakes teacher evaluation is political and economic ignorance about the causal relationship between social well-being, economy, schools, and teachers.

Arguably, the theoretical construct of teacher evaluation only has meaning if placed within a social and/or policy context (or both). As the reader may recall, this was discussed in some detail in Chapter One under both the theoretical framework and as the environment underlying Arizona's legislative action (Task Force on Teacher and Principal Evaluations, 2011). For this reason, some understanding of the positionality within which the evaluation activity takes place is believed to be important for the conduct of this validation study.

To do so, a brief review of selected literature is presented, beginning with a discussion from Dewey (1900) regarding the role of schools and schooling in American

society. This discussion attempts to place the professional practice of teaching, and therefore the study and evaluation, into a larger context originating with the evolving nature of American education back to the early 20th century. This is followed by discussions from McGee-Banks and Banks (1997) about reforming schools in a democratic society and positions the purpose of teaching (and thus teacher evaluation) in a more contemporary social context. Finally, an article by Thomas Good (1999) is reviewed. Titled "The Purpose of Schooling in America," Good argues that the purpose of schools, teaching, and learning is not limited to the core tested subjects (i.e. those used in valued-added measures of teacher effectiveness), but rather concern instilling more abstract attitudes and behaviors coincident with our place within a global interconnected community. This brief review is not meant to be exhaustive, but to provide perspective beyond the immediacy of reviewing evaluation scores and measures.

In his work, titled *The School and Society*, Dewey (1900) reflects on the role and placement of schools in society. He states "… we are apt to look at the school from an individualistic standpoint, as something between teacher and pupil, or between teacher and parent…" (p. 21). However, he suggests that our view of schools and schooling should not be so narrowly focused. Rather, schools should be viewed within a larger perspective that incorporates the needs of community, changes in society, and the expansion of democracy.

To that end, Dewey suggests that schools should not be looked at solely from the point of view of the teacher-child relation. Rather, he reflects that education is inherently a community construct where children grow to become members of the larger society.

Hence, he suggests that individualism and socialism are not competing concepts when it comes to public education. Dewey states that,

> … we are apt to look at the school from an individualistic standpoint. … Yet the range of outlook needs to be enlarged. What the best and wisest parent wants for his own child that must the community want for all of its children. Any other ideal for our schools is narrow and unlovely; acted upon, it destroys our democracy. All that society has accomplished for itself is put, through the agency of the school, at the disposal of its future members. (p. 21)

In these comments, Dewey is arguing for a larger contextual role of schools in society, one that encompasses more than rote learning in favor of instilling broader social, ethical, and moral attributes that may accompany the child into adulthood. From this perspective, Dewey argues that the act of schooling should empower students with individual creativity, curiosity, and potential. Citing examples of students learning through experimentation, he characterizes the process of exploratory learning as critical for empowering individuals to function in an evolving industrial society that requires cooperation and values interpersonal connections. He states that "…a society is a number of people held together because they are working along common lines and with reference to common aims…" (p. 25) and argues schools must be organized to align with this purpose.

In this regard, Dewey is critical of school organization and learning structures that do not serve the larger community context. Indeed, he states that the "… tragic weakness of the present school is that it endeavors to prepare future members of the social order in a medium in which the conditions of the social spirit are eminently wanting" (p. 26). He suggests that when properly structured and implemented, the process of schooling "… has a chance to affiliate itself with life, to become the child's habitat, where he learns through directed living; instead of being only a place to learn lessons" (p. 27).

In describing how the common purpose of learning and community takes place, Dewey repeatedly uses examples of children engaging in lessons that are highly applied and experiential. For example, he describes students learning to work with wool and cotton to weave material. Through hands-on activities, the students become aware of the required raw materials, the physical characteristics of the medium with which they are working, the applications for which such materials are best suited, and the impact they have in society. Dewey argues that by doing applied lessons, students are provided opportunity to integrate individual learning with social connection (p. 30). In this way, Dewey is making comment on the nature of teaching, cognitive learning, and attributes of curriculum and instruction.

Based on his writings, one can only speculate on Dewey's reaction to contemporary approaches to teacher evaluation based standardized tests that (arguably) emphasize non-performance based knowledge and skills. His discussion of transformative teaching and curriculum suggest he might favor more performance-based measures of learning where the student is required to demonstrate not only basic knowledge but the connection of that knowledge to broader societal conditions and contributions. Interestingly, Dewey's reflections sound much like the dialog surrounding current transitions from formulaic standards to new *Common Core* criteria emphasizing problem solving, analysis, and the integration of learnings.

Dewey (1900) makes it a point to reflect that, for centuries, access to education was limited to a small segment of the population, reinforcing societal class and the concentration of political and economic power. However, he notes that this was partly due to the limits of existing technology and economic need. As the industrial revolution

began to take hold, evolving innovations gave rise to such things as paper, the printing press, mass communications, and expanded modes of transportation which collectively enlarged opportunities for commerce, new economic markets, and diverse sources of wealth. By 1900, Dewey contends, these innovations helped democratize the need for public schools stating "… knowledge is no longer an immobile solid; it has been liquefied" (p. 31). From this perspective, the need for a common educated populous rose out of shifting societal structures and the need for persons to engage in collective commerce. Dewey argues that the changing nature and interdependence of modern society requires similar innovations in schools and schooling. He states, "… our social life has undergone a thorough and radical change. If our education is to have any meaning for life it must pass through an equally complete transformation" (p. 33).

With regards to teaching and instruction, Dewey aligns his conceptualization of schools in society with instructional activities in the classroom. Here, he argues that it is not enough for children to learn facts and figures. Rather, students need to be engaged in meaningful activity and hands-on learning. He suggests that …

> … the ideal home would naturally have a workshop where a child could work out his constructive instincts. It would have a miniature laboratory in which his inquiries could be directed. The life of the child would extend out of doors to the garden, surrounding fields, and forests. (p. 37)

Dewey argues strongly that if we were to engage in instructional designs that emphasized experiential inquiry, students would be become more engaged and acquire knowledge and learning that benefit society. He comments that…

> … for the child simply to desire to cook an egg, and accordingly drop it in water for three minutes, and take it out when he is told, is not educated. But for the child to realize his own impulse by recognizing the facts, materials and conditions involved, and then to regulate his impulse through that recognition is educated. This is the difference, upon which I wish to insist… (p. 40)

It seems clear that Dewey believes knowledge should be facilitated through active learning where the outcome (goal) of instruction is not limited to facts and figures. Rather, becoming learned is placed in the context of broader social and community contexts. In this way, the goal of teaching is to instill connection between the individual and the society in which he or she lives. By extension, measures of learning which are limited to rote knowledge seem insufficient as measures of this construct. Thus, Dewey might argue that modern measures of instructional effectiveness require curricular evidences incorporating attributes of creatively, ethics, social connection, and responsibility in addition to the axioms of mathematics and science. The issue for any validation study is to incorporate that perspective in the theoretical construct and the evidences and methodologies employed.

From a more contemporary perspective, McGee-Banks and Banks (1997) present the thesis that contemporary education reform efforts have generally ignored pervasive issues of social inequalities based on race, income, and opportunity to learn. They contend that failure to address these issues threatens the goals and objectives of reform efforts because they ignore the interconnected nature of schools, economy, and social cohesion. The authors suggest that policy makers use school reform as metaphor to ensure future economic and social certainty in a dynamic and increasingly competitive world (p. 184). Yet, unequal social justice and educational opportunities obscure progress toward this outcome. In this way, McGee-Banks and Banks link success of school reform efforts, including the role of teacher evaluation, into a broader socio-political context.

McGee-Banks and Banks (1997) note that since the early 1980s, calls for school reform have resulted in increased attention on accountability systems, standards-based

curriculums, and increased performance expectations. These initiatives are meant to stem the perceived erosion of America's global economic standing and the perception that schools are not preparing students to compete in global economic markets. They contend that …

> …school reform advocates have called for schools to require students to master more facts, use standards to help identify appropriate educational outcomes, and hold teachers accountable for students learning to read, write, and compute. [However] … most school reform proposals focus on quality and rarely discuss equality and justice. (p. 184)

McGee-Banks and Banks (1997) argue that this narrow focus marginalizes large and growing segments of American society by failing to address critical inequities. These groups are defined along racial and economic characteristics. The authors state that "… addressing both quality and equity will result in schools that are reformed in ways that will help students from diverse groups to attain academic and social success" (p. 184). In stating this, they advance a perspective of schooling aligned with Dewey (1900). That is, the purpose of schooling (and by extension, teaching) is to empower individuals to fully participate in, and contribute to, the larger social community.

From this perspective, McGee-Banks and Banks (1997) argue that what is lacking in contemporary reform efforts are strategies of inclusion and opportunity which foster attitudes of community among traditionally disenfranchise populations. They write

> Students must be taught the knowledge, attitudes, and abilities needed to work with people from diverse groups to create civic, moral, and just communities that promote the common good. Diversity will be a salient characteristic of American society as the new century begins and the need to create a unified society in which diversity is honored will be paramount. (p.185)

One wonders if current policy-based definitions of teacher instructional quality incorporate these ideas of "civic, moral, and just communities" into the measures

assembled to infer instructional competency. Their omission might constitute an indictment of the basic teacher evaluation construct as depicted in social policy legislation (Task Force on Teacher and Principal Evaluations, 2011).

McGee-Banks and Banks (1997) advocate that education is situated within an interconnected social-economic-political context. Similarly, Dewey (1900) argues that as the social-economic nature of the community changes, so must the structure and focus of schooling. Each of the authors suggests that educators must play a role in this transformation. McGee-Banks and Banks (1997) distinguish between mainstream and transformative scholars where mainstream scholars "… create dominant ideas, paradigms, and theories …" devoid of action or advocacy (p. 187). In contrast, transformative scholars challenge social conventions and advocate for action. McGee-Banks and Banks note that "… transformative scholars and educators … have historically recognized the relationship between education, action, and a free society … (p. 187). They go on to state that "… educators must not only educate the mind, they must also educate the heart and create a sense of hope, commitment, and possibility among young people" (p. 188).

This position leads the authors to comment on the primary purpose of schools in America and the critical elements required of school reform efforts. They state that school reform needs to "… embody an understanding of a fundamental purpose for schooling: to help students to acquire the knowledge, skills, and values needed to actualize the democratic values stated in the Declaration of Independence, the Constitution, and the Bill of Rights" (McGee-Banks & Banks, 1997, p.188). To accomplish this, the authors reflect that schools need to promote civil discourse among

diverse groups which fosters understanding and compromise. The authors position classroom teachers and education scholars as catalysts for this innovation.

Clearly, the question being raised is whether contemporary teacher evaluation systems incorporate these values and goals as factors to be measured and used in judgments of instructional competence. Is the degree to which teachers infuse students with a sense of social justice, democratic ideals, civility, and moral standing weighted or even accounted for? Do standardized tests operationalize these concepts into numerical values that can be manipulated, ranked, and processed? Can a teacher with low standardized test scores be evaluated highly because his or her students meet these more challenging social-community standards? These are all questions of construct validity because they address the issue of construct definition.

Thomas Good's (1999) article titled "Purpose of Schooling in America" serves as an introduction to a special issue of *The Elementary School Journal* that focused on non-subject matter outcomes of schooling. Good starts out his discussion by questioning whether the traditional public schools and associated curriculum will be adequately responsive to the needs of 21st century society. His notes that while considerable attention has been paid to student's need for knowledge and skills in core subject areas, considerably less attention has been paid to the study of non-subject matter outcomes. Hence, the need for the special issue of the journal.

Like Dewey (1900) and McGee-Banks and Banks (1997), Good (1999) posits that the efficacy of schooling is not limited to expressions of narrowly defined (tested) subject mastery but encompasses larger socio-personal attitudes and behaviors reflective of a global interconnected community. Good formats the question as "… what else should

today's schools be committed to and accountable for in assisting students to become knowledgeable and productive citizens?" (p. 384).

Good (1999) notes that contributing authors address non-tested subjects such as art, music, and physical education, as well as more affective and personal attributes such as social, ethical, and civic attitudes, the ability for critical self-reflection, and motivations toward learning. He notes other topics in the issue concerning critical personal and social issues such as tolerance of diversity, issues of self-worth, environment, and attention to high-risk behaviors such as eating disorders, sex, drugs, and violence.

Good (1999) frames the special issue by providing an historical perspective as to why schools need to add these types of issues to their lexicon of important outcomes. He notes that "… there are many who, unlike Dewey, contend that the *public* school has outlived its utility" (p. 384). However, he goes on to note that at the same there are "…many scholars and citizens [who] currently advocate that the purpose of schooling must include more than simply student academic achievement, especially achievement as measured by exams designed for making international comparisons" (p. 384).

Good advances three factors that have led to claims of inadequate and unresponsive public schools. First, he notes the popular perception that the quality of public education has declined over the past 25 years. Despite evidence from scholars such as Berliner and Biddle (1995) that suggest this perspective to be unfounded, Good (1999) laments that policy makers still cling to it as rationale for contemporary reform initiatives. Good notes the sharp contrast in parent perception of public schools rating their own children's school extremely high but public schools in general much lower (p.

385). From this he concludes that the perception of ineffective schools is primarily media-induced.

A second factor is the popularity of vouchers, choice, and charter schools as solutions to the perceived decline in public schools. Good (1999) notes that "… the issue of public versus private control of schools … is an important part of the current debate about public education" (pp. 385-386). Advanced as a solution to quality issues, Good's inclusion of these factors underscores the individual-social conflict that Dewey talked about over 100 years ago with regard to the purpose of schools in a changing social context. Currently, large inequities along race and income lines in America reflect the lack of attention public policy has played in this regard. Good positions this as a factor in the debate about the role of schooling in society and the types of reforms needed to ensure an educated populous in through the 21st century.

Good's (1999) third factor is the lack of an articulated construct upon which reform initiatives are premised. He notes that "… reform advocates seldom state *how* charter schools and vouchers will transform the curriculum…" and that "… reformers appear to have little if any substantive agenda when it comes to delineating the purpose of educational reform" (p. 387). This lack of clarity allows pundits to proclaim efficacy of reform initiatives without having to provide clear substantiating theoretical foundations, evidences, or validations. This latter factor provides an important context to the school reform debate because it allows discussion to be sanctioned without authority. Because such efforts lack grounding in theory, validation is not possible. For example, the efficacy of vouchers may be proclaimed by data showing higher achievement of participating students compared to those that do not. However, if the reform construct is

premised on improving the educational outcomes of all students, such validation evidence is without merit.

Good's (1999) discussion frames many of the attributes involved in discussions concerning the purpose of schooling and the role education reform initiatives have in a broader social context. The main concept seems to be that without articulating the many factors impacting student outcomes, both from an individual and social perspective, debating the efficacy of school reform and instructional quality become difficult and inexact.

It is argued that each author's perspectives regarding schools in society directly informs on the construct validation environment surrounding contemporary policy-based teacher evaluation systems. Each author brings a unique perspective on the purpose of schooling and its contribution in a larger social context. However, all seem to place emphasis on learnings that go well beyond simple knowledge acquisition and skill building. Rather, each emphasize areas not traditionally tested by standardized testing programs nor aspects found in standardized rubric-based classroom observation (behavioral) systems. Thus, if the current state of curriculum and/or testing ignores or dismisses these broader social learnings (the arts, creativity, problem solving, ingenuity, resilience, moral standing, community connection, and personal contribution, etc.) then the validity of current policy-based teacher evaluation becomes suspect. This researcher finds this line of argument compelling. To the extent that the purposes and contributions ascribed to education are not reified by the empirical measures used to evaluate classroom teachers, the fundamental inferences derived from these measures become misleading. To this end, the conduct of any validation study requires critical examination

of the alignment between the broader (social) context of education and the details associated with developing and interpreting evidence of efficacy.

Validity

Historical Context of Validity Theory

Contemporary thinking in validity theory is set forth in the *Standards for Educational and Psychological Testing* (AERA et al., 1999, 2014). It establishes the professional and ethical criteria for using tests in a wide range of social, behavioral, and psychological settings (Kane, 2001; Linn, 2008; Gorin, 2007). This reference also provides the guidelines for evaluating the quality of tests, testing programs, and the application of test information to consequential decision making (Kane, 2001, 2010; Embretson, 2007).

The introduction to the 2014 edition of the *Standards* notes that "… educational and psychological testing and assessment are among the most important contributions of cognitive and behavioral science to our society, providing fundamental and significant sources of information about individuals and groups…" (AERA et al., 2014, p. 1). Indeed, the publication has gone through five earlier formulations beginning with the 1954 edition, then titled *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (APA, 1954). In each edition, the contemporary thinking on the nature of validity and its components, evidences, and applications has evolved (AERA et al., 1999, 2014; Shepard, 1993; Embretson, 2007; Mehrens, 1997; Kane, 2001).

In 1954, the *Technical Recommendations* defined validity as "… information [that] indicates to the test user the degree to which the test is capable of achieving certain aims…" (APA, p. 13). It differentiates the concept of validity according to the type of

information the user desired from the particular measure. Here, a single test may be used for different purposes (judgments) each requiring different types of validation studies. Indeed, at that time, validity was treated as being comprised of four separate independent concepts: content, predictive, concurrent, and construct. However, Shepard (1993) notes that between 1920 and 1950 correlation studies were the de facto standard for evaluating validity; test scores were compared to some secondary criterion measure in an attempt to answer the question "Does the test measure what it purports to measure?" (p. 410). Kane (2001) comments that at the time, little regard was given to the credibility of the criterion measure itself. Here he notes that "… the criterion-based model is quite reasonable and useful in many applied contexts, assuming that some suitable criterion measure is available… the trouble with the criterion model is the need for a well-defined and demonstrably valid criterion measure" (p. 320).

Publication of the 1954 edition expanded the test-criterion correlation view of validity to four conceptualizations (APA, 1954). Predictive validity compared scores to behaviors/outcomes measured at some subsequent time point. Concurrent validity was defined for studies needing to differentiate performance between groups on a similar criterion. The distinction between concurrent and predictive validity was primarily based on time, since both required comparison to some criterion measure. Content validity focused on evaluating how well a sample of items (behaviors) reflected the construct to which conclusions were to be drawn. And construct validity was framed as an inquiry of the theory underlying the test (pp. 213-214). Evidence for this latter category required two steps. First, the investigator makes predictions "… regarding the variation of scores

from person to person or occasion to occasion? Second, he gathers data to confirm these predictions" (APA, 1954, p. 214).

It is evident that much of the thinking at the time involved some use of correlational studies to assess validity, despite the different conceptualizations. Shepard's (1993) reflection regarding the strong historical reliance on predictive correlation prior to 1954 may be a possible explanation. Similarly, Angoff (1988) tracks the evolution of validity theory leading up to the 1954 publication noting numerous authors of the time depicting validity as the simple correlation among scores on a test with some other measure believe to inform on the same construct (p. 20). Here, Angoff notes the prevailing view of the time that a test could be deemed valid to any external measure to which it correlates, writing "… it was the use of validity in its predictive sense that dominated the scene" (p. 20).

Aside from its attempt to more clearly articulate and differentiate types of validity, the 1954 *Technical Recommendations* (APA, 1954) presented the first operational depiction of construct validity (Shepard, 1993; Angoff, 1988; Kane, 2001). Angoff (1988) reflects that this was no coincidence since Lee Cronbach was the chairman of the committee charged with developing the new testing standards. He, and colleague Paul Meehl, published (what is now considered) a seminal treatise on construct validity theory in 1955, a year after the *Technical Recommendations*. Titled *Construct Validity in Psychological Tests*, Cronbach and Meehl argued that all data aligned to and emerging from the testing activity reflect back on a single underlying trait (Cronbach & Meehl, 1955; Angoff, 1988; Kane, 2001). They define the term construct to be "… some postulated attribute of people, assumed to be reflected in test performance [where] in test

validation the attribute about which we make statements in interpreting a test is a construct" (p. 283).

In doing so, Cronbach and Meehl (1955) begin the shift from evidencing discrete types of validity to assembling multi-faceted profiles of evidence which collectively inform on a construct. They argue that positing the existence of a construct necessarily permits the investigator to generate numerous testable hypothesis. They term this as a *nomological net*, a network of associations or propositions in which the construct exists (p. 290). In turn, these testable hypotheses may be examined using multiple approaches. At the time of their writing, they offered examples such as examination of group differences, traditional correlation studies, factor analysis and related studies of internal structure, changes over time, and studies of process to account for variability in test scores. However, the clear intent was to argue for the assembly of many forms of evidence all informing on the latent construct. In this way, the previously defined predictive, concurrent, and content validity definitions are subsumed under a unifying conceptualization of *construct*. The authors write "… many types of evidences are relevant to construct validity, including content validity, interitem correlations, intertest correlations, test-criterion correlations, studies of stability over time, and stability under experimental intervention" (Cronbach & Meehl, 1955, p. 300). Angoff (1988) extends the contribution he feels Cronbach and Meehl were providing by adding that such investigations were inclusive of both quantitative and qualitative evidences (p. 26). Finally, Cronbach and Meehl (1955) make the strong statement that measures of construct validity cannot be reduced to a single simple coefficient. Rather, it is the

integration of many forms of evidence that either support or reject the presence of the hypothesized construct.

Despite the contribution made by Cronbach and Meehl (1955), the conceptualization of the four forms of validity outlined in the 1954 *Technical Recommendations* persisted through to the 1970s. Kane (2001) notes that in both the 1966 and 1974 editions of the *Standards,* the same four discrete categories of validity continued to be identified (p. 322). Kane (2001) writes "… in essence, then, validity was presented even well into the 1970s as involving several possible approaches" (p. 323). The problem occurs when investigators feel free to choose from a toolkit of evidences to support their validation studies, an approach Kane criticizes as opportunistic. Shepard (1993) reflects that, at the time, Cronbach and Meehl's (1955) argument was too demure and too ambitious while noting that they presented their perspective of construct validity as a weak sister to the other forms of validity (p. 416).

However, the seeds planted by Cronbach and Meehl (1955) were shaping the discourse and gradually began to influence theorists' thinking. Kane (2001) contends that by the end of the 1970s, predictive, concurrent, and content validities were increasingly being viewed as components of, and contributions to, construct validity (p. 324). Linn (2008) comments that preeminent theorists such as Lee Cronbach and Samuel Messick continued to form and promote their conceptualization of construct validity as a broad category encapsulating other forms of validity. Linn (2008) contends that "… Cronbach's [writings] laid the foundation for more of a unitary view of validity and the subordination of content-related and criterion-related evidence to a construct validity perspective" (p. 6).

Shepard (1993) comments that Lee Cronbach's chapter entitled "Test Validation" in the 1971 (second) edition of *Educational Measurement* "… was widely influential among students and specialists in educational measurement" (p. 423). In his writing, Cronbach (1971) opens the chapter with the historically familiar representation of validity by stating that "… validation is the process of examining the accuracy of a specific prediction of inference made from a test score" (p. 443). However, he contends that this is too narrow a perspective and clarifies that validation should be seen as a process of examining all the interpretations of a test, both descriptive and explanatory. Indeed, Cronbach argues that undertaking a test validation study is similar to the evaluation of any scientific theory (p. 443). By doing so, Cronbach breaks from the reliance on empirical, quantitative, correlation studies, and lays the ground work for other forms of evidence which will eventually be seen by theorists to include consequential and ethical aspects of test use, decision making, and impact. Indeed, Shepard (1993) contends that Messick's seminal treatise on validity and the centrality of consequential evidence published two decades later (Messick, 1989a) is based on Cronbach's 1971 thinking.

Despite the evolutionary (historical) importance of his 1971 chapter, Cronbach did not completely sever his thinking from earlier theories. This is evident in his discussion of test use, arguing there are two types: tests used for decision making and those used for description (p.445). In a table early in the chapter he continues to refer to "Types of Validity" and aligns "Content Validity" and "Construct Validity" to evidence intent on uncovering the soundness of descriptive interpretation (p. 446). In addition, he adds the category of Educational Importance as a new primary component for consideration. On the same table he separates criterion evidence (encapsulating notions of

predictive and concurrent) for the purpose of making selection and placement decisions, these latter being connected to criterion correlation studies.

So, while Cronbach's (1971) expansion of validity extends the definition, it does not yet definitively remap the historical perspective into a single unified entity under the banner Construct Validity. However, Linn (2008) comments that "… Cronbach's emphasis on the validation of interpretations of assessment results laid the foundation for more of a unitary view of validity and the subordination of content-related and criterion-related evidence to a construct validity perspective" (p. 6). Regardless, the new (unified) perspective of validity was slow in coming. Kane (2001) notes that the 1974 *Standards* (AERA, APA, & NCME, 1974) continued to represent validity as being partitioned across four discrete forms or types: predictive, concurrent, content, and construct. Up through this time, compartmentalizing validity into discrete types allowed investigators to select from a toolkit of approaches chosen based on the particular question or perspective of interest. In this context, one form of validity evidence served just as well as another with no requirement of perusing multiple vantage points. As mentioned, Kane saw this as highly opportunistic selection of alternative methodological approaches (p. 323). Indeed, test designers were free to offer evidences best reflective of the measurement activity.

Between 1974 and 1989, Kane (2001) reflects that thinking on validity had gradually evolved into three general principles. First, score validity needed be evaluated against well-defined theoretical frameworks requiring extended forms of evidence beyond those provided by criterion and content analysis. Second, there was a belief that theories require pre-specified interpretations against which evidence may be evaluated. And third, there was an expectation that the underlying theory itself was subject to

critical scrutiny and challenge (p. 324). Kane (2001) suggests that these realizations lead to the now accepted axiom that it is not the test or the test score that requires validation, but rather, it is the interpretation of these scores that requires evidentiary foundation (p. 324).

However, the acceptance of construct validity as a unified framework did not fully materialize until publication of Messick's 1989 discussion of validity in the third edition of *Educational Measurement* (Messick, 1989a). Here, Messick authored what might be considered the seminal treatise on validity theory of its time (Shepard, 1993). His eighty one page chapter, simply titled "Validity," synthesized and extended the thinking pioneered by earlier theorists into a coherent framework for understanding the concept of validity. His thesis argued with authority that validity should be seen as a single construct under which all other forms of validity are organized. In addition, he operationalized the inclusion of consequential outcomes and score use as not only a legitimate, but an essential, component of any type of validation study.

In his opening paragraph, Messick (1989a) writes "… Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment." He continues,

> Broadly speaking … validity is an inductive summary of both the existing evidence for, and the potential consequences of, score interpretation and use. Hence, what is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators – inferences about score meaning or interpretation and about the implications for action that the interpretation entails. (p. 13)

Beginning with these words, the remainder of the chapter, and (arguably) the field of measurement theory itself, focused on test score validation from a new and

comprehensive perspective. This said, Messick (1989a) himself recognized the evolving and changing nature of validity theory among leading theorists of the time, noting that the publication of the 1985 version of *Standards* (AERA et al., 1985) "… no longer [referred] to *types* of validity, but rather to categories of validity evidence called content-related, criterion-related, and construct-related evidence of validity" (p. 18). However, it was Messick's 1989 (1989a, 1989b) writings that solidified this perspective and directly impacted the next edition of the *Standards* released in 1999.

Messick (1989a) presented his unified conceptualization of validity as an intersection between two interconnected facets: justification of testing (evidence/consequence) and function (interpretation/use; p. 20). In his chapter "Validity" appearing in the 1989 edition of *Educational Measurement*, Messick (1989a) depicted the dimensions using a 2 x 2 matrix. Figure 9 shows how the author arranged the facets.

	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity + Relevance/Utility
Consequential Basis	Value Implications	Social Consequences

Figure 9. Messick's facets of validity.

Here, the evidentiary basis for both test interpretation and test use, is a single conceptualization of construct validity supported by additional evidence aligned to the specific context in which testing is taking place (relevance/utility). The consequential

foundation for test interpretation is evaluated within theoretical values and/or ideologies defining score interpretation. Finally, the intersection of consequence and test use is evaluated within a context of social values and outcomes.

This representation of validity is not without criticism. Shepard (1993, 1997) criticizes the way Messick expressed his theoretical framework as a validity matrix in which consequential aspects of test score use are depicted as a unique element (lower right corner of the matrix). Shepard's position is that this representation is misleading. While supporting Messick's unifying theory, she states that "…in my view, the matrix was a mistake …" (p. 5). She argues that such a representation leads theorists to view examination of the social consequences of test score use as an independent element subsumed under the larger validity construct. In doing so, it leads to fractured treatment of validity, something that the evolving dialog of validity has discouraged.

Shepard (1997) argues that:

> …facets [of validity] cannot be pulled apart and considered independently …, construct validity resides in all cells, … [and that] the temptation is too great to think that the traditional, scientific version of construct validity resides in the upper, left-hand cell and that consequences in the lower, right-hand cell are the business of moral philosophers and the politically correct. (p. 6)

However, this researcher's reading of Messick's chapter (1989a) tends to disagree with her criticism of the matrix representation, citing his extended discussion which argues for the need to move away from discordant examinations of *types* of validity and to a unified construct. Messick (1989a) writes "…what is needed is a way of cutting and combining validity evidence that forestalls undue reliance on selected forms of evidence …" and that "… these distinctions [depicted in the matrix] may seem fuzzy because they are not only interlinked, but overlapping" (p.20). His intent was to provide a visual representation to

bring clarity to the facets, but at the same time being clear that they are interconnected and overlapping.

In a companion publication occurring at the same time as the release of the 1989 edition of *Educational Measurement*, Messick (1989b) examines the idea of needing multiple sources of data/evidence when examining the validity construct of test scores. Here he clearly states that "…to validate an interpretive inference is to ascertain the extent to which multiple lines of evidence are consonant with the inference" (p. 5). In this way, Messick argues for incorporating all forms of analytic methods, both quantitative and qualitative, and any and all sources of information that serve to inform on warrants made based on test scores (*test* being any type of empirical measure).

Messick (1989b) also argues that validation is fundamentally a theory driven and data driven exercise. In so doing, he crosses the boundary between simply formulating a framework for operationally interpreting scores to a more objective, positivistic premise that validation is a formal process grounded in logic and probability (p. 6). In this way, validation as a mode of inquiry is inherently inductive. It is not a matter of yes/no, right/wrong, but a gradient along weak-to-strong evidence in which evidence is assembled from imperfect sources of measure. Hence, validation necessitates examination of both confirming and disconfirming evidence as well as an evaluation of evidentiary merit of the originating data source.

Consequential Validity

The importance of Messick's contribution to measurement science and validity theory cannot be understated. Arguable, his examination of the limitations for treating content, criterion, and construct validity as stand-alone constructs profoundly influenced

subsequent thinking in validity theory and the implementation of validation studies in applied settings. However, Messick's inclusion of consequential outcomes as a core tenant of validation is not without controversy.

In her article "The Centrality of Test Use and Consequences for Test Validity," Shepard (1997) examines the debate surrounding consequential validity and whether or not it should represent a central component of test score validation process. Reacting to criticism from Wiley (1991), Maguire, Hattie, and Haig (1994), and Popham (1997) that consequential validity detracts from the central focus of the validation process, Shepard argues that consequential aspects of test score use cannot be separated from the nomological net upon which validation studies are based. The term nomological net refers to Cronbach and Meehl's (1955) depiction of validity as being composed of interrelated, observable, components that are representations of the construct of interest. The process of assembling empirical validity evidence depends on the ability to articulate these elements and their association to the construct. Shepard writes "…antagonists and advocates have created a false impression that a new kind of validity was invented in 1989…" (p. 5). Instead, the author argues that "… consequences are a logical part of the evaluation of test use … [and] are not outside the underlying network of relationships that frame a validity investigation" (p. 5). Indeed, Shepard contends that examining consequential aspects of test use is essential in any situation except for the most limited of circumstances.

In contrast, Popham (1997) takes a critical perspective on the inclusion of consequence as a primary component of validity examination. He bases his criticism for incorporating consequential effects as a fundamental element of construct validity on its

disassociation from the primary alignment of test construction with test purpose. He argues is that social consequences of test score interpretation should be seen as an artifact separate from the primary conceptualization of construct validity. In this regard, Popham disagrees with Messick (1989a, 1998), Shepard (1997), and others who advocate for a unified conceptualization of construct validity inclusive of consequential impacts. Acknowledging that his position is counter to many respected theorists, he writes "indeed, one is opposing the heart of the batting order when tussling with folks such as Messick (1989[a]), Shepard (1993), and Moss (1995). But even heavy hitters occasionally strike out" (p. 9). His position is not that consequential impacts of score use are not important, only that such investigation are not the purview of validity studies.

Popham (1997) argues that the foundation of validity is evidence that inferences made from test scores are aligned to the latent construct upon which the test was constructed. He conceptually lays out an ordered association between the latent construct and the constructed measures. This alignment is visualized in Figure 10.

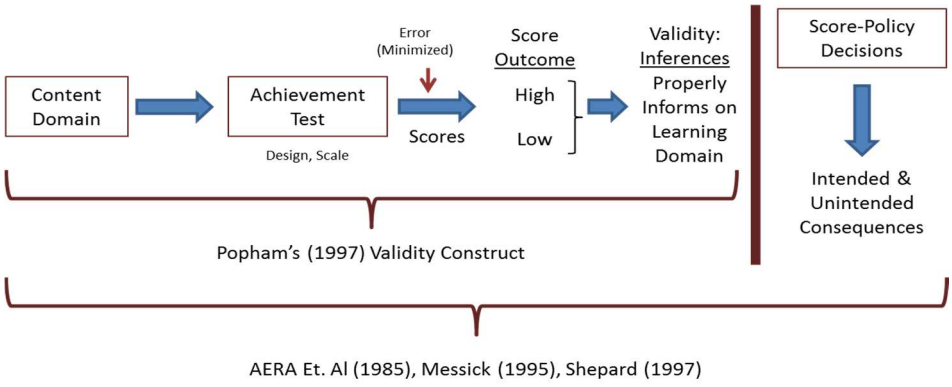


Figure 10. Comparison of validity construct from selected sources and authors. Note: the figure is this researcher's re-construction of Popham's line of reasoning embedded within other conceptualizations of validity theory.

In Popham's (1997) conceptualization, validation studies concern issues of test design, scale construction, domain representation, and characteristics of the scores that suitably inform on the measured attribute. Specifically, Popham states that "… validity refers to the accuracy of score-based inferences" (p. 11). He contends that this simple perspective brings clarity to the conduct of validity examinations, noting that "I have encountered relatively few American educators who understand that validity depends on the accuracy of score-based inferences and is not a property of a test itself" (p. 10). His contention is that embedding consequential aspects of score use distort this representation arguing that "… the clarity of validity-as-score-based-inferences is being threatened by those who would require the concept of validity to shoulder the theoretical baggage that … does not bear on the accuracy of score-based inferences" (p. 11). Implicit is his separation of a theoretical score-to-construct representation from a consequential score-to-decision pathway that Messick (1989a, 1998) first articulated.

To represent his position, Popham (1997) uses the example of a mathematics test where the purpose is to make placement decisions based on differential mastery of mathematical concepts. He begins by assuming that the construction of the test meets all of technical criteria established for such activities including contributions by content and measurement experts. He proposes that all learning constructs have been well articulated and every effort has been made to remove bias and factors causing misalignment or misspecification. The result is a test where

> … everyone who scrutinizes the test, and the definitions of the content domains it was created to represent, reaches the same conclusion. This is a test from which valid inferences can be drawn regarding the levels of … a student's mathematics achievement. (p. 11)

Thus, in this example, a student's score correctly identified his/her position along a continuum of learning within the subject.

Popham (1997) then extends his arguments by proposing a situation where an overzealous school board mandates that any girl not achieving at pre-specified levels be prevented from taking any type of music or art courses and that all boys scoring below a specified threshold be expelled. He states "… such absurd decisions, while deplorable, do not alter one whit the validity of the test-based inference about student's mathematics achievement. Those inferences are just as accurate as they were before the board made its ludicrous test-based decisions" (p. 12). As graphically represented above, Popham argues that the post hoc score-based decisions are outside of the measure's validity examination and its alignment to the desired construct.

This perspective clearly differs from perspectives advanced by Messick (1989a, 1998). Indeed, Popham (1997) questions why such esteemed theorists advance consequential elements as a primary component of construct validity. He writes "why is it, then, that first-rate measurement thinkers such as Messick and others have urged us to add these new consequential trappings …" (p. 12). His answer is that they intended to draw attention to the issue of misuse of test scores, unintended outcomes and effects, and harm caused to individuals as a result of unsound decisions. Popham (1997) concludes that examination of consequential score-based inferences should certainly be part of any context involving decision making. However, this activity should be seen as separate from the foundational tasks of establishing construct validity of test-based measures.

In response, Shepard (1997) reflects that Popham's position permits alignment between measures and their theoretical construct without attention to the consequential

impacts of score use. She argues that this is only possible under the most restrictive conditions where no inferential use of the score is proposed, a condition that only makes sense in a purely intellectual context. Shepard reflects that achievement tests may be viewed as status measures

> … only if they are not used to guide subsequent decisions. For example, when school districts are promised increased funding if they raise their test scores, then the real learning consequences of instructional efforts that follow are at the very center of the validity inquiry. (p. 7)

Her point is that validation cannot be restricted to a convenient set of circumstances identified by the user/researcher. In virtually every conceivable real-context case, test scores are generated for a purpose, with intent to interpret and take action, whatever action that may be. In doing so, warrants associated with that action require validation inclusive of the context and resulting consequences. Shepard (1997) acknowledges circumstances where consequence may not be part of the nomological net characterizing the testing context. Some uses, she says "… are purely descriptive…" (Shepard, 1997, p. 7). But, she adds, "… many test uses have explicit cause and effect relations as part of the test rational…" (p. 7). If the rational includes interpretation and action, then consequential evidence becomes part of the construct validation.

In his article titled "The Consequences of Consequential Validity*,"* Mehrens (1997) joins Popham (1997) in his critique of consequential outcomes as a legitimate element of construct validity. He reflects that "in the good old days" conceptualization of validity was operationally defined by the terms used to describe the investigation of interest. That is, content validity was seen as exploring the adequacy of a sample of items (behaviors) taken from an established behavioral domain. This might be expressed by a finite subset of math items selected from a larger set of items. Content evidences were

meant to convey the adequacy of the sample's representation to the population of all relevant items. Construct validity referred to the suitability of measures to serve as indirect indicators of some unobserved (unrepresented) behaviors. Here, the question concerns suitability of the observed scores to inform on a larger unobserved theoretical construct. Finally, criterion validity referred to the level of association between one behavioral measure and one or more external measures posited to reflect aspects of the same behavior. Knowledge of the first could then be used to conclude performance on the other. These types of criterion perspectives also subsumed concepts of concurrent and predictive validity. Regardless, Mehrens' reflection of the "good old days" is meant to highlight the clarity with which theorists interpreted alternative validity perspectives and each perspective's methodological approach.

Mehrens (1997) laments the trend toward expanding and unifying these conceptualizations of validity by stating:

> Some individuals have promoted combining all such terms into one suggesting that all validity is construct validity and that all evidence is evidence for (or against) construct validity. Such reductionist labeling blurs distinctions among types of inferences. Some people consider that progress. Some individuals are promoting even more "progress" by suggesting that the validity of an assessment should be evaluated based on the consequences. (p. 17)

On the latter point Mehrens (1997) expresses concern for adding consequential elements to validity investigations. In his view, the suitability of a measure to reflect the intended construct is independent of the effect any treatment decision (aka, consequence) has after the fact. In this way, Mehrens aligns with Popham's perspective but expresses even more reservation as to the deleterious effects.

Interestingly, Mehrens (1997) gives merit to the ongoing debate regarding consequential validity by acknowledging the positions of both Shepard (1997) and

Popham (1997). He credits Shepard (1997) for advancing examination of consequential outcomes as a worthwhile and important endeavor regardless of its place in validity theory. Similarly, he credits Popham for taking a critical stand against obfuscating the concept of score-based validity with consequential outcomes resulting solely from poor judgment or misguided policy decisions. To this Mehrens (1997) reflects that "… one can investigate the validity of the inference that a score is a reasonable indicator of the amount of a construct possessed independent of any specific use of the score" (p. 17). He concludes his discussions with a conservative recommendation,

> I suggest that the psychometric community *narrow* the use of the term validity rather than expand it. Let us reserve the term for determining the accuracy of inferences about (and understanding of) the characteristic being assessed, not the efficacy of actions following assessment. (Mehrens, 1997, p. 18)

Full Circle - Validity Components

The reflections and concerns offered by Mehrens and Popham were made in 1997, eight years after publication of Messick's 1989 chapter in *Educational Measurement*. However, the controversy regarding placement of consequential validity as a core component of a unified conceptualization of construct validity has continued to evolve. Mehrens and Popham's perspective that the movement away from a segmented view of validity and has only added unnecessary complexity has grown in stature among some measurement theorists (Brennan, 1998; Kane, 1992a, 2010; Kane & Case, 2004; Lissitz & Samuelsson, 2007). Indeed, the discussion over importance of consequence have given way to bigger questions of the unified construct itself.

Kane as early as 1992 reflected that test validation is a process of test-score interpretation and the goal is to "… support the plausibility of the corresponding interpretive argument with appropriate evidence" (1992a, p. 527). In this way, Kane

began viewing validation as an argument-based approach to validity (Kane, 1992b) and in later writings clarifies that "… to validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on the scores…" (2013, p. 1). Here, the intent is to clearly state the intended interpretations (use, application, context, etc.) and then approach validation by assembling supporting evidence. This is counter to the conceptualizations first offered by Cronbach and Meehl (1995) and articulated by Messick (1989a). In this latter concept, validity is a single construct to which many diverse forms of evidence are required. For Kane's (1992b) argument-based approach, the proposed score interpretation drives the type of evidence assembled. Kane, and others, argue that a unified construction is so general and undefined that it lacks any coherent approach to the process (Kane, 2013, 2013; Lissitz & Samuelsen, 2007*)*. Albeit, Kane's position is not a rejection of Messick's unified theory, but rather a structured approach to validation study. In contrast, Lissitz and Samuelsen reject the unified construct in favor of a more segmented approach (similar to Mehrens, 1997) that emphasizes content validation as the core element (Lissitz & Samuelsen, 2007).

Kane (2001) is also critical of Messick's reliance on a core theoretical foundation for a single unified vision of validity. That is, he identifies a significant drawback to the evolving complexity associated with construct validation. Specifically, if the process of validation requires examination against theoretical interpretations, then clearly it must be shown that a well-defined theoretical framework exists in the first place. Otherwise the entire act of validation is without merit and foundation. Kane (2001) reflects that "… the basic notion of implicitly defining constructs by their roles in a nomological net [i.e. Cronbach & Meehl, 1955] assumes that the network is based on a tightly connected set of

axioms. Educational research and the social sciences in general have few if any such networks" (p. 325). To this end, Kane distinguishes between strong and weak validity paradigms, a concept initially advanced by Cronbach (1988). Strong paradigms require that substantive theories be in place which allow for strong inferential examinations. Lack of such substantive theories degrades the quality of the validation process. Kane (2001) notes that under a strong theoretical paradigm, theoretical assumptions and conclusions are articulated and subjected to empirical challenges. This approach is essentially aligned to the process of theory testing in science (Kane, 2001, p. 327). In contrast, weak theoretical paradigms permit what Kane has remarked as opportunistic strategies of validation where evidence is reliant on readily available data that may or may not inform on the construct of interest.

Offering a more discordant view, Lissitz and Samuelsen (2007) go father in their concern for Messick's unified conceptualization of construct validity. Here the authors comment that "…much of the dissatisfaction with Messick's unitary concept of validity is based on the notion that his rather global view of the topic is impractical…" (p. 437). They contend that content validity should form the center of the validation activity. Lissitz and Samuelsen deconstruct validation into two component parts: internal evaluation of content (test development and the analysis of the test itself) and external purposes and impacts. The authors argue that content validation must come first before external considerations are considered. Their thesis states that "… the internal characteristics should be determined to be the content validity of the test and that these do not [unlike Messick's framework] depend on external factors" (p. 437). Here, they are

echoing the views of Popham and Mehrens from a decade before. In essence, they are re-segmenting validity back to the traditional conceptualizations pre-Messick.

Regardless, the debate continues with Lissitz and Samuelsen's re-conceptualization being challenged by theorist such as Gorin (2007) and Embretson (2007). Regarding this ongoing and evolving discourse, Gorin (2007) comments that:

> Validity is not a box to be checked yes or no. Nor is it a checklist that once marked is a fait accompli. Validity is a judgment, and like all judgments it is relative and ever evolving. It can be, and should be, evaluated in light of new evidence and desired interpretations, making validity and validation an on-going process. (p. 461)

Validity Within Teacher Evaluation Systems

In the context of examining validity questions associated with contemporary policy-directed teacher evaluations systems, the literature base is relatively sparse. That is, there is a relative lack of formal studies that specifically evaluate teacher evaluation systems using a validation construct emphasizing theory-inference evidence (Milanowski, 2011). Most studies investigating validity claims focus on traditional criterion-based measures assessing the relationship between academic performance and assessment of instructional practices (Kimball & Milanowski, 2009; Kupermintz, 2003; Holtzapple, 2003; Milanowski, 2004, 2011). Arguably, a much larger literature base exists exploring technical aspects of estimating instructional effects using student academic performance measures (Au, 2010; McCaffrey et al., 2003; Amrein-Beardsley, 2008, 2009; Baker et. al., 2010; Corcoran, 2010; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Papay, 2011). And a growing literature base is focusing on the consequential outcomes resulting from new policy-driven evaluation systems that directly

impact teacher's professional and personal lives (Amrein-Beardsley & Collins, 2012; Collins, 2012; Shepard 1997).

With regard to validity examinations of teacher evaluation systems, this researcher reviewed selected studies that are most relevant to the focus of the research presented herein. That is, the reviews focus on studies examining evaluation systems that use value-added and standardized measures of professional practice (similar to the Danielson FFT) to judge levels of teacher instructional quality. Each are critically reviewed with regard to their alignment between methodological approaches to validation claims and the theoretical frameworks proposed by Messick (1989a), the *Standards* (AERA et al., 1999, 2014), and Shepard (1993).

This researcher finds that most of these studies address only selected aspects of construct validity with an emphasis on criterion associations between test scores and other facets of the context (Kupermintz, 2003; Kimball & Milanowski, 2009; Milanowski, 2004, 2011). This is especially relevant in the context of Shepard's (1993) observation made twenty years ago that researchers continued to rely on criterion measures as the dominant authority for conducting validation studies, much as they did in the years prior to 1955 (Cronbach & Meehl, 1955). This researcher reflects that despite the evolution taking place in validity theory, applied conduct of validation studies seem to lag behind current thinking on the topic.

Indeed, a recent summary published by the Bill and Melinda Gates Foundation (2013) regarding development of improved methods of teacher evaluation, the authors report (justify) their findings almost exclusively in terms of correlations between test scores and ratings of instructional practice (MET Project, 2010; CCSSO, 2012; Bill &

Melinda Gates Foundation, 2013). Indeed, the results are presented as a form of validation evidence, highlighting the positive correlation found between value-added achievement measures and evaluation ratings. However, closer review reveals that these correlations, while positive significant, are only weak to moderate in size and reflect only a small separation in absolute learning outcomes between high/low rated teachers (CCSSO, 2012; Bill & Melinda Gates Foundation, 2013). It is these types of selected reporting/inferential activities that lead this researcher to comment that few comprehensive validation studies are present in the literature that attempt to deeply explore the theoretical propositions of policy-based evaluation systems using multi-faceted forms of evidence.

In a paper titled “Teacher Effects and Teacher Effectiveness: A Validity Investigation of the Tennessee Value Added Assessment System,” Kupermintz (2003) explored aspects of the Tennessee Value Added Assessment System (TVAAS). In the paper, the author examines the definition used to describe effective teachers, the technical characteristics the TVAAS value-added system, and the impact student background characteristics had on inferences made from the empirical measures of effectiveness.

In the study, the author identifies the main claim of the TVAAS system: that TVAAS is capable of providing policy makers with objective and reliable estimates of teacher effectiveness. Kupermintz (2003) focus is to examine this claim stating "... [we] examine the soundness of such assertions ... [by directing our attention] ... to the manner by which estimates of teacher effectiveness are defined and calculated ... and their validity for purposes of setting educational policy and making personnel decisions" (p. 288). Interestingly, much the same set of questions are examined by Amrein-Beardsley

and Collins (2012) and Collins (2012) but from a completely different perspective. For these studies, manufacturers' claims of inferential suitability are examined based on consequential and stakeholder perceptual evidences using mixed qualitative and quantitative methods.

Kupermintz (2003) approaches the validation activity in manner that closely aligns to Kane's (1992b) argument-based construction of validity evidence. Kupermintz states that

> ... the logical and evidential bases for claims and inferences about scores obtained from any testing procedure are captured by a validity argument. The case for proposed interpretations and inference offered to support a suggested use, rests on favorable empirical findings in light of theoretical propositions regarding the nature of the construct purported to be measured. (p. 289)

In this way, Kupermintz narrows the analysis to the claim that TVAAS test scores may be interpreted as a measure of teacher effectiveness.

Kupermintz (2003) first deconstructs the assumption that the most influential factor on academic achievement is the presence of an effective classroom teacher. He notes the obvious fallacy in associating measures of academic progress to claims of instructional effectiveness: that is, simultaneously using test scores as both the definition and determination of instructional effectiveness. Here he reflects that "... unfortunately, such causal interpretation is faulty because teacher effectiveness is defined and measured by the magnitude of student gains" (p. 289). Kupermintz notes that " ... to turn full circle and claim that teacher effectiveness is the cause of student score gains is at best a necessary, trivial, truth similar to the observation that 'all bachelors are unmarried'" (p. 289).

Given this fallacy in logic, Kupermintz (2003) suggests that validation of the TVAAS system requires exploring alternative hypothesis regarding observed variation in test scores. For example, do student background characteristics, home environment, community factors impact the efficacy of instruction? The essential question becomes whether variability in achievement is influenced by non-instructional contexts that are not fully accounted for in the statistical models. This criticism includes acknowledgment of the technical limitations of the statistical models themselves (Amrein-Beardsley, 2009; McCaffrey et al., 2003; Au, 2010; Corcoran, 2010; Berliner, 2005).

In examining the statistical portions of the TVAAS system, Kupermintz (2003) notes that estimates of classroom gains are normative, comparing the results of one teacher against all other teachers in the school (district) system. He states that "…questions about fairness and equity must be raised if personnel decisions employ normative information that imply in practice different standards or benchmarks in different schools" (p. 290). The problem is that a weak teacher in a weak district (system) may be evaluated more favorably than a weak teacher in a strong district, and that districts (systems) across the state vary widely in their value-added measures. This presents one type of validity concern when making inferential decisions on teacher's tenure based on achievement measures: fairness and consistency of measures used to inform on a common construct. Another issue is that the TVAAS system treats instructional effects as "… independent, additive, and linear…" (p. 290). This means that schools engaging in team teaching, team planning, professional learning communities, and/or cross discipline collaboration, may have individual effects diluted and/or biased depending on the quality and impact of such organizational structures. Kupermintz

comments that "… when the fit between the model and the phenomenon it seeks to represent is poor, validity is threatened…" (p. 290). However, here he is only referring to the statistical model and its assumptions imposed on the data. Arguably, this is not that same as exploring the full construct of teacher effectiveness.

Other technical issues noted by Kupermintz (2003) include issues of missing data and the use of statistical techniques to fill in missing values. The author notes that the statistical impact in this regard varies across districts and thus impacts teachers within those systems differentially. A final concern involves the impact of student, school, and community non-instructional characteristics. Kupermintz is skeptical that the blocking factors used in the TVAAS models completely remove these influences. Citing impacts of non-random assignment to students to classrooms and reports revealing non-zero correlations of student and school covariates to residual gains, he suggests, raises more validity issues.

It seems that Kupermintz's (2003) conceptualization of validity is narrowly defined to simply exploring characteristics of the TVAAS scores, not whether the definition proposed for teacher effectiveness is itself improperly characterized or ambiguous. That is, the study only addresses the claims made by the TVAAS components and not in the underlying construct itself. Essentially, using this type of argument-based approach never questions what is, and is not, effective teaching. This seems counter to the tenets proposed by Messick and other contemporary theorists. Cronbach (1971), Messick (1989a), Kane (2001), and others argue that before validation can take place, a clear articulation of the theoretical construct must be explored and clarified.

In another study titled "Examining Teacher Evaluation Validity and Leadership Decision Making within a Standards-Based Evaluation System," Kimball and Milanowski (2009) explore the validity characteristics of a teacher evaluation system composed of value-added (VAM) measures and observations of professional practice in a large school district in the western United States. Milanowski (2011) also presented a study at the 2011 Annual Meeting of the American Educational Research Organization titled *Validity Research on Teacher Evaluation Systems Based on the Framework for Teaching*. Attention is paid to both these studies because the context and components of each mirror that of this researcher's district evaluation system. This includes the use of multi-level value-added models of residual academic gain as measures of instructional effectiveness and the application of Danielson's Framework for Teaching (FFT) as the basis for rating the professional practice of classroom teachers.

In the first study, Kimball and Milanowski's (2009) approach was to examine the variation in teacher evaluation ratings assigned by school principals, identify principal rating assignments that differed substantively from associated measures of student achievement, and then explore the causal contextual factors potentially impacting the decision making context of those ratings. The authors utilized a sequential explanatory mixed methods design (Creswell, 2009, p. 209) to first identify the quantitative variation in principal evaluation ratings of teachers (within schools). The quantitative methodology served as selection criteria for grouping principal evaluators into high and low deviation groups. Qualitative methods (semi-structured interviews and document analysis) were then employed to investigate the contextual aspects of the rating process and decisions. Their findings failed to reveal any cogent factors impacting deviant rating behavior, but

highlighted the complex environment within which teacher evaluation decisions are made. However, this reviewer argues that the authors' findings ignore the possibility that variation in rating assignments were due to a lack of evaluation program fidelity, inadequate evaluator training, and an overreliance on VAM as a stable measure of instructional effectiveness. In addition, it is argued that the approach was, as with Kupermintz (2003), too narrow to warrant claims of construct investigation.

The research questions initially identified by Kimball and Milanowski (2009) included: (1) How much does the validity of the performance rating relationship vary across evaluators? and (2) Are differences in evaluator decision making related to differences in the strength of the achievement-observation rating relationship? The authors summarize their interest in exploring these questions by stating "we were interested in exploring whether differences in motivation, knowledge and skill, and school context explained why some evaluators' ratings … show stronger relationship with achievement … than other evaluator's ratings" (p. 41). Given this explanation, this researcher questions whether the study constitutes a complete validation study or, as Kane (2001) puts it, an opportunistic approach to researching interesting aspects of an evaluation system.

Kimball and Milanowski (2009) utilized a two-level hierarchical linear model (HLM) to estimate student residual gain (VAM) on standardized reading and mathematics tests in grades three through five for school years 2001 to 2003. A total of 5,683 students and 328 teachers (classrooms) representing 39 school locations were included in the analysis for SY2001-02 and 9,873 students and 569 teachers representing 57 schools in SY2002-03. Only locations (principals) rating five or more

teachers/classrooms were retained. Model specification incorporated student demographic covariates within level 1 (students) while no covariates were specified at level 2 (teachers/classrooms). For each student, residual values by subject were converted to z-scores and averaged together across reading and mathematics within grades and then averaged (aggregated) by classroom. This resulted in a single vector of residual scores per teacher/classroom which could be compared against principal ratings of teacher's professional practice.

This computational approach is very similar to that being used by this researcher's district evaluation system. Here, student residual gain scores are estimated each for reading and mathematics. These two score vectors are aggregated (median) up to the classroom level. The two subject median classroom values are then averaged together to obtain a single vector of combined classroom effects. Finally, the single vector of combined classroom residual scores is transformed to a percentile scale.

Kimball and Milanowski (2009) note that the district they examined also utilized a standardized observational framework for evaluation teacher's professional practice (PP): the Danielson Framework for Teaching (FFT). The FFT is composed of four primary domains and numerous sub-domains. Principals observe and collect evidences of performance on each domain, assigning an integer rating of 0, 1, 2 or 3. The teacher's final PP score represents an average of the four primary FFT domain scores. Again this is very similar to the approach this researcher's district uses to construct PP measures. However, here the Danielson's FFT ratings are summed across elements (not averaged) to construct an overall composite PP score.

Kimball and Milanowski (2009) correlated the VAM and PP scores for each location (school) and compared these to the average correlation of across all locations (r = .22). Each location was then categorized as being high, average, or low based on their deviation from the overall district average value. The authors state "we looked for evaluators with correlations substantially above or below the group correlation" (p. 45). Two groups were eventually identified for follow up: 11 principals in the high group and 12 in the low group. However, this reviewer feels that at this point in their discussion the authors might have provided a deeper review of the limitations inherent in VAM approaches to isolating instructional effects and the impact this might have had on their selection process.

As their next step, Kimball and Milanowski (2009) interviewed principals identified in the high and low correlation groups using a semi-structured interview protocol. Three constructs shaped the interview process. The first concerned the motivation, attitudes, and willingness of each principal to conduct teacher evaluations using the FFT framework. The second concerned the evaluator's knowledge and skill of applying the FFT framework in an observational setting. And the third concerned the school context within which each principal conducted their evaluations. This latter involved reflection on the school's overall socio-economic level, achievement levels, administrator experience, and the relationships (climate) between the administrator and teachers.

Validity. As mentioned, Kimball and Milanowski's (2009) paper is conceptualized as a validity study of a standards-based teacher evaluation system. Its initial approach was to apply a criterion framework for evaluating the validity of the

observational rating (PP) scores assigned by principals to classroom teachers. Without providing any context or justification, the authors position the VAM measures as the fixed criterion against which PP scores are compared. By doing so, they assume PP scores correlating highly with VAM are valid while PP scores with weak or negative associations are deviant. In this context, deviant scores are termed invalid.

However, no theoretical foundation is provided for these perspectives other than previous research (by the authors and others). Findings from these studies are said to reveal generally weak-to-moderate VAM-PP relationships. Kimball and Milanowski (2009) state that this "… suggest that school evaluations of teachers may in fact have some validity as measures of teacher effectiveness, providing some justification for consequential use of evaluation ratings" (p. 35). However, as mentioned previously, this fallacy is problematic with criterion approaches to validity, noting Kane's (2001) reflection that "…the trouble with the criterion model is the need for a well-defined and demonstrably valid criterion measure" (p. 320). In essence, the warrant for making validity claims under this type of criterion analysis is the acceptance of the criterion measure itself. In addition, by ignoring the considerable research published on the methodological and theoretical limitations of using VAM measures as indicators of instructional effectiveness, the foundation of the entire paper's thesis becomes suspect. Indeed, the authors never really examine the construct definition of effective teaching, which is at the core of the validation exercise. By not doing so, their research is really about investigating principal's decision making in the context of evaluating teachers and not about informing on construct of teacher evaluation.

Kimball and Milanowski (2009) go so far as to state that "… if evaluators differ substantively in the degree to which their ratings correlate with student achievement, teachers could receive consequences that are not justified by the general validity evidence cited above" (p. 35). Here, the implicit perspective is that VAM measures are superior to PP (observational) scores if those scores differ. If true, then why not argue that the best measure of teacher effectiveness is simply the VAM measure and dispense with the observational framework? Throughout the entire paper validity is defined by correlation to VAM measures. Without authority, this seems to be a far too restrictive premise and certainly not aligned with the current *Standards* expression of validity (AERA et al., 1999). This researcher feels that additional detail on the HLM model fit and scale reliabilities for both the VAM and PP measures should have also been provided and critically discussed. This would have allowed the reader to reflect on scale issues and view the VAM as a contributing factor to variability measures of evaluator PP.

In the paper, the authors describe the observation framework that the principals utilized for evaluating the PP of teachers (Kimball and Milanowski, 2009). As mentioned, the system, Danielson's Framework for Teaching (FFT), is also utilized by this reviewer's district. As such, this reviewer is familiar with the components and procedures for implementing it throughout a large educational organization. In their description, the authors note that "… training [on the evaluation rubric] did not emphasize interrater consistency. Furthermore, school administrators were not scored for the accuracy of their evaluation ratings, … however, one training session did have principals observe actual classroom teaching in small groups …" (p. 42). In addition, Kimball and Milanowski state that "… in subsequent years, optional training was

available to principals on how to manage the evaluation process (i.e. completing evaluations by their due date)" (p. 42).

From the above, it is evident that the implementing district had little evidence of evaluator consistency (reliability) or system fidelity. Yet, the entire premise of the paper is centered on the assumption that deviation of PP scores from VAM measures reveal substantive contextual issues impacting rater judgment. Indeed, this served as the rationale for conducting a qualitative examination (interviews) of causal factors impacting the principal's activities. However, this logic is suspect since no attention is paid to the possibility that variation in PP scores could be due to a lack of training on the scoring rubric, inconsistent (non-standard) application of the rubric, differing levels of understanding of the system's protocol, and a general lack of fidelity to implementing the FFT framework properly. These factors clearly inject construct-irrelevant variance into the scale's measures that would impact any correlational approach to validity (Fast & Hebbler, 2004; Berliner, 2005; Messick, 1989b, 1995). That the authors do not connect their acknowledgement of inadequate training and lack of evidence of program fidelity weakens that paper's conclusions and reflections.

Methodology – FFT Scales. Kimball and Milanowski (2009) briefly describe how they constructed the final PP scores used in the study. From their description, principals rated each teacher on each of the four FFT domains. However, no mention is made of how, or if, ratings were assigned to the many sub-domain items present in the framework. They state "… we used the simple average of the four composite scores to obtain an overall measure of teacher performance" (p. 43). As described, it appears that the authors chose to throw out a great deal of scale information (and potential variation) by averaging

across the four domains and not summing the values into a total score. If the four domains have a total possible score of 3, then summing across the scale provides a range of discriminating values between 0 and 12. By averaging, the range of the final scores is restricted to be between 0 and 3. Choosing this scaling method directly impacts the correlation analysis performed and makes suspect its interpretation. Finally, it appears that the principals did not include ratings of the sub-domains which would have added information and extended both the range and discrimination of the evaluation scale. For these reasons, this researcher's district sums the individual FFT ratings assigned by evaluators into a total possible PP score ranging between 0 and 66 (22 elements, each with a possible value between zero and 3).

Methodology – Conceptual Framework. Kimball and Milanowski (2009) identify two groups of principal evaluators to follow up on regarding contextual factors impacting rating judgment. They indicate that their conceptual frame for engaging in the interview process involves three concepts: will, skill, and context. Will refers to the intrinsic motivations and attitudes of the principals in the evaluation process. Skill refers to the technical knowledge and fidelity of process implementation. And context refers to the local school community and policy environs impacting the principal and his/her teaching staff. However, they offer no foundational discussion supporting this framework. Indeed, no literature is cited and no theory is proposed as to why these constructs are believed to be important causal factors. Since the majority of the qualitative portion of the paper is founded on this conceptual framework, lack of supporting discussion detracts from the authors' findings.

This reviewer feels that the paper displayed substantive issues with regard to some basic assumptions and premises. The reliance on VAM as a fixed criterion standard without any acknowledgement of the methodological and inferential issues is a serious omission in the text. Ignoring the lack of evidence on interrater reliability and program fidelity as potential major factors impacting scale variance is another. The lack of supporting discussion and foundation on the conceptual framework of the qualitative approach leaves the reader asking why a different perspective wasn't chosen or considered. And methodological issues associated with construction of the evaluation scores raised questions on the metrics used in the analysis. Each of these detracts from the paper's overall soundness and prevents the reader from accepting the author's concluding reflections without reservation.

In his 2011 study titled "Validity Research on Teacher Evaluation Systems Based on the Framework for Teaching," Milanowski focuses specifically on validity evidence associated with the Danielson FFT framework. He begins by suggesting that upwards of 200 U.S. school districts use Danielson's FFT framework or some close variation on its structure, to evaluate the instructional practices of classroom teachers (p. 4). His intent is to summarize the validity evidence associated with the FFT based on selected published and unpublished studies. His selection is limited to studies that examine the relationship between FFT ratings and value-added measures "… of teacher effectiveness" (p. 4) as well as reliability measures associated with rater agreement and consistency of the FFT sub-scales. A portion of the reviews presented rely on data gathered during research collaborations with Kimball (Milanowski & Kimball, 2005) and Heneman (Heneman & Milanowski, 2003). Three additional collaborators are mentioned, White and Gallagher,

and Odden, but common author citations of such are not explicitly apparent in the reference section of the paper. However, within the paper he indicates use of findings from unpublished research collaborations.

To begin, Milanowski (2011) positions criterion-related validity as

> … the idea that if there is an external standard of performance (the criterion) then ratings should correlate with or predict measures of the standard. For many policy makers and educational leaders, value added is the accepted criterion, if not definition, of teacher effectiveness … (p. 9)

Seemingly then, Milanowski forms his validation perspective based on the precept that value-added measures are the de facto representation of the effectiveness construct. From the outset, this researcher argues that by adopting such precept, it allows him to ignore any and all of the technical issues presented in the value-added literature and provides a convenient starting point from which to make claims. Indeed, if any positive relationship is found between value-added scores and FFT ratings, then the FFT ratings are shown to be valid representations of the teacher-quality construct. Indeed, after presenting his analysis, Milanowski states that "…the evidence reviewed in this paper does have value because it shows that evaluation systems based on the Framework can produce reliable ratings that correlate with value added estimates of teacher's contributions to student achievement" (p. 5).

Based on the extensive discussion of validity provided previously in this paper, Milanowski's use of the term validity seems flawed and lacking in clarity: that is, what exactly is being investigated - the underlying theoretical construct or a simple association between two data sets each of which have something to do with education? Had the author approached the activity as an examination of criterion evidence from which the underlying construct of teacher quality would predict a significant, positive (causal)

relationship between teacher ratings and student learning, then the approach would have more closely conformed to contemporary theories of construct validity. Regardless, the author would still be required to critically review the value-added components as substantive representation of the teacher quality construct.

In the report, Milanowski (2011) presents correlations between value-added and FFT ratings from four large school districts. In general, his findings report small to medium positive correlations for some locations for some years. However, the results vary considerably by district and year and not all associations are significantly different from zero ($p < .05$). When significant, the correlations range between .19 to .48 (M=.28). He suggests that complementary evidence is provided by Kane, Taylor, Tyler, and Wooten (2010) but notes that "… it [the external study] used very different analytical strategy …" (Milanowski , 2011, p. 12). Interestingly, the Kane et al. (2010) study cited is unpublished.

Milanowski (2011) comments that Kane et al. (2010) found positive correlations between FFT ratings and value-added measures. However, again, not all of the correlations were positive and some were very weak. Commenting on some of the findings provided by Kane et al. (2010), Milanowski highlights that

> … the difference in achievement for a teacher with an average rating one standard deviation higher is .08 of a standard deviation in math and .10 of a standard deviation in reading. A teacher whose average score would place her at the 'distinguished' level will have students whose student achievement is about one-fifth of a standard deviation higher [than] a teacher at the proficient level" – *distinguished* being a higher performance category that *proficient.* (pp. 13-14)

In contrast to what Milanowski seems to be implying, from this reviewer's perspective the variations in FFT ratings had minimal effects on measures of student achievement. In addition, Milanowski notes that "… slightly larger effects were found for the average of

the eight standard ratings when using student achievement from the same year as the observations were made. The differences were not statistically significant, however." (p. 14). Thus, notwithstanding the credibility issue of inferring value-added as a valid representation of the effectiveness construct, the findings reported do not seem to provide convincing evidence that the relationship between FFT and achievement is strong, consistent, and an inferentially compelling basis for evaluating instructional quality.

Milanowski (2011) goes on to review another collection of studies using data from a single charter school in Los Angeles. Here again, presentation of the correlations reveal a mix of significant and non-significant results depending on year (SY2000-01 to SY2002-03) and subject (reading, math, and language arts). No discussion is offered as to why, within a given year, correlations might vary considerably across subject areas and across years if the latent construct being measured was instructional quality. For example, why would a significant positive correlation ($r = .48$, $p < .05$) exist in reading but not be significant in math for the same year and same group of teachers? Milanowski also reviewed two studies (Schlacter & Thun, 2004; Daley & Kim, 2010) reporting correlational data based on the TAP evaluation system. He reports their findings of positive significant correlations ranging between .21 and .70. However, he acknowledges that the sample (*n*-count) sizes for all these studies were small (he does not report the actual n-counts for any of the studies referenced). He leaves it to the reader to investigate the technical details for each citation. Finally, Milanowski references two additional studies not using the FFT framework, citing correlations between .17 and .28, but no information of significance is provided.

Milanowski (2011) concludes the section of criterion validity reflecting that correlations between value-added and FFT ratings might not be expected to be very large. This is due to the presence of measurement error in value-added estimates, imperfect reliability in evaluator ratings, and misalignment between instruction and the assessments used to measure student academic progress. Regarding reliability, Milanowski reports interrater agreement on FFT for two school systems, Chicago Public Schools and Cincinnati Public Schools. In each, two raters evaluated classroom teachers. Interestingly, no further information is provided on the criteria for determining agreement. However, the author reports that the percent agreement for Chicago (n = 277 teachers) was 52% and 54%, based on two of the four FFT domains. In Cincinnati, the percent agreement was 73% and 79% on the same two domains. Citing similar results from unpublished analysis he has undertaken, the author concludes that substantial interrater agreement can exist if more than one rater are utilized. Again, no statistical measures of rater agreement are presented.

Milanowski (2011) makes the following reflections regarding conducting validity studies into teacher evaluation systems. First, criterion studies need to be completed in more sophisticated ways. The author acknowledges that many factors influence the relationship between achievement and classroom practices. In addition, the gains in achievement from increased competency may not be linear, but may display diminishing returns (that is, the influence of new instructional innovations may be asymptotically limited). Also, just as non-random assignment of students to classrooms may bias value-added estimates, it may also bias evaluation ratings. Finally, multi-year criterion studies might provide stronger evidence of the connection between best practice and student

learning. Milanowski goes on to suggest that research needs to expand to more *proximal* evidence of construct validity (p. 22). By this he means a deeper examination of professional practice beyond the structured frameworks (i.e. Danielson's FFT) currently being utilized. From this, a better understanding of what it means to be a good teacher might be forthcoming.

Overall, this researcher feels that Milanowski's reliance on criterion measures is misrepresented as validity. Notwithstanding his discussion for the need of other forms of evidence, the criterion approaches he presents simply do not adequately attend to the primary construct in question (Messick, 1989b, Kane, 2001, Linn, 2008): that is, teacher instructional quality. While the study provides useful context to the achievement-evaluation association (albeit, weak and inconsistent), only a limited attempt is made to explore the latent construct from more diverse sources of evidence such as content examinations using both qualitative and quantitative methods, consequential evidences aligned to stated purposes, scale reliabilities and consistency measures, and a much more articulated definition of the construct itself. These are precisely the limitations from the literature that informed the design considerations being proposed by this researcher herein, a multi-faceted mixed-method's approach to construct validation that utilizes concurrent forms of evidence to inform on the premise of evaluating teacher instructional quality.

As a contrast to Milanowski's focus, Tuytens and Devos (2009) approached their examination of the validity of a new teacher evaluation system implemented in the Belgium school system from a different perspective. Here, they sought to understand the system in terms of the connection teachers had to the new policy, arguing that "… the

implementation of teacher evaluation in particular is equally problematic as the implementation of any other policy, if not more" (p. 925). In this way, Tuytens and Devos evaluated the perceptual and consequential components of evaluation systems using survey research methods similar to approaches adopted by Amrein-Beardsley and Collins (2012) and Collins (2012). Tuytens and Devos argue that the evaluation process can cause emotions giving rise to defensive routines where "… teachers' emotions can be an obstacle in teacher evaluation" (p. 925).

To understand teachers perspectives', the authors constructed an instrument to measure teacher perceptions. The premise was that by doing so, improvements could be made to the change environment (policy, communication, implementation, trust, acceptance, etc.) that would increase the efficacy of the system toward its intended purpose. They comment that "… individual teacher participation is crucial for good teacher evaluation because it increases the respect teachers have for teacher evaluation and nurtures the quality of the evaluation and the use of received feedback" (Tuytens & Devos, 2009, p. 926). This perspective focuses on understanding the personal environs of evaluation and its impact on people. Their findings showed that teachers were favorable to the activity of evaluation but had concerns over implementation issues. Changes could then be made that permitted higher levels of acceptance, understanding, and trust in the resulting ratings.

Because policy-directed, high stakes teacher evaluation systems are relatively new additions to the education reform landscape, few comprehensive studies exist that evaluate the consequential impact they have on teachers' personal and professional lives. This is despite the contributions of Messick (1989a), Cronbach (1971), Shepard (1993,

1997) and other theorists who have argued convincingly for inclusion of such evidences as a core component of score interpretation and validation. However, Amrein-Beardsley and Collins (2012) and Collins (2012) recently explored aspects of consequential inference using case studies of teachers in the Houston Independent School District. Amrein-Beardsley and Collins (2012) write that "… there lingers a paucity of research evidence to support the attachment of significant consequences to value-added output" (p. 3). Their comment is made in the context of the relatively large volume of research regarding the technical shortcomings of value-added models as reliable measures of teacher effectiveness (Au, 2010; McCaffrey et al., 2003; Amrein-Beardsley, 2008, 2009; Baker et. al., 2010; Corcoran, 2010; Newton et al., 2010; Papay, 2011). Similarly, Collins (2012) remarks that "… despite widespread popularity of [value-added systems], no research has been done to examine how teachers and their practices are impacted by this methodology that professedly identifies effective and ineffective teachers" (p. 3).

Over 20 years ago, Messick's (1989b) contention was that "… the consequential basis of test use is the appraisal of both potential and actual social consequences of the applied testing" [emphasis added] (p. 10). Twenty plus years later, Amrein-Beardsley and Collins (2012) and Collins (2012) finally address the premise by placing the facet of consequence front and center in the discussion of score validity. Whether value-added models of instructional effectiveness are used singularly or in tandem with ratings of professional practice, score validation necessitates such inquiry beyond the positivist view of numerical analysis, distributions, and correlation.

Amrein-Beardsley and Collins (2012) critically examine the evidence underlying decisions to terminate four classroom teachers employed by the Houston Independent

School District (HISD), located in Houston, Texas. Here, HISD constructed two independent measures of teacher competency. The first was based on value-added models of student academic gain generated by SAS®, the developers of the statistical model. The second relied on evaluator (observational) ratings of professional practice. From this standpoint, the components are much the same as in this researcher's local district. However, in HISD, the two measures are interpreted independently while in this researcher's context the two measures are combined to form a single Teacher Instructional Quality (TIQ) Scale.

Amrein-Beardsley and Collins (2012) used a mixed-methods approach to examine evidence related to score correlation, "… reliability, bias, teacher attribution, and validity" (p. 3). However, as used here, the latter reference to validity as a separate area of inquiry raises questions since each of the other evidences arguably contribute to a common understanding of construct validity. Indeed, in the report, the authors systematically examine the adequacy of each source of evidence in light of the termination decision but never address validity as a unique set of contributing data. Regardless, the context of the study was to examine both intended and unintended effects of judgments made based on both the empirical and inferential characteristics of HISD's evaluation system. Their findings include the following:

1. Systemic Issues: Using survey methods, the authors indicate that the majority of teachers related favorably to being evaluated by trained evaluators, but did not trust measures derived from statistical value-added models (note: no information is provided on the type of observational system used by HISD). That is, HISD teachers viewed ratings emanating from the value-added

measures as chaotic and random. As such, trust in the data was lacking. In addition, teachers felt that student attributes impacted the statistical outcomes, making the gain scores unfair and/or suspect. Approximately 46% percent of teachers indicated that their value-added rankings fluctuated when switching grade levels and 55% believed their value-added and evaluator-assigned ratings did not agree. Worse, some suggested that release of value-added scores influenced evaluators to assign commensurate observational ratings (higher/lower), biasing the final performance measure. A minority of teachers raised concerns about being evaluated on tests that did not match their curriculum. Finally, teachers expressed little knowledge about the method or proper interpretation of the value-added measures, nor had they received significant professional development on its application to instructional planning.

2. Correlations: Across the four focus teachers, subject area correlations with evaluator ratings varied considerably, both in size and significance. Some associations were significantly negative, suggesting the higher the academic gains, the lower the evaluation PP rating. In one case, the teacher was awarded Teacher of the Month and then Teacher of the Year, but had a negative relationship between her value added and evaluator ratings. Correlations also varied considerably across years, displaying a lack of consistency. Interestingly, regardless of whether a teacher's value-added measure was statistically different from their peers, the HISD reports displayed the estimated value. Thus, teachers unfamiliar with the statistical measures might easily

misread and misinterpret their associated meaning. All four teachers reported not receiving substantive training in score interpretation.

3. Reliability: The author's report that "… across the four teachers, issues with reliability were evident" despite SAS® claims of model precision and accuracy (Amrein-Beardsley and Collins, 2012, p. 15). Indeed, Amrein-Beardsley and Collins note that "… the probability that three of the four teachers added or detracted value from year-to-year was roughly the same as the flip of a coin" (p. 15). Thus, the belief that integrating at least three years of value-added data to improves reliability of measure and validity of inference seems suspect. In addition, the SAS® claims of reliability even when teachers move between grades, subjects, or classes, was not borne out by the evidence presented by the study.
4. Bias: As mentioned, interviews with the HISD teachers indicated they believed student factors biased the value-added measures of effectiveness. For example, one of the four terminated teachers taught some of the highest-need students in the district. HISD teachers also felt that having high proportions of special education (SPED) students negatively impacted their gain scores. Teachers that taught under looping structures argued that they maxed out their value-added growth in the first year. Similarly, teachers in high need ELL classrooms, where many students get transitioned back to mainstream classes, believed their aggregate gain scores were negatively impacted. High transient teachers believed that their scores were inconsistent because they were constantly adapting to new conditions, students, schools, and/or subjects.

5. Teacher Attribution: Here, the authors question the ability of valued-added systems to isolate and localize instructional effects to a single teacher. This is especially true under conditions of team teaching, co-teaching, enrichment and/or remediation programs. In addition, one terminated teacher was not the teacher of recorded for half of the students she taught because she moved between grades mid-year. Another teacher taught alongside an enrichment teacher who assisted about half of her students. Another taught along with a reading teacher four days per week. Interestingly, the SAS® company makes claim that such distortions are remediated through linking procedures that apportion instructional impacts appropriately. However, Amrein-Beardsley and Collins reflect that "… breaking up effort across teachers using percentages and proportions is nonsensical given the interaction effects that occur among and between students and teachers" (2012, p. 17).

Overall, Amrein-Beardsley and Collins (2012) argue that the HISD evaluation system is not suitably measuring teacher instructional quality. Indeed, they state that "…the conclusion here is that there is nothing substantive to evidence that a valid teacher evaluation system … is in place and in use" (p. 18). Despite the evidence, three of the individuals declined to pursue legal recourse and left teaching in HISD. The hearing officer in the fourth case ruled that the termination decision was unwarranted based on unreliable measures, lack of data, and bias in the growth measures influenced by student background factors.

The Amrein-Beardsley and Collins (2012) study is important to this researcher both because of its innovative perspective and for the alignment the HISD system has to

the evaluation methods being implemented in the Arizona district examined herein. However, notable limitations of the study include restricted sample size ($N = 4$ teachers) for some of the analysis. In addition, it was unclear exactly how the component correlations were computed and how the small sample size impacted the computational findings. Finally, this researcher felt that a tighter connection between the teacher perception data and the validation activity could have been examined. Tuytens and Devos (2009) argue that teacher perspectives directly impact evaluation outcomes and system efficacy. Factors such as lack of trust, concern, and disconnect with the process are reflected in empirical evaluation ratings, degrade system integrity, and create alignment issues with intended goals (i.e., unintended consequences). Arguably, construct validation is the focus here, not the individual data or particular form of analysis, and it is the collection of many forms of evidence that inform on the construct. This researcher feels that the perceptual data might have been more completely brought into the validation context in terms of how it might have effected teacher professional practice, collegial relationships with evaluators, and how/if it may have impacted scores.

In her dissertation titled *Houston, We Have a Problem: Studying the SAS Education Value-Added Assessment System (EVAAS) from Teachers' Perspectives in the Houston Independent School District*, Clarin Collins (2012) builds on the Amrein-Beardsley and Collins' (2012) study by focusing on intended and unintended consequences of using value-added models as the basis for measuring teacher competency. Using a mixed-method approach, Collins surveyed HISD teachers in kindergarten through eighth grade on their attitudinal perspectives of the EVAAS system in the areas of reliability, validity, formative use, and consequences related to the

EVAAS system. In addition, she investigated teachers' beliefs regarding the claims made by the company as the reasons/benefits for using the system to evaluate classroom teachers.

To do so, Collins (2012) constructed a 52 item questionnaire that was administered online to over 6,000 classroom teachers. A total of 882 (eligible) teachers completed the survey (for a response rate of approximately 14%). The questionnaire used a combination of fixed-choice, constructed response, and Likert-scale formats. With regard to the latter, 16 Likert-type items were presented emphasizing EVAAS' marketing claims and posited benefits to classroom teachers. Examples of these statements include: "EVAAS helps improve instruction, EVAAS helps increase student learning, and EVAAS helps you become a more effective teacher" (pp. 176-177). All of the statements were worded in the affirmative and scored using a five option attitudinal response scale (S*trongly Agree* = 5, *Agree* = 4, *Neither Agree nor Disagree* = 3, *Disagree* = 2, and *Strongly Disagree* =1). A *Not Applicable* option was also provided and not scored. Response analysis (mean, standard deviations, response counts) included use of the neutral center point *Neither Agree nor Disagree*. Interestingly, Collins notes that "…more than 50% of the teachers disagreed or strongly disagreed with every single statement" (p. 119)

Data/methods validation included receiving feedback from colleagues in the university, the Director of the Houston Federation of Teachers (teacher's union), and meetings with a selected group of HISD teachers. In addition, seven HISD teachers were asked to pilot the online version of the survey. Communications were also maintained with HISD employees in the Department of Research and Accountability for the purpose

of clarifying information and answering questions. A measure of internal consistency was also computed (Cronbach's Alpha = .96) on the final questionnaire response set. Collins discussed the findings for each of the theoretical categories examined by the questionnaire.

*Reliability***.** Collins (2012) examines teacher's belief in the consistency of EVAAS measures over time, the operational definition of reliability implied in this portion of the study. Here, she reports that approximately 50% of teachers felt that their EVAAS scores were inconsistent. Collins notes that teachers perceived EVAAS measures to no better than a "… flip of a coin …" (p. 136). The reasons given for variation varied but included assignment changes in grade levels and/or subject areas and the influence of student characteristics on growth scores. Collins notes the "… consensus among teachers was that gifted, transitional ELL, and special education students were the most difficult student groups to demonstrate high levels of growth as measured by EVAAS…" (p. 137). This implies that the teachers did not trust the system to remove these types of non-instructional influences, despite manufacturer's claims.

*Validity***.** Collins (2012) examined the validity construct by asking teachers to respond/reflect on factors potentially influencing test scores. Findings indicate that issues of student mobility (into and out of classrooms during instructional time) were of primary concern. Here, teachers felt that conditions related to team teaching, multi-grade classrooms, etc., biased interpretation of EVAAS results by obfuscating the one-to-one nature of the teacher-test score alignment. In addition, teachers lacked trust with the methods used by the manufacturer to apportion instructional effects across multiple teachers. Teachers also noted the lack of consistency between EVAAS and evaluator

ratings of professional practice (PP). Indeed, over half of the teachers felt that these measures did not agree (p. 139). Some reflected that "… principals viewed EVAAS as the more objective evaluation score, in that … principals would adjust their [PP] scores to reflect their EVAAS scores…" (p. 139).

Formative use of EVAAS scores. Collins (2012) notes that teachers did not receive their EVAAS scores until after the end of the school year. As such, formative use of the data in terms of adjusting instruction based on the measures could only take place during the next school year, reducing the perceived value of the measures. Indeed, 60% of the teachers indicated that they did not use their EVAAS scores for any type of formative or reflective activity. Teachers indicated that the EVAAS reports were vague, unclear, and teachers were not quite sure how to interpret (p. 142). However, Collins notes that some teachers mention using the scores to identify bubble kids, students to focus on with the greatest chance of increasing their achievement (p. 142). In this way, teachers were, in effect, gaming the system by making decisions in order to maximize the greatest gains on the EVAAS measures. At the same time, more than a third of the teachers were unaware of any training opportunities to learn about the EVAAS measures and/or how to interpret and use them in their instructional settings. One-half did not even discuss the information with their principals and teachers report that principals were unable to provide useful insights on score interpretation and use for instructional planning.

Intended and Unintended Consequences. Collins (2012) study found that a large majority of teachers strongly disagreed with the manufacturer's EVAAS marketing statements and inferential claims regarding the system. Stating "… overwhelmingly,

teacher respondents reported not believing that the EVAAS model has benefitted much of anything …" (p. 145). In addition, the teachers expressed: (1) a desire not to teach specific groups of students because of the difficulty in obtaining significant growth results, (2) the need to teach to the test in order to obtain EVAAS growth points, and (3) a sense of increased competition among colleagues in terms of attaining higher growth measures (p. 146). Collectively, these findings suggest that teachers may be narrowing the curriculum they teach to just the tested subjects and content (specific competencies in mathematics and language arts that are easily assessed using standardized testing approaches). In addition, it appears that teachers have a self-interest to be assigned to populations that they believe give them the best chance of obtaining growth points and not to teach in classrooms with high needs or problematic students. Finally, Collins' data suggest that teachers are incentivized not to engage in collegial community (team) building that might somehow enhance the scores of other teachers and detract from their own.

The findings of Collins (2012) add to the understanding of a specific component of construct validation regarding implementation of high stakes teacher evaluation systems. That is the role of stakeholder perspectives and its contribution to an understanding of the latent construct: here, the adequacy of EVAAS scores as a measure of teacher instructional competency. However, this researcher suggests that even these claims are subject to critical examination. Evidencing stakeholder (teacher) perspectives (opinions, attitudes, etc…) does not, in itself, qualify as construct evidence. Rather, a direct link between stakeholder perspectives and the latent construct must be posited and

investigated. Collins seems makes this point in her recommendations for future studies, noting the many avenues of new research suggested by the findings.

However, this researcher might still question the veracity of the author's closing statement: that

> …from this study, what we know most importantly is that at least in HISD, the "most comprehensive and reliable" VAM on the market … is not working as Dr. Sanders and SAS intended and marketed, and instead, is resulting in negative, unintended consequences that appear to be harming the teachers, and by mere association, the students whom these teachers teach. (Collins, 2012, p. 153)

This researcher suggests that, given the type of data collected, this statement might be a bit of a reach without more clearly connecting the perceptual and empirical evidence that behavioral modifications were actually taking place. However, it is noted that this was not the purpose or design of the Collins study. Thus, this critical reflection is not about the quality of the data or the findings themselves, merely the strength of connection being made to the EVAAS construct itself.

For example, teachers' perceptions that the alignment between EVAAS scores and companion evaluator ratings are inconsistent and, in some instances, counterintuitive, is not evidence that this association is, in fact, real. In order for the perceptual data to qualify as construct evidence, it must be shown that the consequences resulting from these perceptual conditions are actually impacting the EVAAS measures and biasing the inferential decisions made from those scores. In essence, teachers may not trust the value-added measures, but that does not prove that the value-added measures do not, in fact, measure instructional quality (notwithstanding the considerable published research on the topic). In the HISD context, Collins' paper supplies compelling insight that such perceptions encourage teachers to game the system, distort the instructional fidelity of

curriculum frameworks, and reduce stakeholder trust in the system. But construct validation, arguably, requires additional examination of the perceptual evidence in terms of its manifested behavior: that is, how did perceptions impact behavioral decisions leading to distortions and bias of the EVAAS measures?

As a final comment on the Collins (2012) study, the author notes in some detail issues impacting the overall response rate to the online questionnaire, approximately 14% of the eligible teacher population (pp. 81-86). The author suggests that this low response rate restricts the ability to "… generalize the findings of this study beyond Houston [ISD]…" (p. 85). However, this researcher might suggest that generalizations even to the HISD population itself might be in question. Collins did examine potential bias between the union and non-union status for teachers responding to the survey. However, no discussion is provided concerning potential bias between responders and non-responders. That is, to what degree are the findings reflective of the greater HISD population? It would seem reasonable to compare the data obtained from the demographic questions asked at the beginning of the survey activity (years teaching, degrees earned, grade levels taught, economic status of students taught, etc…) with similar population level data to ascertain adequacy of generalizations made to the HISD community (i.e. Chi-square tests of homogeneity). Non-parametric measures (Gibbons, 1993) of association might have been applied, assuming that population level information was available. Indeed, this may not have been the case.

Measuring Teacher Effectiveness: Value-Added Models

McCaffrey et al. (2003) examine issues related to value-added modeling (VAM) from both a technical and applied perspective. They note that predominantly as a result of

the No Child Left Behind Act (NCLB) of 2001 "… the use of standardized test scores to hold schools, teachers, and students accountable for performance is now the cornerstone of many education reform efforts in the United States" (p. 1). They cite three primary approaches for operationalizing accountability measures. The first is a cohort to cohort comparative approach where schools and districts are held accountable for improving academic performance from one year to the next. Here, the achievement of students at one point in time is compared to a different group of students at a previous point in time. A second approach measures the proportion of students reaching pre-specified levels of achievement within a given year. This approach is exampled by threshold criteria specified under NCLB known adequately yearly progress (AYP). However, McCaffrey et al. note that a fundamentally different way to measure accountability is to track the achievement growth of a specific cohort of students over time. This third method serves as the foundation for value-added models.

McCaffrey et al. (2003) describe value-added modeling as "… a collection of complex statistical techniques that use multiple years of student's test score data to estimate the effects of individual schools and teachers" (p. xi). However, they noted that the use of value-added models in high-stakes decision-making raises important questions, both from a statistical modeling and measurement perspective. That is, if teachers and schools are to be evaluated using value-added models "… an obvious question involves the amount of sampling error in the estimates and rankings …" (p. 4). Second, they question the impact omitted variables and other modeling specification issues might have with regard to biasing parameter estimates. Third, the authors note that value-added

estimates are sensitive to test construction issues, types and quality of test items, and the resulting scale reliabilities.

McCaffrey et al. (2003) identify two reasons why value-added modeling was attracting interest by both policymakers and practitioners. First, they note that value-added approaches were showing promise in their ability to isolate the instructional effects of teachers both at the classroom and school building levels. Second, recent publications were purporting to show large differences in instructional effects. By isolating these effects, actions could then be formulated to improve instructional efficacy, target professional development, inform policy decision making, and increase student outcomes.

The authors reviewed selected published studies utilizing VAM approaches for evaluating teacher effectiveness. Their aim was to clarify the most important modeling issues, to evaluate the practical impact of VAM as a measure of effectiveness, to spur additional work in the field, and to help inform the public policy debate concerning accountability.

McCaffrey et al. (2003) identify that recent studies show strong instructional effects, and that these effects persist for three to four years. However, they note that these results are based on a relatively small number of studies, some of which were not published in peer-reviewed journals. Additionally, shortcomings in the studies make it difficult to explicitly isolate the size of the effects and they suspect that the magnitudes are somewhat overstated (p. 4). Some of the issues they identify include: (1) a lack of sufficient evidence regarding estimation bias, (2) the impact of missing and incomplete data, (3) failure to fully account for important student characteristics and variables, (4)

scaling issues, (5) restricted samples, and (6) a lack of technical detail from which to evaluate the appropriateness of the methods used.

McCaffrey et al. (2003) comment that accurate inferences of teacher effects require precise estimates. However, modeling decisions, data quality, and other factors contribute to increasing error variance and bring into question the validity of inferences that can be made. In this regard, the authors note a significant number of issues impacting accuracy and validity of VAM approaches. For example, confounding factors make it difficult to disentangle instructional impacts from other student, classroom, or school factors including impacts from previous teachers. Ignoring or omitting these factors may bias what analysts conceive as the true impact of instruction. Additionally, the authors acknowledge that any type of longitudinal modeling approach inevitably increases the proportion of incomplete (missing) records. Finally, the authors highlight the limited number of content areas that traditional standardized achievement measures represent. Thus, the types of measures used in VAM might insufficiently reflect of all of the impacts that public schooling have on children.

In their review, the authors are addressing issues of construct validity and the need for close examination of available evidences. From their discussions, it seems clear they feel that making instructional quality decisions based solely on achievement measures is questionable. They note, "… if we are to make inferences about teachers, the outcomes-based definition of effects might be insufficient without additional investigations showing that positive effects correspond to other notions of effective teaching…" (McCaffrey et al., 2003, p. 14). In addition, they state that "…if empirical estimates of teacher affects do not correlate with other generally accepted traits of

effective teachers, we might be concerned that our statistical estimation of teacher affects is to error-prone to be useful" (pp. 14-15). For these reasons, the authors recommend that value-added measures of teacher effects be contrasted with other measures of teacher effectiveness as a means of validating the inferences required. Regardless, empirically evaluating potential sources of bias are critical. The authors conclude that "the research base is currently insufficient to support the use of value-added measures for high-stakes decisions" (p. xx).

Papay (2010) explores the impact that alternative same-subject achievement measures have on value-added estimates of teacher effectiveness. He notes that while considerable research has concentrated on the technical specifications of value-added models (VAM), less effort has been focused on the impact of achievement metrics used in these models. In his research, Papay applies various VAM specifications to data from three different reading tests – a state-mandated test, the Stanford Achievement Test (SAT), and the Scholastic Reading Inventory (SRI).

Papay generates estimates of teacher effects in grades three through five across nine different VAM specifications, each incorporating measures from the same three reading tests. The VAMs apply differing specifications of student, teacher, and school covariates. To assess the degree of association, Papay generates value-added metrics across teachers and then compares the results using Spearman (rho) Rank Correlations. His findings reveal positive correlations among the three tests for each of the model forms (a total of 27 comparisons). Papay reports the correlations to be all statistically significant indicating that increasing effectiveness values based on one test scale are matched with increasing values on the other. However, the strength of these correlations

are all small to moderate (below .50 except in three instances). As a result, Papay states "… these inconsistent classifications would have substantial consequences for any policy that rewards teachers based on their value-added scores" (p. 180).

Papay's findings add to the growing issues researchers are uncovering regarding the use of value-added models to estimate teacher effects in a high stakes policy environment (Wayne, 2010). That is, questions of technical adequacy are conflated with the choice of measures used.

Newton et al. (2010) evaluate the stability of teacher effectiveness ranking across various value-added model (VAM) specifications, courses taught, and years. The approach uses a single set of achievement metrics in mathematics and English language arts across five VAM model specifications. In addition, the impacts of various student characteristics are examined under the common assumption that, once they are statistically controlled for, they have limited impact on teacher rankings.

In their research, Newton et al. (2010) generated effectiveness ratings for teachers in six high schools in the San Francisco Bay area based on measures from the California Standards Test (CST). Four OLS regression-based VAMs were constructed to predict current year outcomes after controlling for prior years achievement. In addition, Newton et al. specified a three-level mixed-effect model that nested students within classrooms within schools. Some model specifications incorporated student covariates (race/ethnicity, gender, free/reduced priced lunch, ELL status, and parent education) with additional formulations incorporating a school fixed effect component. Final teacher effectiveness ratings were computed by averaging the difference between predicted and actual achievement (residual) scores within classrooms. The choice of modeling

frameworks was influenced by "… the practical limitations [faced by] many district and state systems…" throughout the country in response to mandated accountability requirements (p. 8).

Teacher rankings produced by the models were evaluated for concordance using Spearman (rho) Rank Correlation and Interclass Correlation Coefficients (ICC). Results indicate that effectiveness rankings are positively correlated but vary substantially across statistical models (rho values between .76 - .95). ICC values for same teachers across different courses were insignificant indicting that the effectiveness based on one course may not be consistent with rating received for another course. Teacher ratings computed for two consecutive years report only a moderate association in language arts (rho values between .34 - .48) and slightly higher in mathematics (.43 to .63). Finally, ex-post examination of teacher rankings reveals that student characteristics continue to influence rank ordering even after statistical adjustment via the VAM estimation process. That is, effectiveness rankings were significantly negatively correlated to a student's race/ethnicity (r = -.26 to -.43), ELL status (r = -.29 to -.48), and poverty (r = -.20 to -.45), and positively correlated to parent's educational attainment (r = .28 to .48).

Newton et al. (2010) quantify the impacts as follows: 12% to 33% of teachers rankings fluctuated by 2 or more deciles across the different VAM models estimated; 54% to 92% of the teacher's rankings fluctuated by 2 or more deciles across courses taught; and 45% to 63% of the rankings fluctuated across the two years of evaluation (p. 15). Finally, correlation to demographic factors suggest that "… teachers who were teaching greater proportions of more advantaged students may have been advantaged in

their effectiveness rankings, or more effective teachers were generally teaching more advantaged students" (p. 11).

The authors state that "… a substantial share of what some would call a 'teacher effect' actually measures other factors that are correlated with student characteristics" and that simply improving the technical attributes (better tests, better data systems, and improved statistical methods) will fail to solve the problems raised here (Newton et al., p. 19). Newton et al. conclude that caution must be used in interpreting value-added results especially in high stakes policy environments.

Wayne (2010) evaluates the reasonableness of using value-added approaches as measures of instructional effectiveness specifically with regard to evaluating the instructional quality of classroom teachers. She begins by acknowledging the increasing trend of publically releasing measures of individual teacher effectiveness to the local community as part of a broader application of school accountability and education reform. Wayne agrees that, on the surface, such measures seem to satisfy a common sense perspective for judging whether a teacher's impact on student learning is or is not adequate. However, Wayne stresses that the growing body of research brings into question the suitability of value-added models (VAM) for such purposes.

Wayne (2010) articulates six critical reasons why VAM are problematic for the purpose of evaluating individual teachers. First, statistical errors have been shown to be unacceptably high when using between one and three years' worth of achievement data to evaluate instructional impact. Indeed, she quotes errors between 25 to 35% suggesting that there is a 1 in 4 chance of incorrectly rating a teacher as being ineffective or effective.

Second, research reveals that VAMs are inherently unstable in assigning year-to-year effectiveness ratings noting that "…if test scores are an accurate measure of teacher effectiveness, effective teachers would rate high consistently from year to year because they are good teachers" (Wayne, 2010, p. 1). However, the large instability of the ratings over time suggests that the scores have less to do with effective instruction and have more to do with the cohort of students changing year to year in any specific classroom.

Third, the day-to-day stability of a student's test score seriously impacts the suitability for their use as direct measures of instructional effectiveness. Wayne notes that 50 to 80% of a student's improvement or decline in a standardized test score can be attributed to a one-time, one day, randomly occurring factors such as eating breakfast on testing day, family issues, peer issues, or issues with the instructor. She states that because of this "… using tests to evaluate teachers ignores the impact of individual daily factors that are completely out of the control of the teacher" (p. 2).

Fourth, non-random assignment of students to classrooms and schools significantly impacts the VAM results even after taking into account individual demographics factors. Non-random assignments refer to classrooms or schools with highly concentrated homogeneous populations segregated by factors such as limited English proficiency, poverty, special education, etc. The implication, Wayne notes, is that teachers "…could be punished, dismissed, or lose tenure purely because of the course they teach or the school they teach in" (p. 2).

Fifth, Wayne (2010) argues that high stake tests used in most VAM approaches do not account for the complexities of learning resulting in imprecise measurements. She points out that learning is not linear and different students are at different locations in

their learning trajectory at any given point in time. In addition, different teachers across different subjects all collectively impact a student's learning. For example, a social studies teacher may have a profound impact on a child's reading and writing ability due to the type of instructional practices he/she uses (daily reading and writing activities, homework requirements, or project-based activities that all support basic literacy skills). However, in a VAM framework, the child's reading measures are assigned solely to the reading teacher, obfuscating the contributing roles each play in the generating the test score.

Finally, Wayne (2010) argues that the impact of out-of-school factors simply cannot be ignored. Conditions situated around access to health care, poverty, parental involvement and education levels, language proficiency, family and personal stress, all negatively affect a child's ability to learn at his/her capacity, regardless of the teacher's credentials and practice. Wayne writes "…we incorrectly assume that teachers have the ability to overcome [these factors]" (p. 2).

Wayne (2010) states that "the reality of standardized tests is that they are too imprecise and inaccurate to measure the effectiveness of individual teachers" (p. 2). She questions why policy makers are not equally concerned about the accuracy of such measures when used to identify high/low performing teachers.

Organizational Change Theory

Arguably, the implementation of any type of organizational evaluation system represents significant change to those involved. With regard to teacher evaluation, the implementation of a state policy-directed approach is, in many way, more disruptive because it is being imposed externally under the context of high stakes accountability. In

the Arizona context, organizational insiders are both the designers and recipients of the change activity: recipients due to the political constraints imposed by legislative action, and designers due to the methodological flexibility provided by loosely-defined policy guidelines. Recognizing that implementation of evaluation systems reflects significant organizational change suggests that review of change theory is warranted. In this regard, a number of change theories are discussed below that this researcher feels provide important context to the evaluation process.

The Concerns-Based Adoption Model (CBAM)

Hall, Loucks, Rutherford, and Newlove (1975) described their perspective on change implementation in the following manner:

> Based on our experiences in the field as practitioners and adoption agents and on our past research efforts, we have found that "change" or innovation adoption is not accomplished in fact just because a decision maker has announced it. Instead, the various members of a user system, such as teachers and professors, demonstrate a wide variation in the type and degree of their use of an innovation. One of the reasons for this variation is the commonly overlooked fact that innovation adoption is a process rather than a decision point… (p. 52)

In this way, Hall et al. approach the implementation of new innovations (change) as a process involving individual stakeholders. Importantly, they argue that successful implementation of organizational change requires acceptance by stakeholders in terms of their emotional connection to the change process. The Concerns Based Adoption Model (CBAM) attempts to formalize the nature of this connection. Loucks and Hall (1979) write:

> … the Concerns Based Adoption Model (CBAM) has been designed to describe change as it affects individuals and to prompt more successful change efforts. CBAM views the teacher as the focal point in school improvement efforts, yet also acknowledges social and organizational influences. (p. 1)

In this way, the CBAM focuses on individuals as a key component of any change initiative. It posits that with any innovation, successful and sustainable implementation requires support for the persons involved. Loucks-Horsley (1996) suggest that when people experience change, the types of questions they have and their relationship with the initiative progresses. They evolve from questions that are more personal and self-oriented into ones of process and procedures, and then finally onto impact and consequence.

The CBAM is made up of two conceptual components. The first is the *Stages of Concern*. This area attempts to describe "… the feelings that individuals experience in regard to the innovation" (Loucks & Hall, 1979, p. 5). It identifies seven types of concerns that individuals experience as they progress through the change process. They include (1) Awareness, (2) Informational, (3) Personal, (4) management, (5) Consequence, (6) Collaboration, and (7) Refocusing. Loucks and Hall depicts this progression as beginning with concerns about self, moving into concerns about task, and then into concerns about impact. At the extremes, Awareness reflects a complete lack of concern while Refocusing reflects an internal acceptance of the change initiative and how it might be improved.

The second component is the *Levels of Use*. This component describes an individual's behavior as they experience the process of change (p. 5). It is composed of eight stages including (1) Non-use, (2) Orientation, (3) Preparation, (4) Mechanical, (5) Routine, (6) Refinement, (7) Integration, and (8) Renewal. Once individuals are introduced to the innovation, they move from non-use to orienting themselves to its structure and impact. As familiarity and experience progresses, their use gradually moves to routine. Refinement reflects a stage whereby individuals modify the innovation

structure to suit their own changing needs and conditions. In a sense they own the innovation and have accepted it as part of their person-professional practice.

The CBAM framework highlights the need to first focus on individual perceptions and perspectives before addressing the how-to-do-it stage of implementation. It also emphasizes the need to support individuals over an extended period of time and identify organizational priorities related to the innovation process. Using the CBAM as a guiding framework reminds leaders of the human dimension of innovation and the need to design implementations that attend to personal attributes of the participants.

Everett Roger's Diffusion of Innovation

Rogers (2003) formulates a theory of diffusion describing the spread of innovations within an organization. His theory is tied to the idea of a tipping point whereby a newly implemented innovation reaches a self-sustaining, self-directing state of existence. This progression is determined by the rate of interaction and acceptance of the recipients to the innovation. Initially, a small minority of stakeholders quickly adapt to the innovation. Rogers designates this group as Innovators (readily accepting the innovation with little resistance). This is followed by a slightly larger group termed Early Adopters, followed by a still larger group designated as the Early Majority.

Roger's (2003) Diffusion of Innovation theory posits a five step pathway for individuals to reach acceptance of new innovations. This pathway includes (1) Knowledge, (2) Persuasion, (3) Decision, (4) Implementation, and (5) Confirmation. This is a continuum where knowledge reflects a person's awareness, followed by increasing familiarity and comfort leading to eventual acceptance and use in their personal-professional activities. The ultimate success of the innovation depends on reaching the

tipping point whereby enough stakeholders have adopted the innovation as to be self-sustaining. Reaching that point requires a succession of innovators and early adopters to be seen by others as trusted opinion leaders which subsequently bring along more hesitant individuals.

Roger's (2003) pathway aligns with the Loucks and Hall (1979) CBAM framework by recognizing that individuals progress through stages of acceptance and that the success of any new innovation is tied to the level of acceptance of the individuals being impacted. Both Rogers' Diffusion of Innovation theory and the CBAM represent critical understandings for leaders implementing new innovations into the workplace.

Finally, like the change theorists discussed above, Kotter's (2011) theory on eight steps to organizational change emphasizes the importance of stakeholders' emotional connection to the change process. In his blog post titled "Before You Can Get By, People Need to Feel the Problem" (February, 2011), Kotter states that people will not consider change until they truly believe there is a problem that needs to be addressed. He argues that people need to make an emotional connection to a problem before they will act. The best way to do this, he suggests, is to express the problem in a way that makes it feel real.

The theorists discussed above all place considerable importance on stakeholder perspectives as a condition for successful innovation and change. The implementation of a new teacher evaluation system clearly impacts many stakeholders, all of whom need to assimilate the basic construct of evaluation if the goals of the system are to be realized. As describe by Amrein-Beardsley and Collins (2012), Collins (2012), and Tuytens and Devos (2009), teacher perspectives can impact the efficacy of the system by influencing the very measures used to describe the instructional construct. Indeed, the personal

perspectives of stakeholders are directly incorporated into the validity theory through consequential aspects of validation posited by Messick (1998), Shepard (1993), and the *Standards* (AERA et al., 1999, 2014).

Community and Change Theory

Wenger's (1998) work on communities of practice, membership, identity, and knowledge transfer across boundaries offers significant insight into organizational change and innovation. Wenger argues that communities exist within social groups and that membership within and between these communities has great impact on how new learnings (change, innovation) take shape. This is especially evident when the change initiative is occurring external to the community such as with imposition of a new teacher evaluation system. Here a state policy-directed evaluation framework is being imposed on the organization while at the same time organizational decision makers are charged with the technical design and implementation details.

In this context teachers are positioned as recipients of the change initiative, principals are positioned as implementers, and central office leadership are the designers and decision makers. Each of the groups may be considered their own well-defined community with their own identity and unique sense of boundary. Wenger comments that "…the boundary of a community of practice is reified with explicit markers of membership, such as titles, dress, tattoos, degrees, or initiation rites" (Wenger, 1998, p. 104). Membership in each constitutes the presence of common trust, allegiance, and alliance that serves to form the boundary through which communication and knowledge transfer with outside entities occurs. Using Wenger's framework, for each, some form of knowledge brokering must occur between communities in order to exchange information

and knowledge (i.e., a new form of evaluation proposed by central office decision makers). Here the term *brokering* means "… connections provided by people who can introduce elements of one [community of] practice into another" (p. 105). To do so, this requires a designated knowledge broker.

According to Wenger (1998), essential to the knowledge broker's role is the degree of legitimacy he/she has within the target community and this legitimacy is closely tied to identity. For example, this researcher is charged with the technical design of the metrics used to evaluate classroom teachers. The Director of Human Resources is charged with overseeing the implementation of the evaluation framework and its use in assessing teacher competence including training of school evaluators. Finally, site principals are changed with observing classroom teachers and rating their performance on each of the 22 FFT elements. Arguably, none of these individuals are located within what Wenger might describe as the teacher community of practice. Similarly, each must broker some aspect of the evaluation system to the teacher community, since they are positioned as recipients in the system. Key to the success of this is the degree to which teachers trust the knowledge brokers and the claims they advance for acceptance of the system, that the evaluation system, its measures, outcomes, and consequences capture the essence of teacher instructional quality in a fair and accurate manner.

Kotter's (1996) theory of change depicts Wenger's concepts in a different, but complementary, way. In his blog post titled "Why Should Anyone Trust Your Vision" (2010, November), Kotter states that the most critical aspect of effective change leadership and innovation is the existence of a culture of trust among colleagues. Without it, he suggests, innovation takes the form of "small numbers decide, large numbers

execute." Indeed, his eight steps to change emphasize the need for innovation leaders (whatever their community membership) to communicate the vision to all stakeholders (Step 4) and empower people and remove barriers (Step 5; Kotter, 1996). To do so, change leaders must broker knowledge between communities, instill trust, and promote a common sense of organizational identity.

Similarly, Hargreaves and Shirley (2009) in *The Fourth Way of Change,* speaks about five pillars of purpose and partnership. While the authors are speaking in terms of social innovations, the meaning has alignment with Wenger's concept of community. Hargreaves and Shirley's first pillar of change is having an inspiring and inclusive vision. This refers to communicating a compelling moral purpose that brings everyone to a common vision without regard to community membership or organizational position. The authors argue that without a common purpose, participants cannot see the overarching need to change and risk being caught up in interpersonal issues of control, power, and ego. In essence, knowledge brokering across boundaries becomes restricted. To overcome this barrier to innovation, the authors argue for broad stakeholder engagement, investment, and instilling a sense of responsibility across all participants. Here again, Hargreaves and Shirley's perspectives argues for a need to support the emotional conditions of change leadership.

Competency and Empowerment

While not explicitly stated as a condition for successful innovation by most change theorists, Wenger (1998) identifies competence as a critical requirement for group membership and knowledge building. He notes that "… membership in a community of practice translates into an identity as a form of competence" and it is the perceived

competence of boundary brokers that can determine the successful sharing of knowledge between communities (p. 153). Thus, a key aspect of successful innovation concerns stakeholders' perception of competency and the trust it instills between stakeholders and communities. That is, a broker must provide a compelling reason for change and convince stakeholders that change is important and necessary. In Kotter's (1996) Eight Steps to Change, Roger's (2003) Diffusion of Innovation, Loucks and Hall's (1979) CBAM, and Wenger's (1998) membership perspectives, each require competence and trust as a characteristic of the innovator.

A common attribute to all of change theories discussed is the importance of engaging and empowering individual stakeholders to ensure successful innovation. Hargreaves and Fink (2006) argue that sustainable innovation is achieved only through distributed, shared, leadership where stakeholders are empowered to adapt and evolve solutions situationally, not just in response to a directive change agent. In essence, to sustain an innovation, the originator of the initiative must seek to empower others to own and develop the innovation as needed. Additionally, within Kotter's (1996) Eight Steps to Change, step number five is explicitly expressed as 'empower others to act' while Rogers' (2003) Diffusion of Innovation specifically tracks the empowerment pathway whereby individuals learn, decide, implement, and then confirm their acceptance of the change initiative. In Roger's final level (confirmation), individuals evaluate the merits of the innovation and then adapt them as their own context.

Comment

This review of the literature was meant to position the processes of implementing a policy-directed teacher evaluation system in multiple perspectives. First, a discussion of schools and society positioned teacher evaluation in a broader social context. Second, the general nature of validity was discussed in order to ground the validation study in a theoretical framework. Third, validity studies of teacher evaluation system were presented and critically reviewed. Fourth, a brief discussion surrounding measures of teacher effectiveness placed the difficulty of developing measures of instructional quality in perspective. Finally, the context of implementing new innovations and systems in organizations was discussed. It is hoped that the information set the context by which this study was conducted and serves a useful foundation for the analytic methods discussed in Chapter 3.

Chapter 3: Methods

The purpose of this study is to examine validity evidence associated with a policy-directed teacher evaluation system implemented within a large suburban public school district located in Phoenix, Arizona. To do so, the study utilizes a single-phase concurrent mixed-methods design to investigate a variety of supporting research questions (Gelo et. al., 2008; Creswell, 2009; Plano-Clark & Creswell, 2010; Greene, 2007). This chapter discusses the mixed methodological approaches used to investigate each of the stated research questions. The discussion is organized as follows:

1st. Perspectives of mixed-methods design applied to the study of teacher evaluation systems (Outline Heading: Mixed Methods Design)

2nd. Overview of the study's primary and supporting research questions with qualitative and quantitative approaches for collecting and evaluating evidence (Outline Heading: Primary and Supporting Research Questions)

3rd. Discussion of the study's localized setting and description of stakeholder participation and sampling (Outline Heading: Localized Setting)

4th. Review of the specific methodological details aligned to each primary and supporting research question (Outline Heading: Measures and Data Analysis Plan)

5th. Discussion regarding the supporting analytic methods (Outline Heading: Analytic Methods Supporting Research Questions). This section is broken down into four discussion components:

a) Application of value-added models as estimates of instructional effects

b) Application of factor analytic techniques to explore structural aspects of Danielson's FFT theoretical Framework
c) Adaptation of Lawshe's (1975) approach for quantifying content representation of the Danielson Framework for teaching.
d) Analytic framework for qualitative investigations.

Mixed-Method Designs

This study aligns quantitative and qualitative methodological traditions according to the unique evidentiary requirements imposed by each research question. In this way, quantitative investigations serve to examine characteristics of numerically-based measures, scales, and ratings, associated with empirical representations of instructional quality while qualitative inquiries examine affective attributes that both reflect and impact the stakeholders involved. Importantly, this researcher believed that this type of mixed methods approach was necessary to better represent/understand the complex nature of the evaluation system. By incorporating multiple perspectives, data sources, and analytic tools, resulting findings are more robust, provide deeper insight, and allow for more informed policy decisions to enhance system efficacy. In this regard, Greene (2007) states that "… the primary purpose of a study conducted with a mixed methods way of thinking is to better understand the complexity of the social phenomena being studied" (p. 20). It is in this tradition that a nuanced mixed methods methodological framework was utilized to examine characteristics of the policy-directed teacher evaluation system.

Construct validation using a concurrent mixed methods design advances multiple, simultaneously occurring, forms of inquiry (Gelo et al., 2008; Creswell & Plano Clark,

2007; Creswell, 2009). Here, a variety of independent data collection pathways are followed, each from a unique analytic perspective. Findings from each are then brought together and interpreted in the context of the validity construct. An example of this construction is depicted in Figure 11.

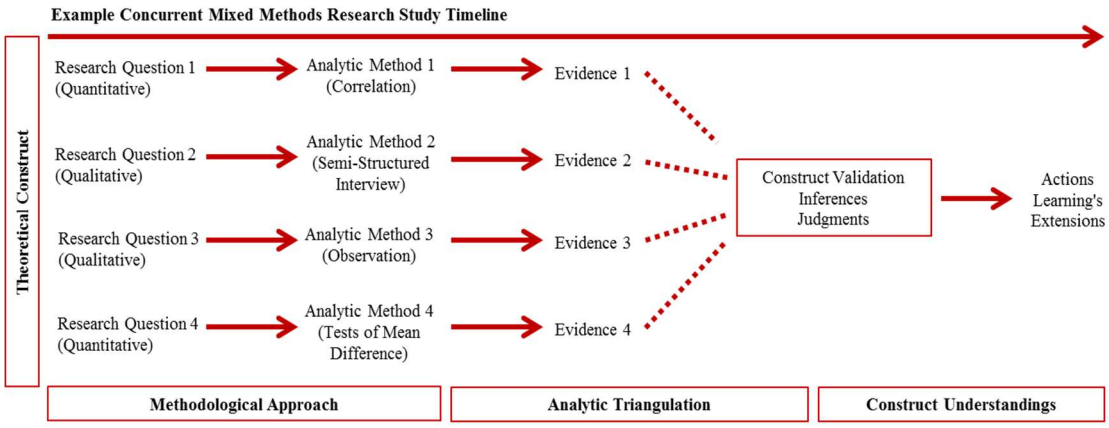


Figure 11. Example of concurrent mixed methods approach to construct examination.

Both Gelo et al. (2008) and Creswell (2009) describe this process of integrating the independent evidences as triangulation taking place in a one-phase model design. The author's reference to a one-phase design suggests that qualitative and quantitative examinations provide unique insights and are conducted simultaneously and independently. That is, they do not necessarily lead to additional investigation using an alternative analytic approach as would be suggested by a multi-phase design (Gelo, 2008, p. 281; Creswell, 2009, p. 210). Regarding triangulation in a one-phase research design, Creswell (2009) states that "… in a concurrent triangulation approach, the researcher collects both quantitative and qualitative data concurrently and then compares the two

databases to determine if there is convergences, differences, or some combination" (p. 213). He goes on to state that

> … This model generally uses separate quantitative and qualitative methods as a means to offset the weaknesses inherent within one method with the strengths of the other. … in this approach, the quantitative and qualitative data collection is concurrent, happening in one phase of the research study. (p. 213)

As noted earlier, contemporary views theorize validity as being a unified, but multifaceted, construct (Messick, 1989a; AERA et al., 1999, 2014). This is consistent with a concurrent mixed method perspective because of its reliance on merging independent sources of evidence and analytic approaches. Indeed, the methodological design adopted herein tailors specific approaches to data collection and analysis based on the nature of the research question specified. It is posited that the collection of research questions, and their associated constructions of empirical evidence, collectively inform on a broadly defined construct described as Teacher Instructional Quality (TIQ).

Primary and Supporting Research Questions

This study advances four primary research questions. The first three interrogate aspects the evaluation system's fidelity to purpose while the fourth is directed at the researcher-practitioner. The four primary questions are:

1. To what degree does the validity evidence generated by the LEA's policy-directed teacher evaluation system support inferences of Teacher Instructional Quality (TIQ)?
2. How does providing an evolving dialog on validity evidence influence organizational policy decisions regarding implementation of the LEA's teacher evaluation system?

3. To what degree have the perspectives and concerns of stakeholders been incorporated into, and influence, the system's implementation?
4. How did the process of engaging in this action-research study impact the investigator as a scholarly researcher and organizational leader?

For each primary research question (RQ), additional supporting questions are posed to more directly guide the method and type of analytic inquiry. It is at the supporting RQ level that specific modes of data collection and analytic methods (quantitative and/or qualitative) are specified along with criteria for processing and interpreting the outcomes. In addition, RQ #1 is further divided into five sub-categories of evidences - criterion, content, consequential, reliability, and construct articulation based on the theoretical components characterizing construct validity (AERA et al, 1999, 2014; Messick, 1989a; Kane, 2001; Shepard, 1997; Linn, 2008). An outline of all primary and supporting research questions and their analytic tradition addressed in this study are provided in Appendix B. Additional detail and discussion for each component is provided in the following sections of this chapter.

Localized Setting

As mentioned in Chapter 1, the information used in this study comes from a moderately large public school district situated within a middle income residential suburb of the greater Phoenix metropolitan area. The district enrolls approximately 24,500 K-12 students across 20 elementary and 4 high schools and employs approximately 1,230 classroom teachers. Approximately 5% of students are classified as English Language Learners (ELLs), 12% receive some type of special education services, and 52%

participate in the Federal National School Lunch Program (NSLP).[4] Eleven of the district's elementary schools qualify for Title I funding from the U.S. Department of Education based on local community poverty rates. Descriptive statistics for selected demographic variables characteristic of the district are provide in Table 1.

Table 1

Selected District Demographics

	Count	Mean	Standard Deviation	Minimum	Maximum	Range
No. of Schools	24					
Elementary	20					
High School	4					
Total Enrollment	24,572	1,024	369	349	2155	1,806
Elementary	17,872	893	173	349	1,167	818
High School	6,700	1,675	400	1,083	2,155	1,072
Special Education	2,901	12%	2.76%	8%	17%	8%
Free/Reduced Lunch	12,516	52%	17.83%	27%	87%	59%
English Language Learners	1,075	5%	4.55%	0%	18%	17%
Gifted	949	4%	2.13%	1%	10%	8%

All classroom teachers employed by the district participate in the annual evaluation system and receive instructional practice ratings from their local school administrator(s). In this district, school principals/assistant principals are responsible for evaluating the practices of classroom teachers using the 2012 Danielson Framework for Teaching (FFT) rubric-based rating system. Non-continuing teachers (teachers with three or less years of

[4] NSLP is a means-tested program based on family income that provides low-cost or free lunch to qualifying students.

in-district teaching experience) are required to receive at least two formal evaluations per year. Continuing teachers (teacher with more than three years' experience) are required to be evaluated at least once per year. However, selected classroom teachers may receive additional evaluations above the required minimum based on individual plans of improvement made in collaboration with local site administrators.

Policy Criteria and Impact on Selecting Teacher Sample

This study explores claims made on the instructional efficacy of classroom teachers based on a state policy-directed instructional rating system constructed from measures of student achievement and professional practice. However, to infer instructional competence from student achievement implies that such measures be aligned with the curriculum content areas being taught by the teacher. Without this, the suitability of such inferences, arguably, falls into question. Within the context of the applied setting discussed herein, a number of issues arise from policy-imposed conditions on measures of student achievement.

First, state legislative policy directives mandate the use of standardized tests that are administered systemically across all like-classrooms in the district. That is, a reading test administered in one fourth grade school/classroom must be the same assessment administered within all fourth grade classrooms across the district. This makes sense in that the intent of the evaluation system is to uniformly evaluate the competency of all teachers not just those in a single school or classroom. Second, state policy requires that assessments used in any evaluation system display suitable data quality standards with regards to documented reliability and validity. To the extent that in-house assessments do

not have supporting psychometric evidence, they cannot be used to judge the instructional efficacy of classroom teachers.

These conditions raise considerable issues from both an evaluative and research perspective. Simply put, most teachers in the district do not utilize standardized, systemic, assessments of the curriculum content they teach. Many assessments utilized in classroom settings are locally constructed and not consistently administered throughout the district. In addition, most classroom level assessments do not meet data quality criteria due to a lack of documented psychometric evidence.

Impact on Teacher Sample

When these conditions are applied, it significantly reduces the number of teachers for which curricula-aligned achievement data is available. Essentially, classrooms (teachers) meeting the criteria are limited to regular education, self-contained, elementary school classrooms in grades three through six who teach the state standards in the areas of reading and mathematics. All students in these classrooms annually take the AIMS test. Science is included for Grade 4 teachers since that subject is tested by the state assessment in that grade level.

These distinctions become important in terms of the sample of participants addressed by each of the primary and supporting research questions in this study. For questions that investigate professional practice ratings, all evaluated teachers may be included. However, examinations involving value-added gains in test scores will focus only on classrooms with content-aligned assessment data (AIMS). Given this background, the following information is assembled to give the reader an idea of teacher participant count by grade level (Table 2).

Table 2

Teacher Participant Count by Grade Level

	Count	% of all district teachers
Total number of K-12 classroom teachers	1,230	
Total number of elementary (K-8) classroom teachers	873	71%
Total number of high school classroom teachers	357	29%
Number of classroom teachers with content-aligned AIMS Reading/Math/Science assessment information (Group A)	317*	26%
Group A who are not a *Guest Teacher*, who have complete data, and a class size greater than 10	238	19%

Note: (*) Regular Education, Self-Contained, Grade 3 - 6

Measures and Data Analysis Plan

This section of the paper will address methodological details including sample participants, measures, data collection, analytic procedures, and inferential criteria. The discussion is sequentially organized by primary and supporting research question. Specific descriptions of the VAM models used in the study will be discussed including model specification, estimation methods, and the data sets employed. Following these discussions, details on the specific analytic techniques/instruments referenced under the various research questions is provided including descriptions of the Content Validity Index (Lawshe, 1975; Wilson, 2012), approaches to interview and observation coding, and application of exploratory and confirmatory factor analytic techniques.

Research Question #1 (RQ1)

To what degree does the validity evidence generated by the LEA's policy-directed teacher evaluation system support inferences of Teacher Instructional Quality (TIQ)?

This Research Question is defined by five types of component evidence: (a) Criterion, (b) Content, (c) Consequential, (d) Scale Reliability, and (e) Theoretical Construct Definition. It is theorized that each component contributes to an understanding of the construct validity of the evaluation system (AERA et. al., 1999, 2014). Details for each component will be provided using a common outline format.

RQ1A: Criterion evidence. The policy-directed teacher evaluation framework posits that measures derived from assessments of VAM and PP collectively inform on unique, but interrelated, aspects of the latent (unobserved) TIQ construct. Here, PP is positioned as an instructional input facilitating student learning. In contrast, VAM is interpreted as an instructional outcome causally connected to PP. Under this paradigm, greater degrees of instructional competence (PP) are assumed to facilitate higher levels of student learning (VAM). Similarly, higher levels of achievement (VAM) are assumed to originate from greater degrees of instructional competence (PP). It follows that significant positive correlations should exist between the two measures. To explore the criterion evidence associated with the teacher evaluation system, five supporting research questions are advanced.

Supporting research question RQ1A(a). To what degree do value-added measures of instructional effectiveness correlate with measures of professional practice (PP)?

- Description: Overall TIQ ratings are comprised of two components: measures of VAM and PP. Under this research question, the correlation between the two

measures is investigated. It is hypothesized that if both VAM and PP domains represent aspects of the TIQ construct, positive and significant correlation will exist between the two measures.

- Measures: Pearson Correlation Coefficients (*r*); Coefficient of Determination (*r*-squared)
- Sample/Participants: Classroom teachers reporting both VAM and PP measures. This will include teachers in regular education, self-contained, classrooms in grades 3 to 6 with 10 or more students.
- Timeline: June – August 2013
- Data Collection: VAM measures are estimated using multi-level statistical models. PP ratings are assigned by school evaluators and placed into an online database. Both data sets are linked using student and teacher identifiers.
- Data Analysis Plan: VAM measures associated with each teacher will be correlated to their corresponding PP rating.
- Reliability Plan: All VAM models generate measures of model reliability. Pseudo *r*-squared statistics will be referenced as well as measures of model fit and variance explained. Comparisons will be made between conditional and unconditional model variance. Scale reliability statistics will be referenced for PP measures including inter-item correlations.

Supporting research question RQ1A(b). To what degree do measures of PP assigned by qualified evaluators correlate with teacher's self-assessment of PP? [Reader's Note: This question was not formally evaluated because teacher's timeline for completing a formal self-assessment was delayed]

- Description: All participating teachers complete a self-assessment using the same rating instrument as the evaluating principal. This research question investigates the degree that teachers' self-assessments align with evaluator ratings. It is hypothesized that positive associations will exist between the two measures, indicating that teacher professional awareness is consistent with input and mentoring being offered by the school instructional leaders.
- Measures*:* Pearson Correlation Coefficients (*r*); Coefficient of Determination (*r*-squared)
- Sample/Participants: Classroom teachers reporting both VAM and PP measures. This will include teachers in regular education, self-contained, classrooms in grades 3 to 6 with 10 or more students.
- Timeline: August - October, 2013
- Data Collection: All PP ratings are maintained in an online database. Self-assessment and evaluator ratings will be linked using a teacher identification number.
- Data Analysis Plan: Self-assessment and evaluator ratings will be correlated to each other
- Reliability Plan: Scale reliability statistics; Descriptive statistics.

***Supporting research question RQ1A(c*).** To what degree do VAM estimates of instructional effectiveness in reading and mathematics correlate?

- Description: VAM-based instructional effectiveness measures are constructed as a combined impact between reading and mathematics achievement. Both are assumed to contribute to the TIQ construct. It is hypothesized that the

correlation between each subject-level VAM measure will be positive and significant.

- Measures: Pearson Correlation Coefficients (*r*); Coefficient of Determination (*r*-squared)
- Sample/Participants: Classroom teachers reporting both VAM and PP measures. This will include teachers in regular education, self-contained, classrooms in grades 3 to 6 with 10 or more students.
- Timeline: June - August, 2013
- Data Collection: VAM measures are estimated using multi-level statistical models developed by the researcher. These data are housed in a local database.
- Data Analysis Plan: VAM Measures for reading and mathematics will be correlated for all teachers, grades 3-6, instructing in self-contained, regular education classrooms with 10 or more students.
- Reliability Plan: All VAM models generate measures of model reliability. Pseudo r-squared statistics will be referenced as well as measures of model fit and variance explained. Comparisons will be made between conditional and unconditional model variance.

Supporting research question RQ1A(d). To what degree do PP sub-scale scores display similar degrees of correlation with VAM measures?

- Description: The overall PP ratings attributed to classroom teachers are constructed as the sum of 22 behavioral elements ratings. These elements are clustered into four theoretical sub-domains: Planning and Preparation,

Classroom Environment, Instruction, and Professional Responsibilities. Sub-Domain scores are defined as the sum of the sub-domain element ratings. Each of these sub-domains are posited to contribute to the TIQ construct. As such, it is hypothesized that positive significant correlations exist between the sub-domain scores and VAM measures.

- Measures: Pearson Correlation Coefficients (*r*); Coefficient of Determination (*r*-squared)
- Sample/Participants: Classroom teachers reporting both VAM and PP measures. This will include teachers in regular education, self-contained, classrooms in grades 3 to 6 with 10 or more students.
- Timeline: June - August, 2013
- Data Collection: VAM measures are estimated using multi-level statistical models. PP ratings are assigned by school evaluators and placed into an online database. Both data sets are linked using student and teacher identifiers.
- Data Analysis Plan: VAM Measures will be correlated against each of the four PP sub-domain scores.
- Reliability Plan: All VAM models generate measures of model reliability. Pseudo r-squared statistics will be referenced as well as measures of model fit and variance explained. Comparisons will be made between conditional and unconditional model variance. Scale reliability statistics will be referenced for PP measures including inter-item correlations for each sub-domain.

Supporting research question RQ1A(e). To what degree are high, middle, and low VAM estimates of instructional effectiveness able to differentiate PP?

- Description: The TIQ construct assumes that both VAM and PP measures are able to differentiate and distinguish instructional competence of classroom teachers. In addition, it is posited that these measures are positively significantly related. Here, VAM measures are considered to be outcome measures resulting from the instructional characteristics of teachers. It is hypothesized that low/middle/high VAM measures originate out of low/middle/high levels of instructional competence. To the degree that this is true, low/middle/high VAM designations should be positively significantly related to low/middle/high measures of PP.
- Measures: ANOVA tests of mean differences; F-Test; post-hoc multiple comparison tests of significance (Bonferroni)
- Sample/Participants: Classroom teachers reporting both VAM and PP measures. This will include teachers in regular education, self-contained, classrooms in grades 3 to 6 with 10 or more students.
- Timeline: June – August, 2013
- Data Collection: All VAM and PP information is contained in the district's evaluation database.
- Data Analysis Plan: To select the sample, VAM estimates will be generated for all teachers in grades 3 to 6, self-contained, regular education classrooms with 10 or more students with valid VAM and PP measures. The distribution of VAM will be divided into low/middle/high categories based on the percentile location of the classroom VAM measure: Low VAM (below 10^{th} percentile), Middle VAM (45^{th} to 55^{th} percentile), High VAM (90^{th} percentile

and above). For each of these three categories, average PP scores of teachers will be computed. ANOVA tests will be run on the mean PP scores between categories. Post hoc tests will be conducted to discern specific location of group mean differences.

- Reliability Plan: ANOVA F-Test will be used as an omnibus test of overall significance followed by post-hoc test of significance between sub-category means.

RQ1B: Content evidence. Content validation explores the premise that the items/instruments used to assemble empirical measures adequately represent and/or cover the construct of interest. In the case of teacher evaluation, content evidences involve examining the suitability of the instruments used to assemble measures of instructional competence, specifically, the Danielson FFT framework. This framework organizes PP across four behavioral domains consisting of 22 measureable elements. The theory assumes that distinguishable characteristics of instructional competence may be exposed based on ratings assigned by evaluators to each of the domain elements. Content examinations explore the degree to which (1) the data support the theory-based premise that instructional competence may be distinguished across the four behavioral domains (planning, classroom environment, instruction, and professionalism), and (2) the rating elements present adequate coverage and representation of latent TIQ construct. These two examinations may be investigated using factor analytic (exploratory and confirmatory) techniques and representations provided by content experts. The procedures for exploring this evidence are provided below, organized by each supporting research question.

Supporting research question RQ1B(a). To what degree do empirical ratings of PP correspond with the theoretical FFT construct?

- Description: Danielson's framework posits that instructional competence is comprised of four distinct behavioral domains: planning, classroom environment, instruction, and professionalism. In turn, each of these behavioral domains may be independently distinguished and measured. It is this distinguishability attribute that permits assessment/judgment of instructional strength or needed improvement. To the extent that the empirical data support the presence of four distinguishable domains, the structure of the FFT framework supports its use in planning, mentoring, and professional development.
- Measures: Factor extractions; Factor loadings; Chi-square; Item correlation Matrices
- Sample/Participants: Classroom teachers reporting both VAM and PP measures. This will include teachers in regular education, self-contained, classrooms in grades 3 to 6 with 10 or more students.
- Timeline: June – August, 2013
- Data Collection: Evaluation ratings are maintained in the district's evaluation database. These data reflect the evaluation ratings assigned by local evaluators (principals) to each of the FFT 22 elements
- Data Analysis Plan: Exploratory and confirmatory factor analysis will be used to explore alignment between the Danielson FFT theoretical framework and the empirical data. Factor extractions and element loadings will inform on

underlying structures inherent in the data. Covariance measures between theorized factors will be used to examine distinguishability assumptions. Confirmatory models will be estimated using Mplus (V. 7) structural equation modeling software. Exploratory models will be estimated using SPPS (V. 21; principal components factor extractions based on item correlations; Varimax orthogonal and direct Oblimin oblique rotation)

- Reliability Plan*:* Chi-square; KMO and Bartlett's test of sphericity; Extraction explained variance; Eigenvalues; Cronbach alpha

Supporting research question RQ1B(b). To what degree does the factor analytic structure of empirically-based PP scores differ between less experienced and more experienced teachers?

- Description: The Danielson FFT Framework is used throughout the organization to evaluate the instructional competence of all teachers without regard to distinguishing characteristics of teachers such as course content, grade level, location, experience, etc. The research question presented here specifically investigates whether the FFT theoretical structure remains consistent for less/more experienced teachers. To operationalize this, a distinction is made between continuing and non-continuing teachers. Based on the district's use of the terms, continuing teachers reflect individuals with more than three years of in-district teaching experience. In contrast, non-continuing teachers are defined as those with three or less years of teaching experience. To the extent that the factor structures revealed by the empirical data are similar across both groups, similar inferences in instructional

competence may be supported. However, if the factor structures differ substantively, use of the data to infer common areas of instructional strength/weakness may not be justified.

- Measures: Factor extractions; Factor loadings; Chi-square; Item correlation matrices
- Sample/Participants: Classroom teachers reporting both VAM and PP measures. This will include teachers in regular education, self-contained, classrooms in grades 3 to 6 with 10 or more students.
- Timeline: June – August, 2013
- Data Collection: Evaluation ratings are maintained in the district's evaluation database. These data reflects the evaluation ratings assigned by local evaluators (principals) to each of the FFT 22 elements
- Data Analysis Plan: Exploratory and confirmatory factory analysis will be used to explore alignment between the Danielson FFT theoretical framework and the empirical data. Factor extractions and element loadings will inform on underlying structures inherent in the data. Covariance measures between theorized factors will be used to examine distinguishability assumptions. Confirmatory models will be estimated using Mplus (V. 7) structural equation modeling software. Exploratory models will be estimated using SPPS (V. 21; principal components factor extractions based on item correlations; Varimax orthogonal and direct Oblimin oblique rotation).
- Reliability Plan: Chi-square; KMO and Bartlett's test of sphericity; Extraction explained variance; Eigenvalues; Cronbach alpha;

Supporting research question RQ1B(c). To what degree do the 22 elements contained within the theoretical FFT framework adequately represent the latent TIQ construct?

- Description: The Danielson FFT Framework is composed of 22 individual elements, each of which is assigned numerical ratings by participating evaluators. It is theorized that the 22 elements collectively provide an adequate (full) representation of the latent TIQ construct. That is, substantive components of the TIQ construct are not omitted, underrepresented, or misrepresented based on the item content. In addition, it is assumed that each element of the FFT framework offers important and meaningful information that contributes to distinguishing degrees of instructional competence. To explore this research question, a Content Validity Index questionnaire is utilized to assess the degree that the FFT elements provide useful information on the TIQ construct (Lawshe, 1975; Wilson 2012). In addition, interviews are used to assess stakeholder perspectives regarding the suitability of the FFT Framework to both identify and distinguish between levels of instructional quality.
- Measures*:* Content Validity Index (CVI) questionnaire; Coded stakeholder interviews.
- Sample/Participants: (CVI) Two stakeholder groups were identified to complete the CVI questionnaire - Instructional growth coaches ($N = 24$, $n = 14$) and members of the district's Teacher Evaluation Committee ($N = 12$, $n = 9$).

(Stakeholder Interviews): Random sample of teachers (N= 238, n = 7), purposeful sample of principals (N = 24, n = 8), purposeful sample of district policy members (N = 5, n = 4), and purposeful sample of state policy members (n = 3).

Teachers: Teacher selection was based on the value-added (achievement) percentile locations derived for all classroom teachers in grades 3 through 6. These individuals were divided into three performance groups based on estimated value-added residuals: High (90^{th} + Percentile), Middle (45th to 55^{th} Percentile), and Low (10^{th} or less Percentile). Three individuals from each of these groups were randomly identified (total of nine individuals). However, during the data collection period, interviews for two of the selected teachers became problematic and did not take place, leaving a total of seven individuals for which interviews were completed.

Principals: The site principals for the nine randomly sampled teachers were identified. During the data collection period, scheduling of one individual became problematic and did not take place.

District Policy: Four out of the five district policy team were identified to participate in the study due to their decision authority and familiarity with the teacher evaluation process.

State Policy: Three state-level policy leaders were identified based on their involvement in the legislative process and availability during the data collection period.

- Timeline: August – October, 2013

- Data Collection*:* Content Validity Index Questionnaire; Stakeholder semi-structured interview protocols;
- Data Analysis Plan*:* Sample participants will complete the Content Validity Index rating instrument. In addition, all sample participants will be interviewed using a semi-structured interview protocol. The interview protocol will focus on the suitability of the Danielson FFT to adequately identify and distinguish between degrees of TIQ.
- Reliability Plan: Cronbach alpha; Interview member checking; Secondary reviewers of coded interviews.

Supporting research question RQ1B(d). Do perspectives differ among stakeholders regarding the capacity for VAM and PP measures to adequately represent and differentiate the instructional quality of classroom teacher?

- Description: This research question explores the consistency of stakeholder perspectives regarding the teacher evaluation system's capacity to identify and differentiate between degrees of TIQ. To the extent that stakeholder perspectives are consistent, the organization may collectively progress with the system's implementation and application. To the extent that stakeholder perspectives differ substantively, future direction and overall system efficacy might be questioned. Finally, the findings from this question will serve as input to the change process within the organization.
- Measures: Coded stakeholder interviews
- Sample/Participants: Same as for supporting research question RQ1B(c)
- Timeline: August – October, 2013

- Data Collection*:* Stakeholder semi-structured interview protocols
- Data Analysis Plan*:* All sample participants will be interviewed using a semi-structured interview protocol. The interview prompts will include exploration of the evaluation system structure (VAM and PP measures) and its ability to identify and distinguish degrees of instructional quality.
- Reliability Plan: Item/prompt field testing; Interview member checking; Secondary reviewers of coded interviews.

RQ1C: Consequential evidence. The process of construct validation necessitates inclusion of consequential evidence that assesses alignment between intended and unintended outcomes (Messick, 1989a; AERA et al., 1999, 2014; Sheppard, 1993; Linn, 2008). Arguable, contemporary policy-directed teacher evaluation represent a high stakes, consequential, activity that warrant such investigation including critical reflection on the impact related actions have on the personal and professional lives of teachers and whether actual outcomes correspond to stated (policy) intent. In doing so, connections may be made between the proposed theoretical construct and the practical implementations of professional practice, data collection, and inferential judgments. Herein, engaging in this examination involves triangulating policy representations of TIQ with those of participating stakeholders. To the extent that these representations are aligned indicates organizational consistency in the system's impact. Arguably, such consistency permits policy makers to move forward toward intended system outcomes. However, evidence of inconsistent perspectives suggests structural problems that lead to differential outcomes and unintended consequences.

Supporting research question RQ1C(a). In what way has implementation of the teacher evaluation system affected the PP of classroom teachers (Instruction, student learning, professional capacity building, job satisfaction, etc…)?

- Description*:* The district's policy-directed evaluation system mandates a prescribed approach to measuring TIQ. This approach is composed of two primary components: test scores (VAM) and ratings of professional practice (PP). Recognition of these components, as well as the elements each uses to reify instructional behavior, is known to participating teachers. This research question investigates how the component structure of the evaluation activity impacts/shapes/directs classroom behavior with regard to teaching practice, student learning, content emphasis, etc.
- *Measures:* Coded interview responses
- Sample/Participants: Same as for supporting research question RQ1B(c)
- Timeline*:* August – October, 2013
- Data Collection: Stakeholder semi-structured interview protocols
- Data Analysis Plan*:* Coded interview responses
- Reliability Plan: Interview protocol prompt field testing; Interview member checking; Secondary reviewers of coded interviews;

Supporting research question RQ1C(b). Do the perspectives of efficacy and system affect differ across stakeholder groups?

- Description: This research question builds on the data collection specified under *Supporting Research Question (a)* by exploring whether the perspectives differ among stakeholder groups. The data collection and

sampling design permits disaggregation of the information collected by stakeholder groups.

- Measures: Coded interview responses
- Sample/Participants: Same as for supporting research question RQ1B(c)
- Timeline: August – October, 2013
- Data Collection: Stakeholder semi-structured interview protocols
- Data Analysis Plan: Coded interview responses
- Reliability Plan: Interview protocol prompt field testing; Interview member checking; Secondary reviewers of coded interviews;

RQ1D: Reliability evidence. Reliability of measure is a necessary (but insufficient) condition for claims of validity (AERA et. al, 1999, 2014; Wainer & Braun, 1988). In the teacher evaluation system under consideration, two scales are used to construct measures of overall instructional quality: VAM and ratings of PP. To examine validity evidence of the evaluation system, the statistical characteristics of the VAM and PP scales must be explored and documented. Measures with higher scale reliabilities provide smaller statistical errors, greater precision of measure, and stronger support for subsequent inferential claims. For VAM-based measures, estimates of model fit, standard errors, and tests of model significance may be examined. For PP ratings, scale reliability statistics based on measures of internal consistency may be computed.

Supporting research question RQ1D(a). What are the reliability indices for the PP and VAM scales used to form measures of TIQ?

- Description: Documentation of the psychometric and statistical characteristics of the measures used in teacher evaluation systems is required to support subsequent examinations of validity evidence.
- Measures: PP scale and sub-scale item correlations and measures of internal consistency (Cronbach alpha); tests of VAM regression assumptions; VAM prediction error (standard errors of measure); Model fit and deviation statistics,
- Sample/Participants: Classroom teachers reporting both VAM and PP measures. This will include teachers in regular education, self-contained, classrooms in grades 3 – 6 with 10 or more students.

  (PP): Data is maintained in the online evaluation database for all participating teachers. (VAM): Model statistics generated for all unconditional and conditional models used to estimate student achievement residuals will be reviewed
- Timeline: June – August, 2013
- Data Collection: VAM models estimating achievement residuals for all students in grades 3 to 8, and grade 10.
- Data Analysis Plan*:* VAM Models: Review of the fit statistics between unconditional and conditional VAM model specifications including construction of pseudo *r*-squared, variance reduction, intercept reliability, and model deviation measures.
- PP Scales: Scale reliabilities (internal consistency), inter-item correlations

- Reliability Plan*:* Inclusion of all relevant model statistics and scale reliability measures

RQ1E: Theoretical construct articulation. Contemporary perspectives posit validity as a unified construct (Messick, 1989a; AERA et al., 1999, 2014; Kane, 2001). The primary purpose of conducting validity examinations is to secure evidence as to whether empirical data reflect and inform upon the construct of interest. However, without a clear operational definition of that construct, validity examinations cannot be undertaken. Absence of a clearly defined construct prevents formulation of a tangible basis for comparison or conclusion. For this reason, central to the study of teacher evaluation systems is the need to first clearly define the concept of Teacher Instructional Quality (TIQ). Here, the supporting research questions seek to understand stakeholder perspectives on TIQ, investigate whether these perspectives differ between stakeholder groups, and provide the foundation from which to evaluate the collective evidence emanating from the evaluation system.

Supporting research question RQ1E(a). What is the theoretical construct definition held by stakeholders regarding high quality teaching?

- Description: This supporting research question explores the operational definition of Teacher Instructional Quality (TIQ) across the various stakeholder groups participating in the evaluation process. It attempts to establish a baseline operational definition of TIQ within the context of the implementing organization.
- Measures: Coded interview responses of stakeholder perspectives on Teacher Instructional Quality (TIQ)

- Sample/Participants: Same as for supporting research question RQ1B(c)
- Timeline: August – October, 2013
- Data Collection: Semi-structured interview
- Data Analysis Plan: Coded interview responses
- Reliability Plan: Interview protocol prompt field testing; Interview member checking; Secondary reviewers of coded interviews;

Supporting research question RQ1E(b). Do the theoretical construct definitions differ by stakeholder group? By VAM group?

- Description: This supporting research question explores whether the perspectives of TIQ differ between stakeholder groups. To the extent that a common understanding of TIQ exists between groups, implementation of the evaluation system across the organization will be consistently perceived. However, if perspective varies among groups, then use of a single framework design may result in implementation, fidelity, and consequential issues.
- Measures: Stakeholder perspectives of Teacher Instructional Quality (TIQ)
- Sample/Participants: Same as for supporting research question RQ1B(c)
- Timeline: August – October, 2013
- Data Collection: Semi-structured interview
- Data Analysis Plan: Coded interview responses and comparison of results by stakeholder membership.
- Reliability Plan: Interview protocol prompt field testing; Interview member checking; Secondary reviewers of coded interviews;

Supporting research question RQ1E(c). What are stakeholder perspectives regarding the purpose and intended outcome of teacher evaluation? Does this perspective differ across stakeholder groups?

- Description: This supporting research question explores stakeholder perspectives regarding the purpose and intended outcomes resulting from the teacher evaluation system
- Measures: Stakeholder perspectives of Teacher Instructional Quality (TIQ)
- Sample/Participants: Same as for supporting research question RQ1B(c)
- Timeline: August – October, 2013
- Data Collection: Semi-Structured Interview
- Data Analysis Plan: Coded interview responses and comparison of results by stakeholder membership.
- Reliability Plan: Interview protocol prompt field testing; Interview member checking; Secondary reviewers of coded interviews;

Research Question #2 (RQ2)

How does providing an evolving dialog on validity evidence influence organizational policy decisions regarding implementation of the LEA's teacher evaluation system?

The purpose of conducting validity examinations is to explore the suitability of consequential inferences based on measures of Teacher Instructional Quality (TIQ). Arguably, such evidence has value if used as input to organizational decision making to improve system efficacy. This research question examines the impact that emerging validity evidence impacts organizational decision making and facilitates changes in

system design, implementation practices, and inferential decision making. The analytic approach is qualitative, employing semi-structured interview protocols across participating stakeholder groups. In addition, field notes and reflections from the researcher-practitioner's journal provide context and documentation of impacts. The primary research question is guided by two supporting research questions outline below.

Supporting research question RQ2(a). To what extent do policy-level stakeholders value the collection and review of validity evidences as an important input to the system's ongoing development?

- Description: This supporting research question investigates the degree to which policy-level decision makers value validity evidence as part of the process of designing and implementing the teacher evaluation system. Such findings support the alignment between policy goals and actual implementation practice.
- Measures: Coded interview responses; Coded journal entries
- Sample/Participants: Same as for supporting research question RQ1B(c)
- Timeline: March – December, 2013
- Data Collection: Semi-structured interviews; Research journal
- Data Analysis Plan: Coded interview responses; Coded journal entries
- Reliability Plan: Interview protocol prompt field testing; Interview member checking; Secondary reviewers of coded interviews and journal entries.

Supporting research question RQ2(b). To what extent does validation evidence prompt changes in organizational decisions regarding system implementation?

- Description: This supporting research question investigates the degree to which developing validity evidence prompts changes and adjustments to system implementation, supporting systems activities such as professional trainings, criteria for consequential decision making, and other aspects of system application.
- Measures: Coded interview responses; Coded journal entries
- Sample/Participants: Same as for supporting research question RQ1B(c)
- Timeline: March – December, 2013
- Data Collection: Research journal; Semi-structured interviews: Policy (5), Teacher Evaluation Committee Members (8); Observational notes and reflections from workshops, trainings, meetings, collegial discussions.
- Data Analysis Plan: Coded interview responses; Coded journal entries;
- Reliability Plan: Interview protocol prompt field testing; Interview member checking; Secondary reviewers of coded interviews and journal entries.

Research Question #3 (RQ3)

To what degree have the perspectives and concerns of stakeholders been incorporated into, and influence, the system's implementation plan?

Implementation of high-stakes teacher evaluation systems impact both teachers and administrators across the organization. How policy makers design and implement these systems determines how well the innovation is accepted and ultimately their ability to affect change in professional practice. Organizational theory suggests that sustainable change is enhanced when stakeholders are engaged in the change process. Empowerment of stakeholders serves as an important predictor of change efficacy. To the extent that

stakeholders feel empowered, engaged, and respected during innovation implementation supports alignment between system goals and actual outcomes. In this regard, validity examination requires review of stakeholder perceptions regarding inclusion, empowerment, and commitment.

In spring 2011, the district assembled a Teacher Evaluation Committee composed of policy, site principals, and classroom teacher representatives. The charge of the committee was to work through the design and implementation issues related to the new system. Committee membership was established (in part) to (1) ensure representation from each stakeholder group, (2) ensure flow of information throughout the larger organizational community, and (3) provide a conduit for stakeholder input. During spring 2012, the Committee held 24 site-based faculty meetings to present technical and policy elements of the new system and to collect feedback and suggestions from classroom teachers. Additional question and answer workshops were held at each of the four district high schools during the fall 2012. An online question submission system was set up in spring 2012 allowing teachers to submit comments or questions. As follow-up, in September 2012 members of the Evaluation Committee collectively reviewed and responded to these questions; responses were posted to the online evaluation portal. Finally, the teacher evaluation committee met seven times between spring 2011 and spring 2013 to discuss technical issues and make necessary policy decisions regarding system implementation.

Given this background, Research Question #3 investigates stakeholder perspectives of empowerment and inclusion in the design and implementation of the

teacher revaluation system. The primary mode of inquiry is through stakeholder interviews aligned to each of the supporting research question outlined below.

Supporting research question RQ3(a). Do stakeholders perceive that system design and implementation decisions have incorporated their perspectives, concerns, and viewpoints?

- Description: This supporting research question investigates the degree to which stakeholders believe they have had opportunity to provide input into the design and implementation of the teacher evaluation system and that system decisions have taken stakeholder perspectives into consideration.
- Measures: Coded interview responses; Coded journal entries
- Sample/Participants: Same as for supporting research question RQ1B(c)
- Timeline: August – October, 2013
- Data Collection: Semi-structured interviews
- Data Analysis Plan: Coded interview responses
- Reliability Plan: Interview protocol prompt field testing; Interview member checking; Secondary reviewers of coded interviews and journal entries.

Research Question #4 (RQ4)

How did the process of engaging in this action-research study impact the investigator as a scholarly researcher and organizational leader?

During the process of designing and implement the agency's Teacher Evaluation System, many technical, policy, and social/community related issues will be encountered. This includes issues of positionality, power, and policy interacting at many levels in the organization. Conflicting pressures of policy adherence, technical integrity, and conduct

of scholarly research make it more difficult to engage in applied research in a real-world setting. Hopefully, learnings derived from this activity will benefit the research-practitioner in implementing future research projects, especially with regard to navigating and managing organizational innovation and change.

Supporting research question RQ4(a). What barriers or impediments were encountered during the course of the study? How were they overcome and/or handled?

Supporting research question RQ4(b). What were the salient learnings from this study? How did the researcher grow professionally and personally? How will these learnings be incorporated into future research & leadership activities?

- Description: This supporting research question investigates the challenges encountered during the research activity including approaches for resolution. The questions also concern personal learnings and growth as a researcher and scholar.
- Measures: Field notes and journal entries
- Sample/Participants: Researcher-Practitioner
- Timeline: January – October, 2013
- Data Collection: Field notes and journal entries
- Data Analysis Plan: Coded field notes and journal entries
- Reliability Plan: Reflection with colleagues

Analytic Methods Supporting Research Questions

Multi-Level Value-Added Models of Academic Growth

This research study uses a statistical modeling approach to estimate the academic progress of students. Specifically, multi-level (hierarchical) linear regression techniques (Raudenbush & Bryk, 2002) are employed to generate predictions of student achievement which are then compared to actual achievement levels. These *residual* difference scores (actual minus predicted) are interpreted as the *value-added* resulting from intervening instruction over the course of the school year (McCaffrey et al., 2003; Amrein-Beardsley, 2008; Darling-Hammond, Amrein-Beardsley et al., 2012; Braun, 2005). All of the statistical models were specified and run by this researcher-practitioner using the Mixed (Linear) Models procedure within Version 21 of the Statistical Package for the Social Sciences (SPSS) and Version 6.03 of HLM: Hierarchical Linear and Non-linear Models for Windows.

Value-added regression models have been increasingly used to examine the instructional effects of classroom teachers (McCaffrey et al., 2003; McCaffrey & Hamilton, 2007; Braun 2005; Darling-Hammond, Amrein-Beardsley et al., 2012; Corcoran, 2010). McCaffrey et al. (2003) state that value-added models "… attempt to estimate how much teachers or schools add to the achievement of entering students …" (p. 2). However, VAM may be more generally described as an analytic method that uses statistical regression techniques to predict an outcome of interest based on its correlation with other determining variables. In their seminal text *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Cohen and Cohen (1983) describe the general class of regression techniques as "… a highly general and therefore

very flexible data-analytic system that may be used whenever a quantitative variable (the dependent variable) is to be studied as a function of, or in relationship to, any factors of interest (expressed as independent variables)" (p. 3). The *independent variables* to which the authors refer are also often called covariates and models incorporating covariates are sometimes referred to as covariate adjustment models (McCaffrey et al., 2003).

Regression models are essentially correlation models because they estimate the degree to which covariates (independent variables) are correlated to the dependent variable (Cohen & Cohen, 1983). However, there is an important distinction between regression and correlation. This distinction lies in the general assumptions made on the dependency between variables. Indeed, correlation models make no a priori assumption on directional causality. In this way, correlation models express a simple symmetric association between variables. In contrast, regression techniques require the researcher to explicitly express one or more variables as a being dependent on, or determined by, another set of variables. Here, the researcher is interested in predicting the dependent variable from a set of independent variables (Stevens, 1996). Many times, application of value-added (regression) models in educational settings attempt to do this by employing one or more covariates that are believed to explain (or *predict*) student achievement. In this regard, value-added models attempt to isolate instructional impacts by controlling for numerous non-instructional factors believed to affect student learning (McCaffrey et al., 2003; McCaffrey & Hamilton, 2007; Betebenner, 2008; Amrein-Beardsley, 2008).

To place this study's use of value-added models in context, it is useful to review the basic formulation for deriving estimates of student academic progress. To do so, first consider a simple model form of linear regression using one covariate:

$$Y_i = \alpha + \beta_1 X_i + \varepsilon_i$$

where Y_i represents the dependent variable, X_i is the independent (covariate) variable posited to correlate with Y_i, and α and β_1 represent statistically estimated parameters that describe the characteristic of the linear relationship (here, α indicates the intercept and β_1 the slope of the regression function). The ε_i term represents errors of prediction – the difference between the (actual) observed value of Y_i and the value estimated from the regression equation for a given value of X_i. This concept is depicted in Figure 12.

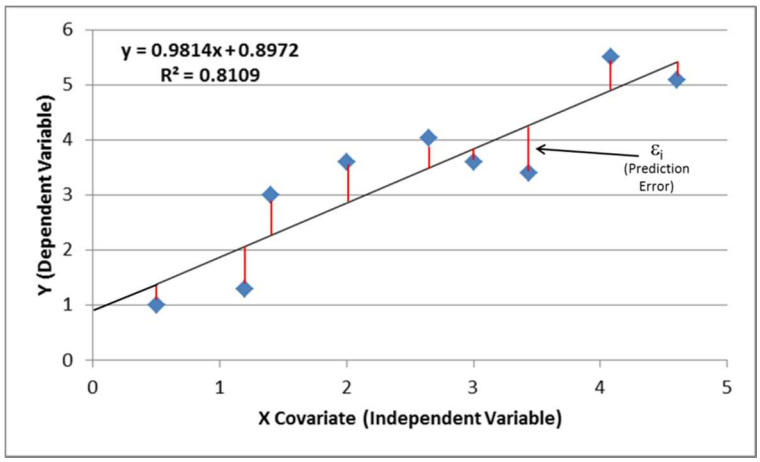


Figure 12. An example of a regression line representing the best fit between one covariate and the dependent variable.

Figure 12 depicts the relationship between two variables, Y and X for nine cases. The (regression) line reflects the equation which minimizes the collective error (ε_i) between observed (actual) and predicted values of Y given each value of X. For this reason the regression equation is often referred to as the *best fit line* because there is no other linear relationship which minimizes the discrepancy between the estimated and predicted values of Y (Cohen & Cohen, 1983; Stevens, 1996). Indeed, it becomes the task

of the scientist to develop models that minimize these prediction errors within the analytic context under investigation. Stevens (1996) writes

> … the *linear combination* of the *x*'s which is maximally correlated with *y* is sought. Minimizing the sum of squared errors of prediction is equivalent to *maximizing* the correlation between the observed and predicted *y* scores. This maximized Pearson correlation is called the *multiple correlation*. (p. 72)

Similar to examining the association (simple correlation) between two variables (*r*), the higher the value of a regression model's multiple correlation (R), the lower the predictive error. The most common computational representation of this is the squared multiple correlation coefficient (R^2; Cohen & Cohen, 1983; Stevens, 1996). R^2 represents the proportion of total variation in Y that is accounted for by the independent (covariate) variables (Cohen & Cohen, 1983; Stevens, 1996). Higher values indicate improved prediction capabilities and lower prediction error. In the simple example presented above, the $R^2 = .8109$, indicating that the approximately 82% of the variation in Y may be explained by the model's equation.

In an educational value-added context, the approach is to account for as much non-instructional background effects as possible (maximizing R^2, minimizing prediction error) such that deviation between predicted and actual scores are isolated to the intervening instruction (McCaffrey et al., 2003; Braun, 2005; Amrein-Beardsley, 2008). In this regard, value-added models often employ a variety of student background characteristics (i.e., English Language Proficiency, Special Education Status, poverty indicators …) and prior academic achievement in order to predict current period test scores. A simple example of this would be a regression model that uses a student's prior achievement to predict current test score. This could be expressed as:

$$TS_i^{Current} = \alpha + \beta_1 (TS_i^{Previous}) + \varepsilon_i$$

where $TS^{Current}$ is the student's current test score and $TS^{Previous}$ is the student's prior test score. As before, the ε_i term serves as a measure of prediction error and α and β_1 are the estimated equation parameters. As one might suspect, these types of model specifications may become quite complex depending on the context and purpose of the research activity, the consequences involved, and the need to be as precise as possible (McCaffrey et al., 2003; McCaffrey & Hamilton, 2007; Amrein-Beardsley, 2008; Amrein-Beardsley, 2012; Schochet & Chiang, 2010; Sanders, 1994, 1998)

As mentioned, the difference between a student's actual test score and his/her predicted score is called the *residual score* (Cohen & Cohen, 1983; Stevens, 1996; Lord & Novick, 1968). In the educational value-added context, the residual scores are interpreted to represent the effect that classroom instruction has on student learning after accounting for background characteristics and other intervening factors (McCaffrey et al., 2003; Amrein-Beardsley, 2008; Darling-Hammond, Amrein-Beardsley et al., 2012; Braun, 2005). In general, if a student's actual score is greater than predicted, a positive instructional effect is inferred. If the actual test score is less than predicted, then a negative instructional effect is inferred. As an example, consider a value-added regression model that includes multiple student background characteristics (X_{ij}) and prior academic achievement ($TS_i^{Previous}$). This hypothetical model takes the form:

$$TS_i^{Current} = \alpha + \beta_1 (TS_i^{Previous}) + \sum_{j=0}^{n} \beta j (X_{ij}) + \varepsilon_i$$

where i identifies the i^{th} student, j reflects a vector of (n) student background factors, and ε_i again represents the residual discrepancy between predicted and actual achievement. Estimating the model produces a predicted score for each student (i) which is compared to the actual outcome and the difference (residual) is interpreted as the *value-added*

produced by classroom instruction. A graphic representation of this is provided in Figure 13.

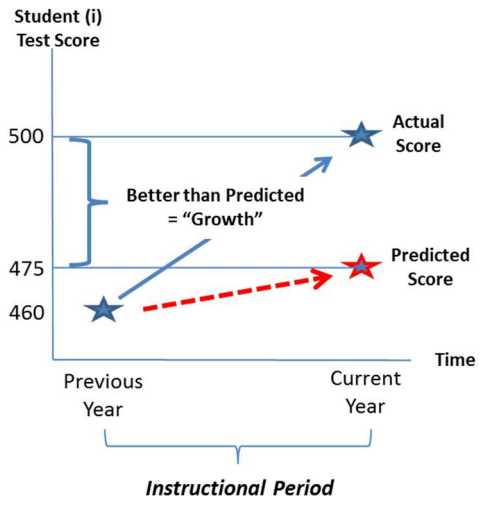


Figure 13. Depiction of predicted versus actual scores in a value-added context.

Here, the instructional effect is assumed to be positive because the student's actual score was higher than predicted after controlling for background factors and previous test history. This simplified example depicts the basis for generating the student academic progress (residual) scores used in this study. However, in actual practice, a more complex model was specified in order to obtain more precise estimates. The final model employed is discussed in the next section.

Value-Added Multi-Level Linear Regression Models (VAMLRM)

As indicated, this study uses multi-level value-added linear regression models (VAMLRM) to predict current year student achievement outcomes. A VAMLRM approach was adopted to improve the precision of the model estimates by accounting for the hierarchical (multi-level) structure of the data. Here, it is recognized that students are

geographically grouped within established classrooms, classrooms within grade levels, grade levels within schools, etc. This nesting structure is commonly encountered when conducting organizational research. However, this type nested data structure poses substantive methodological issues when attempting to obtain accurate estimates of organizational (instructional) effects (Raudenbush & Bryk, 2002; Subedi, Swan, & Hynes, 2011; McCoach & O'Connell; 2012). Specifically, if correlations in the dependent variable exist between individuals grouped within fixed organizational units (i.e., schools), non-trivial bias may be introduced into the regression estimates. That is, estimation of unbiased regression parameters require residual errors be independent, normally distributed, and have a constant variance (Raudenbush & Bryk, 2002; Cohen & Cohen, 1983; Stevens, 1996).

To understand this, consider students grouped within two contrasting schools: the first characterized by affluence (wealth), the second by high levels of poverty. Also, assume for the sake of argument, that these non-school wealth factors impact the achievement of the students within each school resulting in lower overall scores in the less affluent location and vice versa for the higher income community. Here, there will be a positive relationship of achievement between schools based on their level of affluence. As a result, the individual (student) residual errors generated by the regression model will not be independent of each other and result in biased parameter estimates. In addition, it is quite possible that the test score variance differs between the two groups further inflicting bias in the estimated parameters.

Stevens (1996) writes,

> … in the linear regression model it is assumed that the errors are independent and follow a normal distribution with constant variance. … The independence assumption implies that the subjects are responding independently of one another. … if independence is violated only mildly, then the probability of a type 1 error will be several times greater than the level the experimenter thinks he or she is working at. (p. 93)

In the example, the lack of independence is non-random and creates bias in regression estimates leading to greater measurement error, posing significant threats to model interpretation (Raudenbush & Bryk, 2002; McCoach & O'Connell, 2012). For models used in the evaluation of teacher effectiveness, such conditions might lead to inappropriate (policy) inferences impacting consequential decision making.

To correct for this, the nested nature of the data need to be explicitly accounted for. Indeed, Raudenbush and Bryk (2002) show that by explicitly modeling the nested (multi-level) structure inherent in such data, efficient and unbiased parameter estimates made be produced. This is why multi-level modeling frameworks are employed in many organizational research designs. This is the principal rationale for using multi-level modeling techniques in this study's validity examination of teacher evaluation.

Value-Added Multi-Level Regression Model (VAMLRM) Specification

For this study, student achievement is estimated using a two-level linear model specification where students are nested within schools. A graphical depiction of the nested data structure is provided in Figure 14.

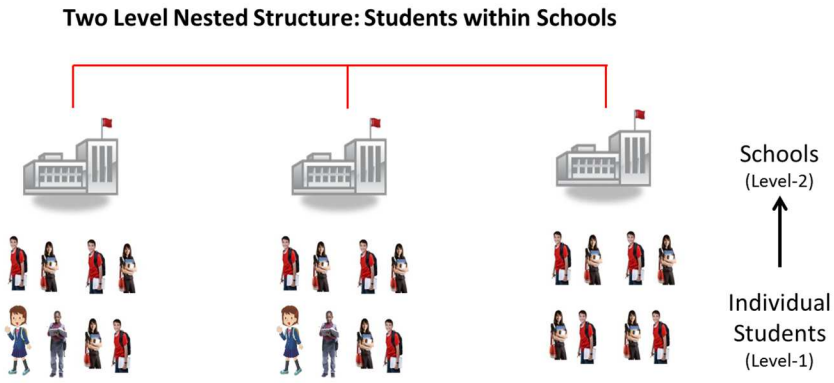


Figure 14. A depiction of a two-level data structure of students nested within schools.

In this design, Level-1 reflects student characteristics incorporating multiple background covariates and prior academic achievement. Level-2 attempts to make adjustments for differences in school/community factors. Here, the specification presumes that school (community) influences need to be explicitly accounted for in order to obtain more precise student-level outcome predictions. By doing so, more of the non-instructional variation in the achievement data is removed leading to more precise estimates of instructional effect.

The dependent variables used in the VAMLRM framework are current year reading and mathematics (scale) scores from the state's annual standardized assessment (AIMS) system. Separate models are estimated for each grade level (3 to 6) and subject area. The covariates incorporated into each level of the model are shown in Figure 15.

Level-1: Student Covariates:	
DAYS ABSENT	Number of days enrolled during the school year
LUNCH	Membership status in the National School Lunch Program (NSLP). This was a dummy coded variable where 1=Yes and 0=No.
GIFTED	Membership status in the district's Gifted Students program. This was a dummy coded variable where 1=Yes and 0=No.
SPED	Membership status in the district's Special Education Program. This was a dummy coded variable where 1=Yes and 0=No.
ELL	Whether or not the student was receiving services under the district's English Language Learner program. This was a dummy coded variable where 1=Yes and 0=No.
AIMS (2011)	Prior year Reading & Mathematics (total) scale scores on the state AIMS assessment
Benchmark (2012)	Reading & Mathematics (total) Scale Scores from the district's internal benchmark testing program. Scores from the current year (spring) assessment are utilized.
SAT10	Prior year Reading & Mathematics (total) scale scores on the state administration of the Stanford Achievement Test (SAT10). This measure is only used in the VAM model for current year grade 3 students to reflect their prior grade 2 achievement. The state of Arizona does not administer the AIMS assessment to grade 2 students.
Level 2: School Covariates:	
LUNCH	Percent of students participating in the federal free/reduced price lunch program

Figure 15. Covariates incorporated into each level of the model.

Table 3 provides a summary of the individual models estimated for this study and the mix of background factors used.

Table 3

Summary of Models Estimated by Grade Level and Subject

		Program Membership					Achievement History (Scale Scores)		
Subject	Grade Level	SPED (Yes/No)	Gifted (Yes/No)	ELL (Yes/No)	Lunch (Yes/No)	Days Absent (#)	AIMS Prior (-1)	SAT 10 prior (-1)	BM #3
Reading	3	✓	✓	✓	✓	✓		✓	✓
	4	✓	✓	✓	✓	✓	✓		✓
	5	✓	✓	✓	✓	✓	✓		✓
	6	✓	✓	✓	✓	✓	✓		✓
Math	3	✓	✓	✓	✓	✓		✓	✓
	4	✓	✓	✓	✓	✓	✓		✓
	5	✓	✓	✓	✓	✓	✓		✓
	6	✓	✓	✓	✓	✓	✓		✓
Science	4	✓	✓	✓	✓	✓	(Math)		✓

As shown, the grade three reading and mathematics value-added model utilized a different mix of prior year achievement factors. This is because current year grade three students did not take an AIMS test in grade 2. In Arizona, students in grade 2 are assessed using the Stanford 10 nationally norm-referenced test.

Each covariate has a presumed influence on current year achievement. Specifically, larger numbers of days absent is presumed to be negatively correlated with achievement. Designations related to English language proficiency (ELL) are expected to have negative association to test scores, as is the LUNCH variable since it is being used here as a proxy for the negative effects of poverty on learning outcomes. At the school level, communities reporting higher proportions of students receiving services under the National School Lunch Program are also presumed to report lower overall achievement.

In contrast, students having a gifted designation are expected to report higher achievement levels. Finally, higher past achievement scores are presumed to positively correlate with current year test scores. All the achievement scores are expressed in terms of scale scores obtained from Item Response Theory (IRT) scoring procedures used by the testing companies.

Level-1: Reflecting the covariates identified above, the Level 1 (student) specification takes the following functional form:

$$\text{Subject\_AIMS\_SS\_2012}_{ij} = \beta_{0j} + \beta_{1j}*(\text{DAYS\_ABSENT}_{ij}) + \beta_{2j}*(\text{GIFTED}_{ij}) + \beta_{3j}*(\text{SPED}_{ij}) + \beta_{4j}*(\text{LUNCH}_{ij}) + \beta_{5j}*(\text{ELL}_{ij}) + \beta_{6j}*(\text{Subject\_BB\_SS\_2012}_{ij}) + \beta_{7j}*(\text{Subject\_AIMS\_SS\_2011}_{ij}) + r_{ij}$$

where i represents the i^{th} student and j represents the j^{th} school. The $Subject\_BB\_SS\_2012_{ij}$ term represents the district's benchmark test scale score for student i in school j for the subject area tested (reading or mathematics) and $\text{Subject\_AIMS\_SS\_2011}_{ij}$ reflects the same student's 2011 AIMS subject area scale score. The β_{0j} through β_{7j} terms reflect the estimated model parameters while r_{ij} is the level 1 error term and is assumed to be well behaved and normally distributed with mean of zero and variance σ^2.[5] That is, $r_{ij} \sim N(0, \sigma^2)$. Finally, the intercept term β_{0j} represents the expected current year AIMS scale score when all covariates are assumed to be zero

Level-2: The Level 2 (school) specification incorporates adjustments on the constant term for the proportion of students participating in the National School Lunch Program (students receiving free or reduced priced lunch). This variable is included as a proxy indicator for community wealth. The explicit Level 2 functional form is as follows:

[5] Test of conformance to model assumptions were carried out for all estimated specifications. No substantive violations were apparent.

$\beta_{0j} = \gamma_{00} + \gamma_{01}*(LUNCHPCT_j) + u_{0j}$
$\beta_{1j} = \gamma_{10}$
$\beta_{2j} = \gamma_{20}$
$\beta_{3j} = \gamma_{30}$
$\beta_{4j} = \gamma_{40}$
$\beta_{5j} = \gamma_{50}$
$\beta_{6j} = \gamma_{60}$
$\beta_{7j} = \gamma_{70}$

where $LUNCHPCT_j$ represents the percent of students in school j participating in the Federal NSLP. The u_{0j} term represents the random error associated with the school-level effects [$u_{0j} \sim N(0, \sigma^2)$] and γ_{00} is the grand mean of current year AIMS scores across all schools. By combining the two level equations, the mixed multi-level model becomes:

$$Subject\_AIMS\_SS\_2012_{ij} = \gamma_{00} + \gamma_{01}*LUNCHPCT_j + \gamma_{10}* DAYS\_ABSENT_{ij} + \gamma_{20}* GIFTED_{ij} + \gamma_{30}* SPED_{ij} + \gamma_{40}* LUNCH_{ij} + \gamma_{50}* ELL_{ij} + + \gamma_{60}* Subject\_BB\_SS\_2012_{ij} + \gamma_{70}* Subject\_AIMS\_SS\_2011_{ij} + u_{0j} + r_{ij}$$

This two level structure is referred to as a random intercepts model where the (school) slope effects remained fixed and the intercept term randomized (Raudenbush & Bryk, 2002, p. 111).

Computing Value-Added Measures of Instructional Effectiveness

VAMLRMs were used to generate current year AIMS reading and mathematics residual scores for students in grade three through six. In order to compute classroom level measures, student residual scores were linked to teacher/course/classroom assignments based on student identification numbers maintained in the testing and course schedule databases. Once linked, median student score residuals were computed based on the AM Attendance (homeroom) course name. In the district, classes in grades three to six are self-contained, indicating that the homeroom teacher instructs all subjects to the same group of students throughout the day. It was decided to use median rather than

mean aggregates after examining variance measures of residual scores within classrooms. The presence of extreme values in some classrooms suggested that utilizing median rather than mean value calculations was warranted.

Measures of Value-Added Model Adequacy

For each estimated VAMLRM, a variety of examinations of model adequacy were undertaken. These include the following: (1) generation of overall model fit statistics, (2) comparison of covariance parameter estimates, and (3) examination of multi-level model assumptions.

Model Fit Statistics: AIC (Akaike's Information Criteria) statistics are useful for comparing alternative model specifications and for inspecting whether a particular specification form improves on the modeling activity. For AIC, smaller values indicate improved models. Cavanaugh (2009) describes the use of AIC statistics as a model selection criterion

> … that assesses the propriety of a fitted model by gauging how well the model balances the competing objectives of conformity to the data and parsimony… The smaller the value of the criterion, the better the fitted model balances these objectives… Given a set of fitted candidate models, the model corresponding to the minimum value of the criterion is preferred. (Slide 33)

Mazerolle (2004) provides general rules of thumb for interpreting the change in AIC values between two competing models. He suggests that changes in AIC less than two indicates "… substantial evidence…", while values larger than two indicate progressively less evidence.

Pseudo R^2 Statistic: The method of computing model fit for regression techniques based on ordinary least squares is not appropriate for multi-level models. OLS techniques compute model fit (R^2) based on minimizing prediction error of the fitted model based on

a single error term. However, the structure of multi-level models identifies more complex error structures. In this regard, iterative maximum likelihood procedures are used to derive the various multi-level model parameters (Raudenbush & Bryk, 2002; McCoach & O'Connel, 2012) and these procedures do not permit computation of an R^2 statistic. Rather, multi-level models decompose variance components at each level in the framework. Using these data, pseudo R-squared measures may be computed to reflect the amount of explained variance when comparing one model specification against another. Usually this is done by comparing a model containing covariates (the conditional model) against a model containing no covariates (the unconditional model).

Covariance Parameter Estimates: Multi-level models generate variance parameters for each model component. These parameters are characterized by a Chi-Square distribution. Tests of significance may be undertaken for each component, assessing conformance to hypothesized values and deviation from zero. For example, it is desirable for all level-2 group variance to be accounted for in the model. The level-2 variance component may be tested to see if it is statistically different from zero or whether additional error variance still needs to be modeled. In addition, a variety of Inter Class Correlation (ICC) and variance reduction statistics may be computed to assess the degree of variation accounted for between model components.

Multi-Level Model Assumptions: Multi-level models imposed similar assumptions as those required by ordinary least squares (OLS) regression models (Raudenbush & Bryk, 2002; ATS Statistical Consulting Group, 2013; Templin, 2008; Stevens, 1996). However, because multi-level models have more complex error structures and level dependencies, additional assumptions are imposed. The examination

of fidelity to these assumptions is important for determining adequacy of results, especially if the outcomes are to be used in high stakes, consequential settings as present in teacher evaluation systems. Tests of the following assumptions are run for the multi-level model used in this study.

(1) *Linearity*: The relationship between dependent and independent (interval) measures are approximately linear and supports use of linear models to represent the correlation structure of the data. This condition is assessed by graphically examining the scatterplot and associated correlation between independent and dependent variables.

(2) *Normality*: This condition requires that estimated model residuals are normally distributed with a mean of zero and variance σ^2 [$r_{ij} \sim N(0, \sigma 2)$]. In addition, the distribution of the standardized residuals (Z_{ij}) is assumed to be $Z_{ij} \sim N(0, 1)$. Adherence to this condition may be assessed by examining the distribution statistics of the estimated residuals including review of residual histogram and P-P Plots of cumulative density.

(3) *Homogeneity of Variance*: This condition assumes that the variance components are equivalent across all level-2 groups and that the level-1 variance is not heteroscedastic (i.e., the degree of variance changes across predicted values of the dependent variable). This may be assessed by examining residual plots against predicted values as well as computing residual variance measures within bounded groups of the predicted or dependent variables.

Factor Analytic Techniques of Content Validation

An important line of inquiry for any validity study is the degree to which observed measures align with the theoretical construct under investigation. AERA et al. (1999) reflect that the term *construct* has historically been defined as "… characteristics that are not directly observable, but which are inferred from interrelated sets of observations" (p. 5). However, the authors note that this causes confusion, and propose a broader perspective by defining construct as "… the concept of characteristic that a test is designed to measure" (p. 5) and that "… it is always incumbent on a testing professional to specify the construct interpretation that will be made on the basis of the score or response pattern" (p. 5). These perspectives are retained in the 2014 edition of the *Standards* (AERA et al., 2014, p. 11). In this regard, factor analysis provides researchers an analytic tool kit for examining the validity of these interpretations based on the correspondence between empirical measures and the theoretical construct they are purporting to measure.

A Priori Construct Articulation

For this study, district's policy makers have adopted a very clear and direct interpretation of the observational component of teacher's instructional practice. As discussed in Chapter 1, the district's teacher evaluation system utilizes Danielson's Framework for Teaching (FFT) as the basis for evaluating the professional practice (PP) of classroom teachers. Here, evaluators (observers of instructional activities) assign ratings to each of 22 behavioral elements organized across four domains. Each of these rating categories and related domains are explicitly delineated with unique, descriptive, traits and characteristics, replete with exemplars, behavioral artifacts, and graduated

levels of proficiency. Evaluators, trained in the intricacies of the evaluation system, collect evidences aligned to each element and use pre-defined criteria (rubrics) to assign numerical (integer) scores between 0 and 3, the higher the score the better the performance. By adding the scores across 22 elements, an overall performance score is attained. The range of total possible scores is 0 to 66. In addition, summing element scores within each sub-domain provides an additional level of detail. And since each element is characterized by a very explicit (unique) behavioral trait, examination of performance at the individual item level on the FFT is easily accommodated.

The organization of the FFT and its associated rating scales is purposeful, descriptive, and encourages a wide range of reflection and inference. Indeed, Danielson (2007) writes in the introduction to the second edition of *Enhancing Professional Practice, A Framework for Teaching*,

> … the success of the framework is a reflection … of both the recognition of the vital importance of high-quality teaching and an awareness of its complexity … that combination of factors has led both practitioners and policy makers to embrace a definition of teaching that is simultaneously clear and succinct … and respectful of the intricacies of the work. (p. v)

Thus, it is not without intent that the FFT attempts to provide implementers with a very clear definition of the Teacher Instructional Quality (TIQ) construct. Indeed, this is what it was designed to reflect. Danielson (2007) states:

> Each of the framework's four domains refer to a distinct aspect of teaching … the components within each domain form a coherent body of knowledge and skill that can be the subject of focus independent of the other domains. (p. 26)

Arguably, this perspective, if taken at face value, provides permission for practitioners to utilize and infer sub-scale information for the purpose of reflection, inference, and consequential action in the context of the evaluation process. However, for the researcher

the questions of interest become *To what degree do the empirical data support this construction*? *Is there evidence that, indeed, the data (ratings) substantively identify and distinguish instructional performance according to the behavioral components outlined by the FFT? Is it justified to differentiate specific behavioral traits based on the element/domain ratings and then imposed action plans targeted to these areas?* In essence, *does the empirical data align to the theoretical structure proposed by the framework and is there evidence to support using such data to make inferential claims on instructional quality*? All of these questions concern the broad context of construct validity. As such, any examination of validity evidence should include study of the structural integrity of the theory (AERA et al., 1999, 2014; Cronbach & Meehl, 1955; Messick, 1989a).

AERA et al. (2014) discusses the alignment between empirical evidence and the posited theoretical structure of a measure by stating

> … the conceptual framework for a test may imply a single dimension of behavior, or it may posit several components that are expected to be homogeneous, but that are also distinct from each other … the extent to which item interrelationships bear out the presumptions of the framework would be relevant to validity. (p. 16)

This, then, is the basis for investigating the content representation of the FFT implementation where the essential question is *To what degree do empirical ratings of teacher PP correspond with the theoretical FFT construct*? Factor analytic techniques provide one method of investigating this question.

Factor Analysis - Context

Factor analysis refers to a class of analytic techniques that identify the pattern of item correlations among variables (Sullivan, 1979; Kim & Mueller, 1978; Ferguson & Takane, 1989). While computationally it is simplistic in its basic approach, factor

analysis is conceptually tied to the examination of theoretical constructs. More specifically, factor analytic techniques may be used to investigate the underlying structure in observed data for the purpose of interpretation and/or comparing data to a construct of interest. In this way, factor analytic techniques allow the researcher to explore meaning and/or confirm theory. The former is often referred to as exploratory factor analysis (EFA) while the latter confirmatory factor analysis (CFA; Kim & Mueller, 1978; Corbell, Osborne, & Grable, 2008; Webber, Rizo, & Bowen, 2012).

In EFA, a priori assumptions are not made about the unobserved (latent) constructs or factors. Rather, the researcher is interested in uncovering the factor structures from a larger set of test items. EFA assesses the strength of the item correlations to reveal these structures. Depending on how the items align themselves, the researcher assigns interpretive descriptions to the revealed factors based on the nature of the correlated items. For example, consider a test of mathematics that contains six items. The researcher is interested in understanding whether all six items are measuring mathematics ability. The first three items represent numerical computations and the last three are context-based word problems. The test is administered to a large group of students and item correlations are computed. Figure 16 reports these hypothetical values.

	1	2	3	4	5	6
1	1					
2	0.87	1				
3	0.74	0.82	1			
4	0.11	0.20	0.23	1		
5	0.23	0.17	0.19	0.90	1	
6	0.08	0.10	0.22	0.89	0.88	1

Figure 16. Example math test item correlations.

In the example, all of the first three (numerical) items are highly correlated to each other. Similarly, the last three (word-based) items are strongly correlated. However, none are strongly correlated outside of their item type. Here, EFA uses the strength of the correlations to (in this example) extract two primary (unobserved) factors. Upon review of the correlated items, the researcher might assign a descriptive interpretation to each factor. The first might be "Numerical Ability" and the second "Word Problem Ability. Regardless, there is some evidence that the test does not measure a single construct. Rather there are two independent underlying factors reflecting unique characteristics. This might suggest that the test be scored against two sub-scales instead of a single total score. Figure 17 diagrams the hypothetical factor structure of the six item math test.

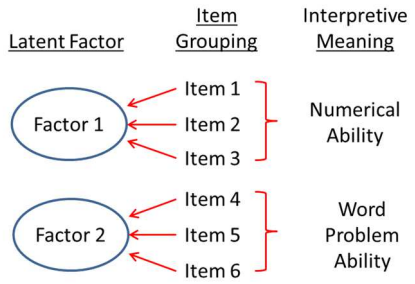


Figure 17. Hypothetical factor extraction from a six item math test.

In contrast to the above, sometimes the theoretical structure of the latent construct(s) is pre-specified such as with the Danielson FFT framework. Here, the FFT structure is a priori composed of four sub-domains with specific elements aligned to each. Figure 18 reports the FFT item-factor structural alignments.

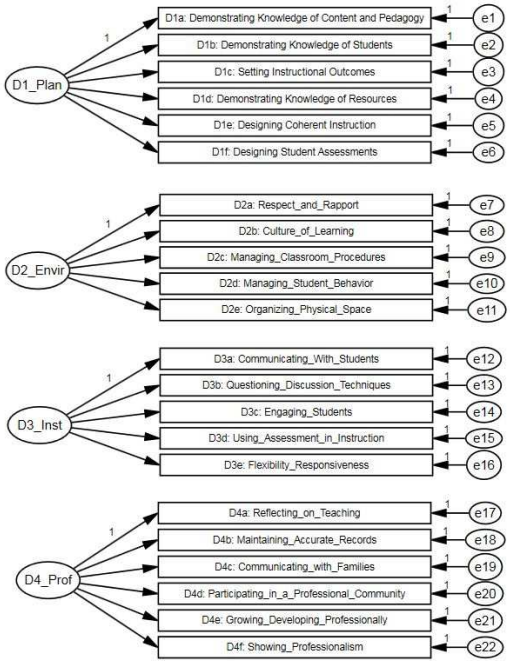


Figure 18. Rating element alignments for the Danielson Framework for Teachers.

Domain 1 is described as *Planning and Preparation* and is composed of six rating elements. Domain 2 is Classroom Environment composed of five elements. Domain 3 is Instruction made up of five elements. And Domain 4 is Professional Responsibilities comprise of six elements. Given the fixed structure, the researcher may choose to run a confirmatory factor analysis (CFA). Here, the researcher specifies the item-factor alignments beforehand and then forces the factor analysis to compute the item-factor correlations using these alignments. If the resulting forced item-factor correlations are weak, evidence exists that the theorized structure may not be present. In this case the

researcher may begin to search for new factor structures and re-interpretations of the theoretical model by adjusting the item-factor alignments.

For this study, both EFA and CFA investigations are conducted on the 22 FFT element ratings for classroom teachers. The EFA analysis was conducted in Version 21 of SPSS using the FACTOR procedure. The CFA analysis is run within the Mplus (v.7) software application. The purpose of the EFA analysis is to examine the factor structure inherent in the observed evaluation data and compare that solution to the one specified by the Danielson framework. In contrast, the CFA analysis starts with the assumption that the data is already organized according to the theorized factor structure and then tests conformance of the data to that assumption.

The EFA estimation is conducted using a principal components extraction process based on the item correlations found in the empirical data. Two forms of factor rotation, Varimax and Direct Oblimim, are performed to assess the stability and interpretation of the final solution (Sullivan, 1979; Kim & Mueller, 1978; Ferguson & Takane, 1989; Brown, 2009; Conway & Huffcutt, 2003). Here, factor rotation attempts to simplify the underlying structure such that each is defined by a subset of highly correlated items (Kim & Mueller, 1978; Conway & Huffcutt, 2003). The goal of rotation is to make the items loading onto each factor clearer and more pronounced to support interpretation (Brown, 2009).

However, determining which method of rotation is best is the subject of ongoing discussion amongst researchers (Brown, 2009; Conway & Huffcutt, 2003). Varimax is an orthogonal rotation method which "… attempts to maximize the variance of squared loadings on a factor … to produce some high loadings and some low loadings on each

factor" (Conway & Huffcutt, 2003). However, orthogonal rotations assume that the extracted factors are uncorrelated. In contrast, the direct Oblimim approach represents an oblique rotation which removes the assumption of zero factor correlation. Citing the literature, Conway and Huffcutt (2003) argue that " … if factors are really correlated (a likely situation), then orthogonal rotation forces an unrealistic solution that will probably distort loadings away from a simple structure, whereas oblique rotation will better represent and produce better simple structure" (p. 153). Given the preeminence in the use of Varimax as a method of rotation found in the literature (Conway & Huffcutt, 2003), both orthogonal and oblique approaches are used in this study as a method to assess the consistency and agreement of factor structures.

For this study EFA factor extraction criteria are initially based on eigenvalues greater than one. However, it has been suggested that these criteria alone may not produce an accurate assessment of the true underlying structure (Conway & Huffcutt, 2003). In this regard, review of associated scree plots and variance components are used to examine/confirm final extraction solutions including adjustments to the extraction criteria by specifying alternative numbers of fixed factors in the estimation process. Additionally, the Kaiser-Meyer-Olkin (KMO) and Bartletts Test of Sphericity (Chi-Square) statistics are used to assess characteristics of the data set. KMO tests whether the correlations among the items are too small to support an EFA approach. Values of the KMO should be above .6 indicating sufficient interdependence (ATS, 2012). Similarly, Bartletts Test is formed as a chi-square (χ^2) statistic assessing whether or not the correlation matrix is an identity matrix (i.e., diagonal values equal to 1 and off-diagonal values of 0). The null hypothesis assumes the correlation matrix is no different from an

identity. Significant values of χ^2 lead to rejection of the null hypothesis, suggesting the data may be applied to factor analytic techniques. Final EFA factor solutions and interpretations include review of communalities, factor loadings after rotation, and related variance components are discussed in Chapter 4.

The CFA analysis is run within the Mplus (v.7) software application. Here, the purpose is to explicitly model the structure specified by the Danielson FFT framework and assess the evaluation data's alignment to that theoretical structure. The CFA structural model including specification of factor covariance's and unique (element) error terms is depicted in Figure 19.

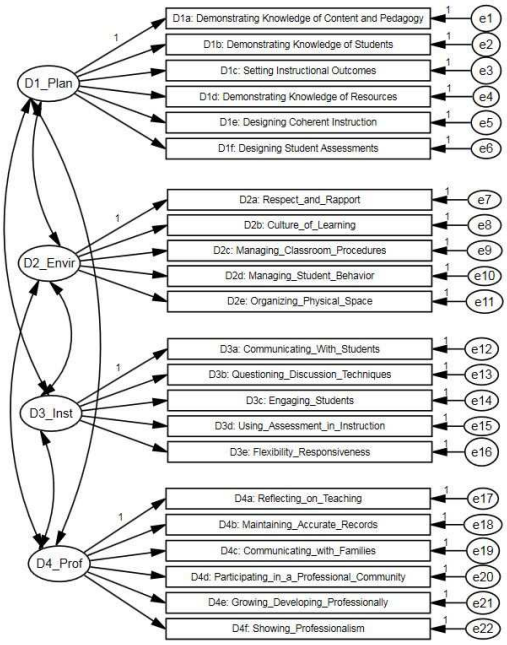


Figure 19. CFA structural representation of the Danielson FFT framework.

The CFA model is estimated using maximum likelihood procedures which employ iterative processes to derive parameter values (correlation, variances) that most closely match that of the specified structure. To assess model fit, χ^2 measures are referenced to test whether the covariance structure of the data matches that implied by theory (Kim & Mueller, 1978; DeCoster, 1998; Webber et al., 2012). In addition, comparative model fit (CFI) and root mean squared error residual (RMSER) statistics are referenced. CFI values at or above .95 and RMSER values below .05 are considered indicators of good model alignment (Webber et al., 2012; Arbuckle, 2005). Review of parameter estimates such as factor loadings (correlations), variances, factor covariance, and the lack of multiple alignments of items to factors also served as indicators of model fit.

For both EFA and CFA analysis, review of final factor scale reliabilities serve as measures of internal consistency which support inferences of construct adequacy. Factor scale reliabilities based on Cronbach's alpha (α) are computed for items defining each factor. Kline (2011) suggests adopting the following general rules for interpreting α within a factor analytic context: $\alpha \geq .70$ indicates adequate reliability, $\alpha \geq .80$ indicates good reliability, and $\alpha \geq .90$ indicates excellent reliability (p. 70). Arguably, suitability of the scale reliabilities for making consequential inferences depends on the context of the research (Webber, Rizo, & Bowen, 2012). In the context of this study, assessment of scale reliabilities associated with teacher evaluations seems to warrant higher standards of adequacy.

Content Validity Index

For the observational component of the district's teacher evaluation system, the professional practices of teachers are assessed using Danielson's FFT rating instrument. The instrument identifies 22 individual elements for which Principals/evaluators assign a rating value ranging from 0 (lowest) to 3 (highest). Ratings from each item are then added together to form an overall instructional quality score. All items are weighted equally, assuming that each represents an important aspect of high quality instruction. In this regard, higher scores are interpreted to reflect more effective instructional practices. Taken on face value, use of the rating instrument assumes that each and every item provides important information regarding a teacher's professional competency.

Arguably, if the purpose of teacher evaluation is to differentiate instructional quality, evidence is required that the measurement instruments provide adequate representation of this construct. In this regard, measurement theory has historically referred to examinations of content representation as content validity (Lawshe, 1975; Cronbach & Meehl, 1955; Messick, 1989a; Embretson, 1983). Here, the focus is on assessing the degree to which test items sufficiently cover the domain of interest. AERA et al. (2014) incorporates construct representation as a component of validation study and notes

> ... construct underrepresentation refers to the degree to which a test fails to capture important aspects of the construct … [and] … important validity evidence can be obtained from analysis of the relationship between the content of a test and the construct it is intended to measure. (pp. 12 and 14)

Content evidence, therefore, contributes to the process of construct validation by examining the representativeness of measures to the underlying concept. Carmines and Zeller (1979) describe this as the "…extent to which an empirical measurement reflects a specific domain of content …" (p. 20).

To illustrate, consider that an elementary school curriculum specifies that students be able to perform *basic mathematical operations* defined as addition, subtraction, multiplication, and division. To assess the degree to which students have mastered these skills, a classroom teacher might construct and administer a test made up of numerous multiple choice items shortly after the appropriate period of instruction. In this context, higher total scores are interpreted to reflect higher degrees of mastery. In addition, examination of the individual item responses might provide insight on which skill areas students required assistance. However, such inferences are appropriate only if the test's assembly of item content suitably represent the curriculum definition of *basic mathematical operations*. If the test were to completely omit items for multiplication and division, the resultant scores would be only partially representative of the desired construct. Similarly, if the test included items not aligned the instruction (i.e. irrational numbers) the resulted scores would also misrepresent learning.

Carmines and Zeller (1979) outline three critical components for understanding content validation. First, the underlying construct must be fully defined with regard to its component makeup. In the example above, this would correspond to operationally defining basic mathematical operations as consisting of the four areas of numeracy. In the context of teacher evaluation, it involves articulating all factors characterizing the concept of instructional quality. Second, any measurement instrument must consist of a suitable representation (or sample) of the desired construct. For the test of basic mathematical operations, the items must provide adequate representations across addition, subtraction, multiplication, and division. Similarly for a measure of instructional quality, rating items must be representative of the underlying construct. Third, the items

themselves must display attributes that ensure quality of measure. That is, items must be well constructed and display high reliability. This researcher conceptualizes the elements discussed by Carmines and Zeller as an ordered sequence depicted in Figure 20.

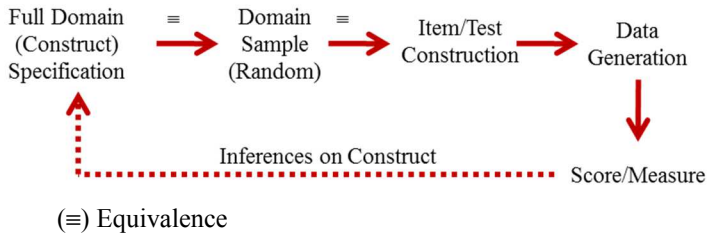


Figure 20. Carmines and Zeller's ordered sequence of test content validation.

In his chapter titled "Validity: An Evolving Concept" appearing in Wainer and Braun's text, *Test Validity*, Angoff (1988) reflects "… the demand for determination of validity is satisfied by a review of the test by subject-matter experts and a verification that its content represents a satisfactory sampling of the domain – an exercise in content validity" (p. 22). This approach is further supported in the *Standards for Educational and Psychological Testing (*AERA et al., 2014) which states "… evidence based on content can also come from expert judgments of the relationship between parts of the test and the construct. … These or other experts can then judge the representativeness of the chosen set of items" (p. 14).

Interestingly, this approach to content validation was operationalized by Lawshe in a 1975 paper published in *Personnel Psychology* titled "A Quantitative Approach to Content Validity*"* (1975). Here, Lawshe reflected on the relative lack of professional standards in establishing content validity of instruments used in employment decisions. At the time, Lawshe reflected that

> …until professionals reach some degree of concurrence regarding what constitutes acceptable evidence of content validity, there is a serious risk that the courts and the enforcement agencies will play the major determining role. Hopefully, this paper will modestly contribute to the improvement of this state of affairs. (p. 563)

This study adopts the Lawshe (1975) approach for assessing the content validity evidence based on experts' reviews of the Danielson FFT domain elements. Under the Lawshe approach, content experts evaluate the contribution that each test (rating) item makes to defining and understanding the construct. Specifically, Lawshe argues for use of a content evaluation panel composed of persons knowledgeable about the construct the test was designed to measure (Lawshe, 1975, p. 566). To do so, each panel member is provided a copy of all items on the test and asked to evaluate each using the following rating rubric: Is the skill or knowledge being measured by the item (*1) Essential to defining the construct, (2) Useful but Not Essential,* or *(3) Not Necessary* (p. 567). Responses from all judges are combined and the number indicating *Essential* is tallied. From this, a Content Validity Ratio (CVR) is computed for each item using the following formula:

$$\text{CVR} = \frac{N_e - \frac{N}{2}}{\frac{N}{2}}$$

where N_e is the number of panelists rating the item as *Essential* and *N* is the number of panelists (Lawshe, 1975, p. 567). Using this formulation, when less than one-half of the panelists rate an item as *Essential*, the CVR is less than zero; when one-half rate the item *Essential*, the CVR is zero, and when more the one-half rate the item *Essential*, the CVR is positive (the range of possible CVR values is -1.0 to +1.0). Lawshe (1975) suggests that items that are rated *Essential* by more than half the panelists have "… some degree

of content validity. The more panelists (beyond 50%) who perceive the item as *essential*, the greater the extent or degree of its content validity" (p. 567).

Lawshe's (1975) approach is not unlike a factor analytic metric to assess whether particulate items correlate (contribute) to an underlying construct. Here, the CVR forms a basis for selecting construct correlated items from a larger set of items that are most indicative of the underlying construct. In addition, Lawshe provides criteria for acceptance/rejection of item-construct alignment based on the number of rating panelists and a ninety-five percent level of confidence ($p = .05$). This information is provided in Table 4, below.

Table 4

Lawshe CVI Minimum Values (p = .05)

Number of Panelist	Minimum CVR Value
5	.99
6	.99
7	.99
8	.78
9	.75
10	.62
11	.59
12	.56
13	.54
14	.51
15	.49
20	.42
25	.37
30	.33
35	.31
40	.29

After the CVR is computed for each prospective item, a Content Validity Index (CVI) may be computed for the overall test. Lawshe (1975) defines this as the average CVR across all items. Lawshe states "… operationally [the CVI] is the average percentage of overlap between the test items and the job performance domain" (p. 568). However, Lawshe does not provide threshold criteria for this measure. The final judgment of whether or not an instrument has a sufficient number of highly aligned items is left up to the researcher.

Regardless of approach, evaluating content evidences is an inherently subjective exercise (Wilson, Pan, & Schumsky, 2012; Allen & Yen, 1979; Lawshe, 1975; AERA et al., 1999, 2014). Allen and Yen comment that "…content validity is established through a rational analysis of the content of a test, and its determination is based on individual, subjective judgment" (p. 95). However, speaking about the subjectivity of the CVR/CVI process, Lawshe (1975) reflects that

> … if [expert] panelists do not agree regarding the essentiality of the knowledge or skill measured to the performance of the job, then serious questions can be raised. If, on the other hand, the *do* agree, we must conclude that they are either "all wrong" or "all right." Because they are performing the job, or are engaged in the direct supervision of those performing the job, there is no basis upon which to refute a strong consensus. (p. 567)

As applied in this study, a content evaluation panel (subject matter experts) was identified including instructional support coaches ($N = 24$) and members of the Teacher Evaluation Committee ($N = 12$). These individuals possessed a thorough understanding of the Danielson FFT framework and the observational instrument used to evaluate classroom teachers. Each were asked to complete a Content Validity Assessment Questionnaire, rating each of the FFT's 22 elements on its importance for contributing to

identification of teacher's competence. Descriptions of Lawshe's original scoring categories were adapted to better align with the teacher evaluation activity. The full instrument is provided in Appendix C.

After completing the questionnaire, CVR measures were computed to assess adequacy of each item. An overall CVI measure was computed based on the number/percentage of items that meet criteria (Lawshe, 1975; Wilson et al., 2012).

Analytic Framework for Qualitative Investigations

As discussed, this study employs both quantitative and qualitative approaches for data collection and the assembly of validity evidence related to measures of the TIQ construct. As applied herein, qualitative information are gathered from individual stakeholders through interviews and group observation activities. The data collection and analysis approach aligns closely with the perspectives and techniques used by Grounded Theory (GT) as described in the writings of Glaser and Strauss (1967), Strauss and Corbin (1990), Corbin & Strauss (2008), and Saldana (2009).

Glaser and Strauss (1967) describe GT as "… the discovery of theory from data systematically obtained from social research" (p. 2) while Corbin and Strauss (2008) write that grounded theory is "… a specific methodology developed by Glaser and Straus (1967) for the purpose of building theory from data" (p .1). In this way, GT is an inductive approach that begins with data collection and then moves toward developing a broader understanding of the phenomena being studied. In contrast, deductive reasoning begins with an a priori premise and then collects data to assess its tenability. Glaser and Strauss (1967) describe GT from this perspective: "… our basic proposition is that generating grounded theory is a way of arriving at theory suited to its supposed uses …

[we] contrast this position with theory generated by logical deduction from *a priori* assumption" (p. 3).

A GT approach to understanding phenomena is also rooted in social constructivism (Creswell, 2009; Gergen, 2009). That is, a GT approach begins with the collection of data that is only loosely connected to the broad context of interest. The data is then reviewed and reflected upon, giving rise to initial understandings, relationships, and connections leading to further avenues for investigation and inquiry. Strauss and Corbin (1990) reflect that "… since phenomena are not conceived of as static but as continually changing in response to evolving conditions, an important component of the [GT] method is to build change, through process, into the method" (p. 5). Thus, as the process of data collection, review, and reflection evolves, the researcher continuously constructs concepts, themes, and connecting constructs. In this way, the researcher is constructing meaning from data. Glaser and Strauss (1967) comment that

> … concepts and theories are *constructed* by researchers out of stories that are constructed by research participants who are trying to explain and make sense out of their experiences and/or lives … [and it is] out of these multiple constructions, analysts construct something that they call knowledge. (p. 10)

In this regard, without safeguards, it might be argued that GT lacks suitable scientific rigor upon which to base findings. However, Strauss and Corbin (1990) reflect that "… grounded theorists share a conviction with many other qualitative researchers that the usual canons of 'good science' should be retained … [and that] these cannons include significance, theory-observation compatibility, generalizability, consistency, reproducibility, precision, and verification" (p. 4). They argue that qualitative researchers must protect these standards through the systematic processes they use to collect, analyze, and interpret data.

This is the perspective adopted for the qualitative investigations in this study. Here, selected research questions serve to define a broad category of inquiry (i.e. *What are stakeholder perspectives regarding the purpose and intended outcome of teacher evaluation? Does this perspective differ across stakeholder groups?*). Then a process of interview and observation are undertaken to collect stakeholder perspectives. At each step of the data collection, analysis of the information is conducted which then shapes subsequent inquiry by either altering the initial question, incorporating additional probes and dimensions, or defining new sources of data. The goal is to construct and report stakeholder perspectives and then reflect on how these perspectives influence and/or impact implementation and application of the organization's policy-driven teacher evaluation system.

Importantly, a caveat should be made regarding the qualitative methods employed in this study and those described by Glaser and Strauss (1967). For this study, findings emanating from a GT-adopted approach are not generalizable to a larger context. Rather, the intent of the study is to better understand the local (district) context, impact organizational process, and (hopefully) contribute to the larger body of knowledge regarding validity of teacher evaluation systems. In this context, the process is not technically giving rise to new theory (big "T"). The modes of data collection, coding, and examination borrowed from the GT toolkit provide a structured methodology to derive understanding that may be supported. In this way, the methodological procedures that Glaser and Strauss (1967), Strauss and Corbin (1990), Corbin and Strauss (2008), and Saldana (2009) describe are utilized to support the cannons of "good science" (Strauss & Corbin, 1990).

Qualitative Methods

Three forms of qualitative data collection methods were utilized in this study: individual stakeholder interviews, individual and group observation, and journaling. In all cases, semi-structure protocols were used to guide the initial activity. However, these protocols remained fluid and served only as guide posts. For example, to investigate selected (qualitative) research questions based on stakeholder interviews, a set of guiding questions were developed to ensure alignment to each research question and consistency across participants. These protocols were the most structure of the three qualitative activities. For group observation (workshops, meetings, discussions, trainings, and evaluation sessions), less structured protocols were enlisted. Here, the researcher entered into each session with a short list of constructs of which to remain aware but not be bound by. Similarly, for journaling activities, the researcher maintained a short list of constructs on which to reflect after recording personal impressions and reflections. For the observation and journaling activities, the protocols were written on the last page of the researcher's journal and referenced each time entries were recorded. In this way, the observation/journaling activity starts with open scripting and then reflections/notations made for each of the protocol areas as needed. All data collection instruments/protocols used in this study are been provided as Appendices. An outline of protocol dimensions is discussed below.

Stakeholder interview protocol. The protocols used for each stakeholder interview are presented in Appendix D. Each prompt was constructed to align to a specific supporting research question. All prompts received limited field testing during fall 2012. Here, informal interviews were held with stakeholders and discussed with

colleagues in informal settings. During early spring 2013, the interview prompts were utilized in three formal interview occasions with school principals as well as reviewed by two district office colleagues.

Audio recordings were made of all stakeholder interviews. The recordings were then transcribed by a professional transcription service. Where appropriate, artifacts in the form of documents and photographs were also obtained as part of the data collection process.

Observational protocol. Observation events were documented by the researcher using an open scripting protocol. The intent was to capture as much of the activity, setting, and topic discussion as possible for use in subsequent coding and interpretation. Scripting activities were augmented with researcher's notes, reflections, and areas for further follow-up and/or data collection. The observational protocol and reflection categories are provided in Appendix D.

Journaling protocol. The researcher maintained a personal journal during the course of the study. Journaling activities were intended to supplement understanding of all aspects of the research activity including post-interview/observation events, informal discussions with stakeholders, and the researcher's personal reflections, ideas, and notations. Entries were unrestricted and inclusive of any and all facets believed relevant to the analytic process. In his text *Qualitative Research: Studying How Things Work*, Stake (2010) writes "… here [in a researcher's journal] you should make notes about everything in the research: contact information, calendar, bibliographic references, risks; get it all in one place … put your ongoing speculations, puzzlements, and ponderings … write down your concern[s]" (p. 101). For this study, the researcher's journal was also the

primary data collection instrument used for documenting Research Question 4 - *How did the process of engaging in this action-research study impact the investigator as a scholarly researcher and organizational leader*?

For the journaling activity, a simple set of protocol categories were constructed as a reflective reminder and thinking prompts. As in the observation activity, these categories were meant to be referenced during and after reflective memoing so as to provide some type of minimally consistent structure over time. They serve as reminders and guidelines when completing journal entries. The journaling protocol is provided in Appendix D.

Admittedly, adherence to the use of every prompt, every time, for every entry, was difficult and sometimes problematic. However, the protocols attempted to provide structure to the data collection activity that was taking place over an extended period of time.

Qualitative Data Analysis Plan

Approach of qualitative data collection and analysis. As mentioned, a modified GT perspective was adopted for collection and analysis of the study's qualitative components. In this modified approach, initial data was collected using appropriate protocols. Immediately after a data collection event, review and initial analysis of the information was undertaken. In addition, first reflections from the data were documented. Results from this data analysis stream were "banked," meaning the learnings from the activity are added to an evolving set of findings and reflections. In addition, based on the initial analysis, adjustments/additions to protocols were made for use in the next round of data collection. If new sources or forms of relevant data became evident, these were

added to the data collection profile. Alternatively, some lines of inquiry were dropped. Central to this process was the need to reflect upon just-completed analysis (interviews/observations/journal entries). Strauss and Corbin (1990) state that "… in grounded theory, the analysis begins as soon as the first bit of data is collected … here, analysis is necessary from the start because it is used to direct the next interview and observations" (p. 6). This process is depicted in Figure 21.

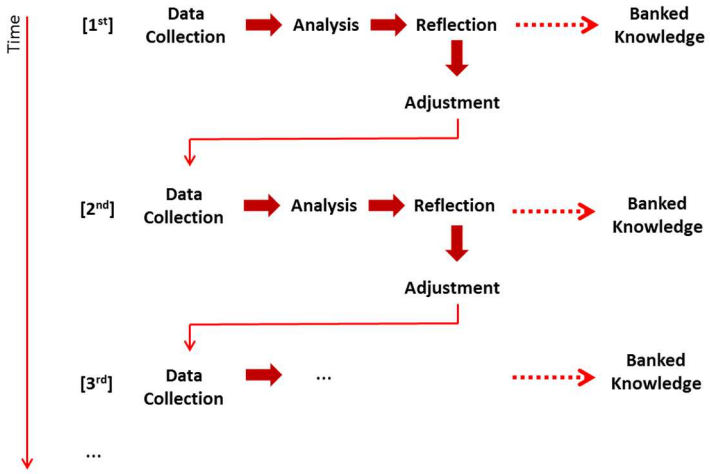


Figure 21. Qualitative [GT] data analysis stream: protocol adaptation and knowledge banking.

In this context, the concept of knowledge banking does not refer to data achieving or the listing of specific codes or memos on a specific data set. Rather, it reflects a running documentation of the researcher's inferences and conclusions on a specific research question emanating from the growing accumulation of information and evidence. To this end, the journaling activity proved very useful for documenting an evolving understanding of the evaluation context and thinking through the many sources of information.

Data analysis methods for qualitative information. All qualitative information in this study was analyzed using methods adopted from GT (Corbin & Strauss, 2008; Stake, 2010; Patten, 2012; Saldana, 2009). In this tradition, information is first reviewed with minimal annotation. Follow-up readings increasingly incorporate memoing techniques encouraged by Corbin and Strauss (2008) where the researcher free-thinks and free-associates about the dialog being examined. In this way, memoing is a first phase action aimed at providing early insights and generating initial concepts which stimulate analytic thought (p. 140). Following this, in vivo and descriptive (open) coding is used to begin identifying simple ideas, context, and dimensions (categories) seen in the data (Corbin & Strauss, 2008, p. 195; Stake, 2010, pp. 150-156; Saldana, 2009, p. 70). As the process progresses, more structured coding schemes are adopted, that is, previous codes begin to be increasingly utilized to identify like concepts. Axial coding is then applied to correlated concepts in order to derive connected themes and constructs and to bring the raw data into a more thematic framework. Finally, theoretical (thematic) code categories are constructed, serving as an umbrella construct for like code groups (Saldana, 2009, p. 163). For each step in the process (in vivo/open coding, axial codes, and thematic constructions), the researcher begins to assign descriptive interpretations to the emerging constructs, keeping cognizant of the inherent interrelationships that gradually define a broad conceptual framework. Figure 22 provides a graphic depiction of this coding process.

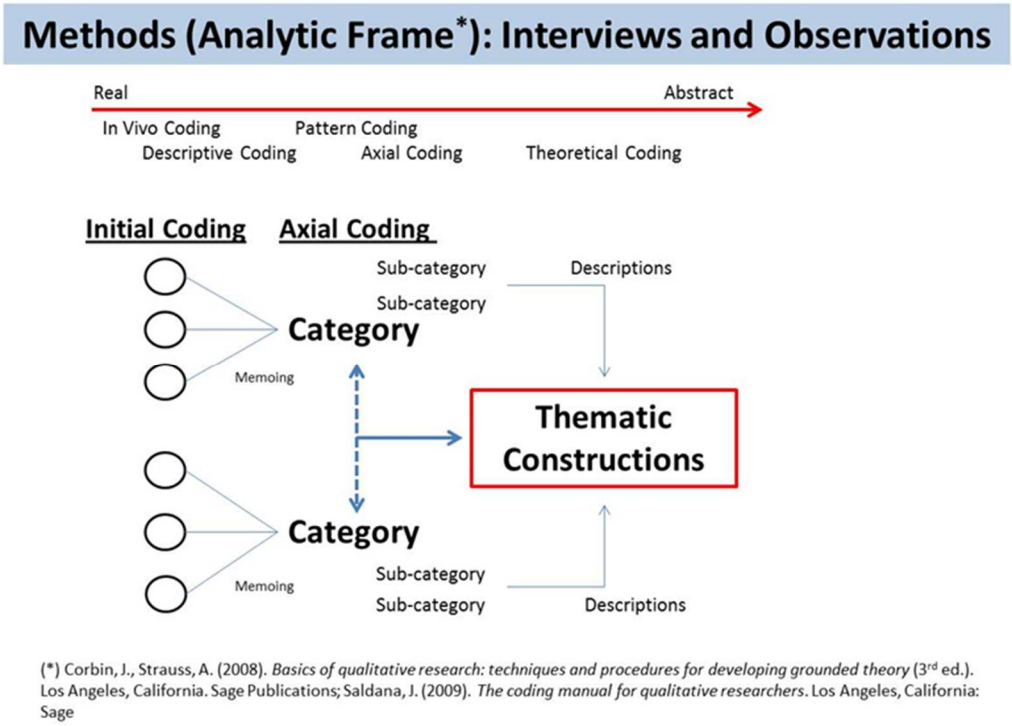


Figure 22. Generalized coding framework for analyzing qualitative information (interviews, observations, journal notations).

As mentioned all stakeholder interviews were captured using a digital audio recording device. The audio recordings were transcribed (verbatim) into Microsoft Word documents by a professional online transcription service. While coding the documents, this researcher listened to selected sections of the audio that appeared unclear or improperly transcribed.

Data quality methods. Ensuring the data quality standards of qualitative investigations was critical to establishing the validity of the resulting inferences made on the information. This study adopted a variety of data quality techniques to examine both the reliability and the validity of the collected data. Specifically, a mix of data, methods, and researcher triangulation, peer review, and member checking were employed

depending on the type of information, the participants involved, and level of the analysis. In this regard, these data quality activities align with perspectives offered by Patton (2012). Here, *data triangulation* refers to the collection of multiple sources of data (stakeholder groups) concerning the same construct. For most of the qualitative research questions being investigated under Primary Research Question 1 (validity evidences), similar interview prompts were utilized across stakeholder groups. In this way, a single data collection method was used to collect similar information from different (stakeholder) participants. Comparing responses between groups served as form of triangulation and confirmation of the derived constructs.

In contrast *methods triangulation* refers to the use of different types of methods to collect data from the same group of stakeholders (same construct of interest, same stakeholder group, using multiple data types). For this study, a number of supporting research questions assembled information using both quantitative and qualitative methods from the same stakeholder groups. Comparison of results from this approach provides a form of reliability evidence.

Researcher triangulation refers to employment of multiple researchers in data collection and analysis to check/protect against bias emanating from a single researcher's perspective. Patton (2012) notes that the use of researcher triangulation "… reduces the possibility that the results of qualitative research represent only the idiosyncratic views of one individual…" and that this helps assure the quality of analysis (p. 157)

As used in this study, researcher triangulation is closely tied to methods of *peer review*. For the qualitative analysis, this researcher employed the services of two professional colleagues to independently analyze and provide comparative reflection on

findings. Both colleagues were employed as research analysts within the school district, one of which was a doctoral student in at a local university. The other had background conducting applied program evaluation including collection and analysis of qualitative survey information. To conduct peer review, a subset of raw interview transcripts was given to each colleague. They were then asked to read/reflect on meaning and themes. The team then discussed comparative findings and interpretations. Consistency was assessed and documented, making note of any changes resulting from the collaboration.

In addition to the above, peer-consultation was utilized. Here, selected stakeholders were asked to read and discuss this researcher's interpretations, findings, and inferences regarding selected qualitative research questions.

Finally, a limited form of *member checking* was utilized. Interview transcripts were sent back to stakeholders, and they were asked to confirm the accuracy of the transcription, ideas, perspectives, and topics discussed. In addition, the individuals were provided opportunity to add, delete, and/or expand any part of the original discussion.

Sample selection of interview participants. Two primary sampling techniques were used in this study: stratified random sampling and purposeful criterion sampling (Plano Clark & Creswell, 2010; Creswell, 2009; Patton, 2012). Random sampling is a probabilistic method of selecting individuals (or units) from a larger population. It is used when the intent is to infer meaning to the larger population based on information obtained from the smaller chosen sample. The term *random* signifies that all members of the population have the same probability of being selected (Hinkle, Wiersma, & Jurs, 1994; Ferguson & Takane, 1989). An extension of this procedure is *stratified random sampling*.

This refers to random selection of individuals from pre-established groups or *strata* (Hinkle et al., 1994; Ferguson & Takane, 1989).

Purposeful criterion sampling (Patton, 2012) refers to the selection of participants based on pre-established, non-random, criteria. For example, in this study, principals were identified based on the site location of the randomly selected teachers. The desire was to allow comparison of perspectives between teachers and evaluators located in the same community environment. In addition, instructional coaches were purposefully (non-randomly) included due to their familiarity with the structural context of the evaluation framework. Similarly, three state-level policy leaders were purposefully identified based on their accessibility and close connection and understanding of the evaluation framework's legislative origins.

The importance of randomness in sample selection concerns matters of both validity and statistics. Plano Clark and Creswell (2010) note that "… the advantage of using probability sampling is that any bias that exists in the population should be equally distributed among the selected participants …" thus making inferences drawn from sample data applicable to the larger unobserved group (p. 183). However, Lord and Novick (1968) comment that

> … in scientific work one frequently draws a random sample, not because this is the most representative kind of sample, but because its statistical properties are well known, whereas the statistical properties of some possibly more representative, but subjectively chosen, sample would not be known. (p. 235)

Regardless, the validity of making inferences from such data is based on the premise that the observed sample is equivalent (in every aspect) to the larger (unobserved) population. Random selection attempts to ensure this type of equivalence (Plano Clark & Creswell, 2010).

As discussed, a number of research questions investigate the perspectives of classroom teachers using qualitative (interview) methods. However, it would not be possible to interview every teacher in the district, and if one could, it would be problematic to process that amount of information. In such cases, it would be advantageous to use sampling techniques that support the premise that inferences drawn from the data, even with a small sample size, might be representative of the faculty at large. Admittedly, qualitative data restrictions and time limitations allow for only a small sample of teachers to be chosen. Regardless, at the teacher level, random sampling techniques help bolster the validity of findings. In this regard, a stratified random sample of teachers was chosen based on the following criteria:

- All teachers in grades 3 through 6, teaching regular education, self-contained classrooms with 10 or more students were placed into one of three groups determined by the percentile location of their classroom's median value-added (residual) gain scores.
- The three value-added groups were defined as High (scores above the 90^{th} percentile), Average (scores between 45^{th} and 55^{th} percentile), and Low (scores below the 10^{th} percentile)
- Within each of these groupings (strata), three teachers were randomly selected. In the event that a teacher chose not to participate, another teacher from the group was randomly assigned. A total of nine teachers were initially chosen using this method.

Principal sample selection. Initial sampling of principals (evaluators) was non-random and based on the teacher sample. The principal of the sampled teacher's school was automatically included in the data collection design.

Policy maker sample selection. Out of five policy-level leaders, four were identified to be interviewed for this study. Selections were based on their decision-making role within the organization and their familiarity/involvement with the evaluation process.

Instructional growth coach (IGT) sample selection. All IGT members ($N = 24$) were identified to participate in the study because they worked directly with classroom teachers and possessed a high level of understanding of the evaluation process.

Evaluation committee member selection. All members of the district's Teacher Evaluation Committee ($N = 12$) were selected for participation based on their role in system design and implementation.

Chapter 4 provides a comprehensive review of the analytic procedures, data sets, and empirical findings. It is organized sequentially by primary and secondary research question. Care is provided to present details of the data sets and the associated analytic methods.

Chapter 4: Data Analysis

Chapter 4 presents a systematic analysis of the data collected for each primary and secondary research question presented in this study. As previously outlined, this study posits four primary research questions. The first question is further sub-divided into five component areas of inquiry. The primary questions and components are formulated as shown in Figure 23.

RQ1: To what degree does the validity evidence generated by the LEA's policy-directed teacher evaluation system support inferences of Teacher Instructional Quality (TIQ)?

Components of RQ1 include:

RQ1-A: Criterion Evidence

RQ1-B: Content Evidence

RQ1-C: Consequential Evidence

RQ1-D: Reliability Evidence

RQ1-E: Construct Articulation

RQ2: How does providing an evolving dialog on validity evidence influence organizational policy decisions regarding implementation of the LEA's teacher evaluation system?

RQ3. Third, to what degree have the perspectives and concerns of stakeholders been incorporated into, and influence, the system's implementation?

RQ4. How did the process of engaging in this Action-Research study impact the investigator as a scholarly researcher and organizational leader?

Figure 23. Primary questions and components.

Under each primary (and component) question, additional supporting research questions were advanced, characterized by specific data collections and analytic

methodologies. A complete listing of all primary, component, and supporting research questions is provided in Appendix B of this report. Figure 24 presents a graphical representation of the main inquiry areas investigated along with their associated analytic approach (quantitative and/or qualitative).

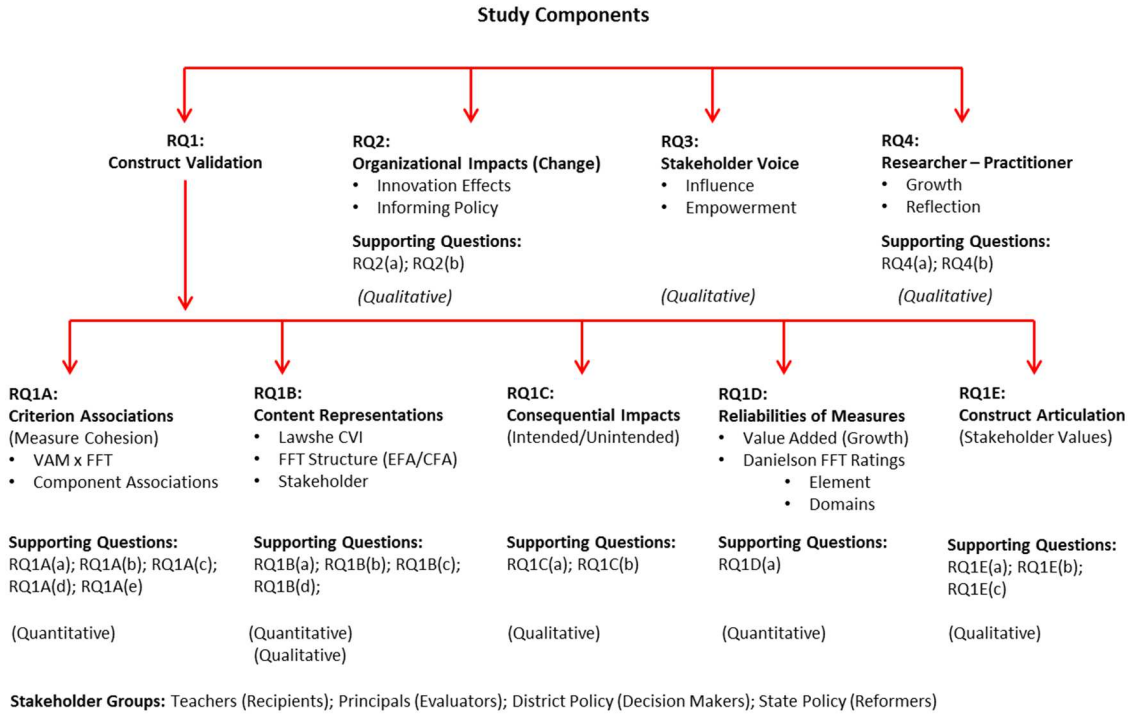


Figure 24. Study components.

Chapter 4 is organized into four main sections. Part 1 presents descriptive summary information for the quantitative data collections including various data selection rules and associated sample statistics. Part 2 discusses characteristics of the qualitative data collections including sample selections and descriptive attributes of the assembled information. Part 3 presents a sequentially organized analysis of each primary and secondary research question. Where necessary, details of the aligned analytic techniques

are reviewed as are any unique features of the data set followed by a collective summary of the findings.

Part 1: Descriptive Summary of Quantitative Data Collections

Background

During SY2012-13, the district evaluated a total of 1,335 certified staff based on the new state-mandated evaluation framework. However, the dataset included a diverse range of employment categories and instructional settings, with only a small proportion defined as self-contained general education classroom teachers. Discussed below, this non-equivalent mix of instructional contexts led to refinement of the sample and a reduction of the records retained for the analysis.

As previously noted, the state framework distinguishes teachers according to two major categories: Group A and Group B. Group A teachers are characterized as having a direct alignment between the content assessed on the state's standardized assessment (AIMS Reading, Mathematics, and/or Science) and the content actually instructed in the classroom. For this group, the achievement measures are presumed to provide a direct measure of instructional efficacy because the content taught aligns closely with the content assessed.

In contrast, Group B teachers lack this direct relationship. For example, art teachers are not instructing on the state's assessed standards in mathematics, reading, or science and no formal standardized measures of the art curriculum are available. However, under state policy, Art teachers must nevertheless be evaluated by standardized achievement measures. Arguably, this causes the linkage between a Group B teacher's

instructional efficacy to be less well-defined due to the misalignment between instructed and measured content.

In addition to the distinctions between Group A and Group B teachers, the district's evaluation system applied the same evaluation procedures to a variety of specialized instructional settings including some non-instructional job categories. For example, instructional growth coaches (IGTs) serving in supporting professional development and mentoring roles were evaluated using the same procedures and rubrics applied to general education teachers. In addition, special education teachers serving small groups of special needs students were evaluated using identical procedures, rubrics, and achievement criteria.

It is argued that this heterogeneous mix of instructional settings introduces an unknown amount of construct-irrelevant variance, potentially obfuscating inferences made from the information (Haladyna & Downing, 2004). That is, variance in the measures may be due to differences in job descriptions and instructional settings not associated with substantive differences in instructional competency. Since the purpose of this study is to explore the construct validity of a system intended to assess the instructional competency of classroom teachers, it is important to apply a working definition to the context.

To this end, this study focuses on Group A general education teachers characterized by the availability of achievement measures which are directly aligned to the content instructed within the classroom. This definition restricts the data set to elementary education teachers teaching in general education settings for the full academic

year that instructed in content areas assessed by the state's standardized achievement test (AIMS): reading, mathematics, and science.

Excluded from the analysis were certified staff operating within specialized instructional program areas such as self-contained SPED, pre-school, alternative and supplemental assistance, and a variety of instructional support activities including content-specific intervention specialists, gifted specialists, technology support Specialists, instructional growth coaches, Title 1 data specialists, counselors, psychologists, and individuals serving in temporary or supporting administrative positions.

Quantitative Information: Data Reduction and Selection

As mentioned, the original data set contained evaluation information for a total of 1,335 certified staff employed within the district during SY2012-13. Of these, 238 individuals meet the homogeneity criteria regarding Group A membership and the availability of both formalized professional practice (Danielson) and aligned achievement growth (VAM) scores. Teachers (classrooms) with achievement measures of fewer than ten students were also excluded from the database. Table 5 summarizes the grade level distribution of classroom teachers retained for further analysis.

Table 5

Grade Level Distribution of Teachers

Position Description	Frequency	Percent
3RD GRADE ES	78	32.8
4TH GRADE ES	68	28.6
5TH GRADE ES	67	28.2
6TH GRADE ES	25	10.4
Total	238	100.0

Teachers assigned to classrooms in grades 3, 4, and 5 accounts for 90% of the retained observations. Departmentalized instruction is conducted at some school locations in grade 6. This accounts for the reduced representation at this grade level.

Teacher status. The organization's evaluation requirements differ depending on the number of years a teacher has been employed in the district. Teachers employed for more than three years are identified as *continuing* and are required to have at least one formal evaluation per year. Teachers employed three years or less are termed *probationary* and are required to have at least two formal evaluations per year. This distinction (years of experience) is explored as part of a supporting research question and is referred to in the dataset as *Teacher Status*. Table 6 reports the distribution of continuing and probationary teachers in the data set.

Table 6

Distribution of Teachers by Employment Status

Teacher Status	Frequency	Percent
Continuing	141	59.2
Probationary	97	40.8
Total	238	100.0

As shown, approximately 60% (n = 141) of the records represent more experienced teachers. In addition, the dataset includes a field designating the total number of years each teacher has been with the district (*Years in District*). Summary distribution statistics on this dimension are reported in Table 7.

Table 7

Teacher: Years in District- Descriptive Statistics

Descriptive Statistic	Value
Number of Teachers	238
Mean Yrs. In District	5.28
Median Yrs. in District	5.00
Std. Deviation (Yrs. In District)	3.845
Std. Error (Mean Yrs. In District)	.249
Range	21
Minimum	1
Maximum	22
25th Percentile	2.00
50th Percentile	5.00
75th Percentile	8.00

The mean/median number of years employed in the district is approximately five with values ranging between one to 22 years. One quarter of the teachers report two or less years of experience while approximately 25% report eight or more years' experience. Table 8 and Figure 25 display the frequency distribution for the *Years in District* variable.

Table 8

Teacher: Years in District – Frequency Distribution

Years in District	Number of Teachers	Percent of Teachers	Cumulative Percent
1	53	22.3	22.3
2	24	10.1	32.4
3	19	8.0	40.3
4	13	5.5	45.8
5	19	8.0	53.8
6	21	8.8	62.6
7	23	9.7	72.3
8	24	10.1	82.4
9	13	5.5	87.8
10	12	5.0	92.9
11	5	2.1	95.0
12	3	1.3	96.2
13	3	1.3	97.5
14	2	.8	98.3
18	2	.8	99.2
20	1	.4	99.6
22	1	.4	100.0
Total	238	100.0	

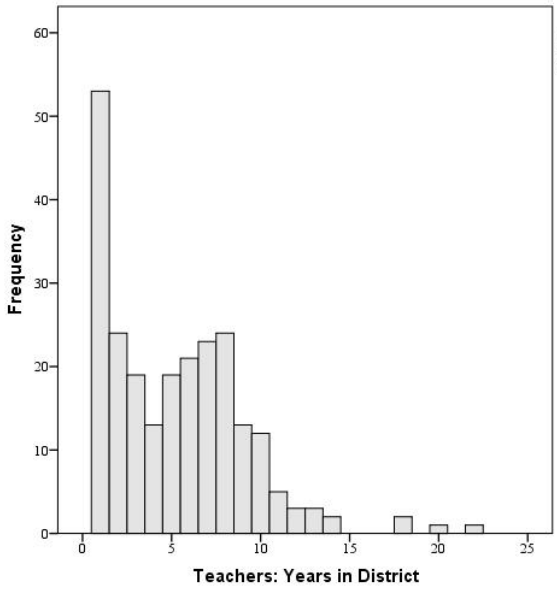


Figure 25. Histogram of teachers' experience (years in district).

As shown, 54% report less than 6 years' experience in the district with 93% reporting ten years or less experience.

School locations. During SY2012-13, the district operated 20 elementary schools within a 140 square mile service area. The community characteristics (income, population, age of housing stock, etc.) vary considerably including nine elementary facilities designated as Title I schools based on the proportion of students qualifying for the National Free/Reduced Priced Lunch Program. Table 9 reports the distribution of Group A teachers in the dataset across the twenty elementary schools.

Table 9

Distribution of Group A Teachers Across Elementary Locations

School Location Place Holder	Title I Facility	Number of Teachers	Percent of Teachers
1	Y	14	5.9
2		9	3.8
3		8	3.4
4		12	5.0
5	Y	6	2.5
6	Y	11	4.6
7	Y	15	6.3
8	Y	10	4.2
9	Y	12	5.0
10		11	4.6
11		14	5.9
12		8	3.4
13		20	8.4
14	Y	16	6.7
15		14	5.9
16		15	6.3
17	Y	13	5.5
18	Y	11	4.6
19		9	3.8
20		10	4.2
Total	9	238	100
	Mean	11.9	5
	Minimum	6	2.5
	Maximum	20	8.4
	Range	14	5.9
	St. Dev.	3.31	1.39

The number of teachers (evaluation records) across elementary locations ranges between six (3%) and 20 (8%), with a mean of approximately 12 teachers per school. A total of 108 (45%) teach at low-income, Title I, facilities.

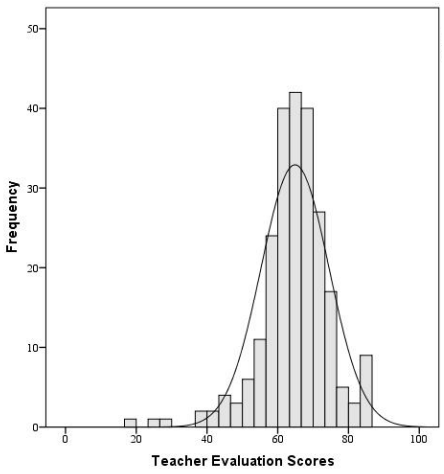
Evaluation component scores. As previously discussed, each teacher's overall evaluation score is constructed as a weighted average of his/her Professional Practice (PP) and VAM score. PP scores are expressed as a percentage of total possible PP points while VAM measures are represented as averaged median percentiles of all students in the classroom or aggregation group. For each teacher, the PP score is weighted by a factor of .67 and the VAM scores by .33. The resulting combined-score scale ranges from approximately zero to 100. Table 10 reports descriptive distribution statistics for the Overall, Professional Practice (PP) and Academic Growth (VAM) scores.

Table 10

Descriptive Statistics – Total Evaluation Scores

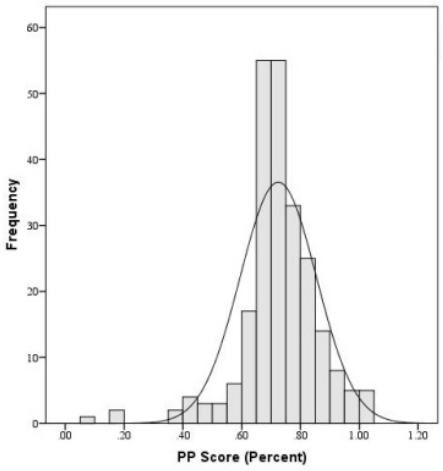
	Overall Evaluation Score (Total Possible = 100)	Professional Practice Score (Percent of Total Possible Points)	Academic Growth (VAM) Score (Averaged Median Percentile)
Valid *N*	238	238	238
Missing *N*	0	0	0
Mean	64.884	72.436	49.658
Std. Error of Mean	.623	.8416	.484
Std. Deviation	9.612	12.983	7.459
Median	65.480	72.514	49.828
Skewness	-1.003	-1.139	-.321
Std. Error of Skewness	.158	.158	.158
Kurtosis	3.550	4.525	.365
Std. Error of Kurtosis	.314	.314	.314
Minimum	18.602	8.890	22.149
Maximum	85.743	100.020	69.105
Range	67.141	91.130	46.955
25th Percentile	60.473	66.679	45.143
50th Percentile	65.480	72.514	49.828
75th Percentile	70.498	78.972	55.046

The data indicate that the mean (M = 64.9) and median (Mdn = 65.5) Overall Evaluation Scores are similar with values ranging between 19 and 86. The mean PP and VAM scores were 72 and 50, respectively. Review of skew and kurtosis statistics suggests that the Overall and PP score distributions are substantively non-normal reporting skew values below -1.0 and kurtosis values above 3.0 (Brown, 2011; Bulmer, 1979; Ho, 2014,).[6] In comparison, the distribution of the VAM scores are more symmetric and more closely approximates a normal distribution. This is important because violations of normality become problematic depending on the type of test statistic being used to assess various research hypotheses. Figures 26 to 28 report histograms associated with each measure.

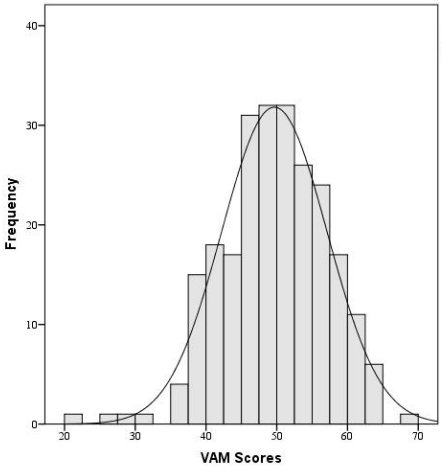


Figures 26. Distribution - evaluation scores.

[6] Perfectly normal distributions are characterized by Skew = 0 and Kurtosis = 3.



Figures 27. Distribution - Professional Practice scores.



Figures 28. Distribution - VAM scores.

Professional Practice element rating scores. Professional Practice ratings are assigned to each of the 22 behavioral elements contained in the Danielson FFT evaluation framework. Each element is rated using a four-option integer scale with values ranging

from *zero* (lowest performance) to *3* (highest performance). Table 11 reports the distribution of assigned ratings for each of the 22 elements.

Table 11

Distribution of Assigned Professional Practice Ratings

	Number of Ratings				Percent of Total Ratings				
FFT Behavioral Component	(0)	(1)	(2)	(3)	(0)	(1)	(2)	(3)	Total
D1a Demonstrating Knowledge of Content and Pedagogy	2	22	164	50	1%	9%	69%	21%	100%
D1b Demonstrating Knowledge of Students	2	10	153	73	1%	4%	64%	31%	100%
D1c Setting Instructional Outcomes	0	18	171	49	0%	8%	72%	21%	100%
D1d Demonstrating Knowledge of Resources	1	14	179	44	0%	6%	75%	18%	100%
D1e Designing Coherent Instruction	0	20	167	51	0%	8%	70%	21%	100%
D1f Designing Student Assessments	1	19	184	34	0%	8%	77%	14%	100%
D2a Respect and Rapport	0	4	136	98	0%	2%	57%	41%	100%
D2b Culture of Learning	3	13	135	87	1%	5%	57%	37%	100%
D2c Managing Classroom Procedures	2	6	156	74	1%	3%	66%	31%	100%
D2d Managing Student Behavior	1	11	174	52	0%	5%	73%	22%	100%
D2e Organizing Physical Space	1	3	188	46	0%	1%	79%	19%	100%
D3a Communicating With Students	0	10	134	94	0%	4%	56%	39%	100%
D3b Questioning Discussion Techniques	1	41	160	36	0%	17%	67%	15%	100%
D3c Engaging Students	0	22	157	59	0%	9%	66%	25%	100%
D3d Using Assessment in Instruction	2	19	182	35	1%	8%	76%	15%	100%
D3e Flexibility Responsiveness	2	13	181	42	1%	5%	76%	18%	100%
D4a Reflecting on Teaching	0	16	160	62	0%	7%	67%	26%	100%
D4b Maintaining Accurate Records	1	12	172	53	0%	5%	72%	22%	100%
D4c Communicating with Families	2	11	158	67	1%	5%	66%	28%	100%
D4d Participating in a Professional Community	1	11	159	67	0%	5%	67%	28%	100%
D4e Growing Developing Professionally	2	18	175	43	1%	8%	74%	18%	100%
D4f Showing Professionalism	2	7	153	76	1%	3%	64%	32%	100%
Totals	26	320	3598	1292	0%	6%	69%	25%	100%

Scale: (0) Unsatisfactory; (1) Basic; (2) Proficient; (3) Distinguished

As shown, the distribution of rating options is not evenly dispersed. Indeed, nearly 70% of all FFT ratings assigned to the 238 Group A teachers were *Proficient (2)*,

followed by 25 percent *Distinguished (3)*. Only 6% of assigned ratings were at the *Basic (1)* level, and less than .5% were *Unsatisfactory (0)*. Ninety-four percent of all assigned ratings were in the combined *Proficient (2)* and *Distinguished (3)* categories.

Construct Validity Index Questionnaire. Members of the Teacher Evaluation Committee and IGT groups were asked to complete the Construct Validity Ratio (CVR) questionnaire (Lawshe, 1975; Wilson et al., 2012). These individuals were asked to complete the CVR due to their in-depth understanding of the component FFT elements and their use in the evaluation process. In essence, these groups serve as "subject matter experts" regarding the appropriateness of the FFT framework to provide suitable representation of instructional quality (Lawshe, 1975).

As previously discussed, the CVR asks individuals to rate the relative importance of each FFT element for its ability to identify the construct of "good/effective" teaching. Response options for each element include *Not Important*, *Important,* and *Very Important.* On the CVR, the proportion of *Very Important* responses serves as a measure of construct identification and coverage. That is, the more items rated as being *Very Important,* the better the instrument aligns to the underlying posited construct of good/effective teaching (Lawshe, 1975).

The CVR also asked respondents to provide personal reflections on what it means to be a good/effective teacher. These qualitative responses served as additional input to a number of research questions identified in this study. A copy of the CVR questionnaire is provided in Appendix C. Additional sampling details and data collection for the CVR are discussed under Research Question *RQ1B (c): To what degree do the 22 elements*

contained within the theoretical FFT framework adequately represent the latent TIQ construct?

Part 2: Descriptive Summary: Qualitative Data Collections

Introduction

As discussed in Chapters 1 and 3, this study identified four primary stakeholder groups influencing the policy, design, and implementation of the district's teacher evaluation system: classroom teachers, principals, district policy leaders, and state policy leaders. Conceptually, teachers are seen as recipients, principals as evaluators/implementers, district policy staff as internal decision makers, and state policy persons as external decision makers. Two additional stakeholder groups were also identified in the study design: members of the District's Teacher Evaluation (TEval) Committee and instructional growth teachers/coaches (IGTs). Members of the TEval committee are seen as architects of the evaluation system, possessing detailed knowledge/understanding of the system's implementation plan and purpose. IGTs are viewed as supporters of the system and had working knowledge of the Danielson evaluation rubrics and how they are applied in classrooms on a day-to-day basis.

Following the methods outlined in Chapter 3, interviews and questionnaires were used to gather perceptual data from each stakeholder group. Table 12 reports group summary information including population, identified sample size, final data collection counts, and sample collection rates.

Table 12

Response Rates for Qualitative Data Collections

Stakeholder Group	Total Population	Identified Sample for Study	Data Collection	Sample Response Rate	Method
Classroom Teachers	235	9	7	78%	Interview
Principals/Evaluators	20 (ES)	9	8	89%	Interview
District Policy	6	5	4	80%	Interview
State Policy	5	5	3	60%	Interview
Subtotal:	266	28	22	79%	
Teacher Evaluation Committee	9	9	9	100%	Survey
Instructional Growth Coaches	24 (ES, HS)	24	14	58%	Survey
Subtotal:	33	33	23	70%	
Grand Total:	299	61	45	74%	

Population Notes: The District Policy group includes Superintendent (n = 1), Assistant Superintendents (n = 4; Human Resources, Business/Finance, Curriculum and Instruction; and Education Services & Operations), and the Director of Human Resources (n = 1); (ES): Elementary School; (HS) High School

A total of 22 in-person stakeholder interviews were completed. This represented 79% of the originally identified sample. In addition, information was received from all members of the Teacher Evaluation Committee and approximately 70% of the instructional growth coaches cadre. These latter two groups were asked to complete the Construct Validity Ratio (CVR) questionnaire (Lawshe, 1975). The questionnaire asked respondents to reflect on the meaning of being a good/effective teacher (paralleling a prompt used during the in-person interviews) and the relative importance of each behavioral component used in the Danielson rating framework (via the CVR). A more

complete discussion of the CVR, the respondent groups, and the information collected is provided in a later section of Chapter 4.

Coding and analysis of the interview data proceeded as outlined in Chapter 3 using techniques adapted from grounded theory research (Corbin & Strauss, 2008; Saldana, 2009). This involved a purposeful progression of data review, initial reflection, and memoing (annotation), followed by first applications of in vivo/open coding. Subsequent passes delineated broader categories which connected/aligned codes into larger conceptual domains. The derived categories were then united into themes/constructs providing the basis of inferential interpretations and reflection within each of the posited research questions.

Each of the coding steps was completed on information collected from each stakeholder group by interview prompt. All of the stakeholder interviews utilized a pre-defined semi-structured protocol aligned to each qualitative research questions posited for this study. In essence, the pre-defined constructs served to organize major portions of the interview process. In some cases, interview prompts contributed understandings to multiple qualitative research questions. Finally, the specific phrasing of selected interview prompts was adapted for each stakeholder group to better align with their roles/perspectives of the evaluation process. The alignment between each qualitative research question and its contributing interview prompt is provided in Appendix G.

The initial coding process lead to identification/refinement of nine global constructs: *Purpose of Education, Role of Teachers in Education, Purpose of Teacher Evaluation, Definition of Good/Effective Teacher, Ability of the Evaluation System to Identify a Good/Effective Teacher, Test Scores as a Measure of Instructional*

Effectiveness, Impact of Evaluation System on Professional Practice, Teacher/Educator Voice in Evaluation System Design and Implementation, and *Ways to Improve the Evaluation System*. This last construct was not initially aligned to a primary or supporting research question. Rather, it grew out of the ensuing stakeholder discussions. The global constructs served as the basis for aligning multi-faceted evidences to specific research questions. A complete rendering of the global constructs including all codes, identities, annotations, and supporting exemplars extracted from the narrative data is available as a separate document to this study. An online version of the global construct document may be accessed at http://disslinks.blogspot.com/

The global construct development process occurred in the following order of activities:

Phase 1: Initial review, coding, code refinement

1. Load raw document transcriptions into the HyperResearch application program.
2. Begin initial read, review of transcript information; Begin initial in vivo/open coding identification and code annotation; Begin initial memoing and construct reflections.
3. Construct/organize an initial coding structure/hierarchy based on pre-identified interview protocols and emerging codes/constructs; Continue application of in vivo/open coding identification and annotations.
4. Construct more formalized conceptual code-categories (axial coding); code refinement, re-coding, code re-organization.

Phase II: Construction of Global Code Groups

5. Identification of Global Code Groups within HyperResearch; re-organization of code identifiers under Global Code Groups; Continued code identification, code refinement, annotation, and reflective memoing.

Phase III: Restructuring of Qualitative Database

6. Restructuring database using Global Code Groups as primary organizers; Transform database structure from *Case x Global Code-Group* to *Global-Code-Group x Stakeholder-Group* format. This was accomplished in HyperResearch by nesting the application's Case/Code filters within the Report Builder tool. Table 13 reports the reorganized structure of the HyperResearch database.

Table 13

Restructured Interview Database – Global Code Group Designations by Stakeholder Group

Global Code Group Descriptive Category	Teachers	Principals	District Policy	State Policy	TEval Committee & IGTs
Purpose of Education	✓	✓	✓	✓	
Role of Teachers in Education	✓	✓	✓	✓	
Purpose of TEval Systems	✓	✓	✓	✓	
Define Good/Effective Teacher	✓	✓	✓	✓	✓
TEval Identify Good/Effective teacher?	✓	✓	✓	✓	
Test Score as Measure of Good/Effective teacher	✓	✓	✓	✓	
TEval Impact on Professional Practice	✓	✓	✓	✓	
Teacher/Educator Voice in TEval System	✓	✓	✓	✓	
How to Improve TEval System	✓			✓	

7. Export *Global-Code-Group x Stakeholder* data structures into Microsoft Word; This resulted in one file document for each *Global-Code-Group x Stakeholder* (Total of 35 files).

Phase IV: Conduct New Round of Analysis

8. Conduct a new round of analysis: coding, code refinement, annotation, memoing within each Global-Code-Group x Stakeholder file; Application of MS Word's Track Changes, text coloring, and comment insertion to position final list of codes, categories, annotations, and reflections onto each case entry within each file.

Phase V: Extract & Re-group codes, categories, comments, reflections

9. From each Global Code Group, extract codes, categories, annotations, and reflections for each stakeholder group into new document – this combined all stakeholder information for the specific Global Code Group into one file.

Phase VI: Finalize Categories, Themes, and Assertions

10. Summarize collection of reflections, categories, codes, annotation into themes.
11. Condense Global Code Group themes within and across stakeholders; Formulate extensions, interpretations, assertions.

Phase VII: Align evidence-based themes, assertions to study research questions

The final *Global Construct Codebook* is available as a supporting document to this study. It may be accessed at http://disslinks.blogspot.com/ The *Global Construct Codebook* contains all codes, identities, annotations, and aligned stakeholder exemplars assembled for each of the nine global constructs. These assemblies served as the primary input to examine each of the study's primary and secondary research questions.

Data verification and reliability checks. All 22 in-person interviews were audio recorded and transcribed using a professional transcription service. During the analysis activity, selected portions of the transcribe materials were cross-referenced with the recordings. This primarily occurred when the transcription accuracy seemed questionable (i.e., unclear, out of context, and/or where the transcriber had indicated a problem). While conducting the initial reading of all transcription documents, edits were limited to issues of pagination, paragraphing, punctuation, grammatical adjustments, and corrections due to incorrect word-language interpretations.

Member Checking

Copies of all interview transcriptions were forwarded back to each participant. Participants were asked to review the transcription and provide corrections, clarifications, and/or additional reflections (this was completed via email communication). While some participants acknowledged receipt and appreciation for the opportunity for review the information, few provided substantive additions to their initial comments. None of the interviewees substantively changed or withdrew aspects of their original reflections.

The qualitative data were also subjected to three methods of researcher-triangulation (see Chapter 3). First, selected original (uncoded) interview transcriptions were forwarded to two professional colleagues serving as co-researchers to the organization's internal Teacher Evaluation System Program Evaluation Project.[7] Each co-researcher reviewed select portions of eight interview transcripts, two from each stakeholder group (teacher, principal, district policy, and state policy). In addition, these

[7] This research study was being conducted along with a formal program evaluation of the district's teacher evaluation system. It is noted this researcher also serves as direct supervisor to the co-researchers.

individuals reviewed all qualitative responses obtained from the Lawshe Content Validity Ratio (CVR) survey, administered via online questionnaire to instructional growth coaches and Teacher Evaluation Committee members. The two co-researchers were asked to independently review and informally code/reflect upon the same set of information. Table 14 reports the stakeholder narratives selected for co-research review.

Table 14

Narrative Transcript Re-Coding (Co-Researchers)

Stakeholder Group	Participant Codes		Construct Topic
State	402	403	Test Score as Good TIQ Measure
District	301	303	Purpose of TEval
Principal	202	205	Define Good/Effective Teacher
Teacher	101	108	TEval Identify Good Teacher
Lawshe CVR	(All Survey Response)		Define Good/Effective Teacher

Second, a sample of four previously coded/annotated transcripts were provided to the co-researchers. The co-researchers were asked to read the coded passages and reflect on the codes, categories, themes and related annotations that this researcher had identified. In addition, each co-researcher was asked to review and comment on summative themes and assertions aligned to a selected subset of Global Code Groups.

Co-researcher review of the uncoded and coded information was undertaken sequentially. The random number generator function within Microsoft Excel was utilized

to rank order participants within each stakeholder interview groups: lowest rank order identified transcripts for uncoded review and then for coded review.

After the co-researchers completed their reviews, the team met to discuss findings, compare codes, categories, and common themes. The intent was to obtain feedback on the reasonableness, clarity, and interpretation of the coded sections and discuss some of the derived perspectives. The dialog served to both validate some aspects of the process and raise critical reflection on others. This researcher used this information to engage in additional rounds of reading, coding, and reflection.

Third, findings from both the quantitative and qualitative analysis were presented to members of the district policy team, the district's Teacher Evaluation Committee, and other colleagues with knowledge of the evaluation system. Feedback, confirmation, and reflection from this activity were intended to help inform on the reasonableness of the interpretative findings.

Finally, reflections from all the data triangulation and review activities were included as part of this researcher's journaling process. Here, informal reflections and collegial discussions resulting from the data review process provide additional data regarding credibility, relevance, and validity of findings.

Interview Data Summary

Appendix G summarizes each of the qualitative data collection activities conducted across stakeholder groups including interview duration, number of transcribe words, and transcriptions selected for review by co-researchers.

Research Journal

This researcher maintained a journal as part of the study's data collection activities. These journal reflections served two purposes. First, journaling provided a means to track/record reflections concerning the research process, make notes about needed adjustments, upcoming activities, reactions to discussions, meetings, and/or observations, and to generally provide a method and location to keep track of the dynamic nature of the activity.

Second, the research journal served as a primary source of data from which to evaluate questions posited under Research Question #4: *How did the process of engaging in this action-research study impact the investigator as a scholarly researcher and organizational leader?* This primary question was divided into three supporting questions:

1. *What barriers or impediments were encountered during the course of the study? How were they overcome and/or handled?*
2. *What aspects of the action-research process provided the most growth and/or learning for the researcher-practitioner?*
3. *What were the salient learnings from this study? How did the researcher grow professionally and personally? How will these learnings be incorporated into future research and leadership activities?*

To answer these questions, journaling served as a primary data collection/recording technique. Three types of journaling methods were developed during the course of the study. The first method was to maintain a written (hard copy) journal. Here, a journal notebook accompanied the researcher to selected meetings, observations,

and interviews conducted as part of the study over the course of approximately 16 months (September 2012 to December 2013). This included professional evaluation training activities, administrative and evaluation committee meetings, policy discussions, and many other formal and informal sessions held as part of the district's implementation process. In many instances, notes were made at the time of the event(s), while in other occasions entries were made after attending/experiencing the event. Indeed, for many of these sessions, this researcher served as an active facilitator/participant and the journal provided an opportunity to reflect after the fact. At the close of the study, the hard-copy journal spanned 166 handwritten pages.

The second journaling method involved construction of a personal online private blog site. This was instituted approximately half way through the research study as a means to record reflections using web-enabled devices (laptop, smartphone, tablet, or desktop computer) at times/locations where access to, or use of, a written journal was not convenient. Indeed, after construction of the online blog site, this researcher gradually began to use the tool more often than written journal entries. A main benefit of this approach was the ability to export journal entries in standard text formats for use in qualitative software analysis applications such as HyperResearch. While written journal entries require traditional methods of qualitative coding, insertion into electronic mediums allow for more complex and efficient coding approaches. At time of writing, online blog posts totaled 30 entries, 6,949 words spanning a period between August 2013 and September 2014.

The third journaling activity occurred during the process of writing the final dissertation document, primarily during construction of Chapter 4 (data analysis). As will

be shared in discussion of Research Question 4 (personal reflections from the research process), the process of completing the final write up proved to be a thought provoking analytic process. While writing, reflecting, and re-writing portions of Chapter 4, this researcher began thinking about topics, issues, comments relevant for use in Chapter 6 (reflections). However, this process of writing, reflecting, and then annotating quickly took on a new form of journaling, that is, it became an independent analytic task. At time of this writing, these entries total 3,324 words.

Part 3: Research Questions – Analysis and Findings

Introduction

As mentioned, this study posits four primary research questions (RQ1 to RQ4), each composed of subordinating questions and areas of investigation. In total, 20 separate analytic questions are investigated each concerned with assessing differing aspects of the teacher evaluation construct. Appendix B provides a concise listing of the research questions and related analytic frameworks.

In this section, each primary research question is addressed sequentially beginning with Research Question #1 through Research Question #4. Within each, the supporting research questions are identified and also addressed in sequential order. When necessary, additional aspects of the data are presented along with descriptions of the particular analytic approach. In some instances, the reader is referred to additional appendices for more detailed tabulations related to specific statistical techniques. At the close of this chapter, a summary review of collective findings is provided, serving as the basis for more interpretive discussions and reflections in Chapters 5 and 6.

Primary Research Question #1

To what degree does the validity evidence generated by the LEA's policy-directed teacher evaluation system support inferences of Teacher Instructional Quality (TIQ)?

Primary Research Question #1 examines characteristics of the measures used to construct indicators of teacher instructional quality (TIQ). It does so by utilizing both quantitative and qualitative analytic methods. Quantitative methods assess characteristics of the metrics used to reify the posited construct while qualitative approaches are used to evaluate the theoretical assertions authorizing the evaluation process.

The reader is reminded that the state policy-directed teacher evaluation framework posits that academic growth measures estimated by statistically-based value-added models (VAM) may be combined with independent measures of a teacher's professional practice (PP) to inform on Teacher Instructional Quality (TIQ), a latent (unobserved) construct. Here, PP is positioned as an instructional input facilitating student learning. In contrast, VAM is interpreted as an instructional outcome causally connected to PP.

Under this paradigm, greater degrees of instructional competence (PP) are assumed to facilitate higher levels of student learning (VAM). Conversely, higher levels of achievement (VAM) are assumed to originate from higher levels of instructional competence (PP). If the construct representation holds, it follows that significant positive correlations should exist between the two measures. In this way, each serve as a criterion measure of the other since both are hypothesized to represent selected aspects of the same latent construct. In addition, since the VAM measures are constructed from multiple sub-

components, substantive positive associations are posited to exist within/between these elements.

Like the VAM metrics, the PP measures are composed of multiple sub-scales and components. These sub-components are posited to conform to an a priori set of content representative of high quality instructional behavior. In addition, all component scales (VAM and PP) are assumed reify the latent construct (TIQ) with adequate measure reliability and that these measured components match the theoretical attributes ascribed to good/effective teaching. Finally, the evaluation policy context presumes to both assess and improve the instructional practices of classroom teachers.

To explore these assumptions, Research Question #1 is delineated by five complementary categories of construct evidence: (RQ1A) Criterion, (RQ1B) Content, (RQ1C) Consequential, (RQ1D) Scale Reliability, and (RQ1E) Theoretical Construct Definition (Messick, 1989a, Messick 1989b; AERA et al., 1999, 2014). Each category contains its own primary and secondary research questions which, in turn, provide focus for the data collection and analytic methods employed. Evidences for each will be explored in sequential order of presentation.

Research question 1A: Criterion evidence (RQ1A). The criterion evidence portion of Research Question #1 is examined using correlation and means testing techniques applied to the two main metrics used in the evaluation framework: measures of academic growth (value-added—VAM) and measures of professional practice (PP). The data was assembled for 238 Group A teachers instructing in general education, self-contained classrooms, with ten or more students for which complete data was available during SY2012-13. The academic growth data was constructed using multi-level models

estimating student VAM scores. The PP metrics are based on evaluator ratings of classroom practices across the 22 Danielson FFT behavioral elements. The Criterion Evidence portion of Research Question #1 is articulated by five supporting investigations. Table 15 summarizes the supporting research questions specified under RQ1A (Criterion Evidences).

Table 15

RQ#1A: Criterion Evidence: Five Supporting Research Questions

Item	Description	Approach	Measure
RQ1A (a):	To what degree do value-added measures of instructional effectiveness correlate with measures of professional practice (PP)?	Correlation	r, r^2
RQ1A (b): *	To what degree do measures of PP assigned by qualified evaluators correlate with teacher's self-assessment of PP?	Correlation	r, r^2
RQ1A (c):	To what degree do VAM estimates of instructional effectiveness in reading and mathematics correlate?	Correlation	r, r^2
RQ1A (d):	Do PP sub-scale scores display similar degrees of correlation with VAM measures?	Correlation	r, r^2, χ^2
RQ1A (e):	To what degree are high, middle, and low VAM estimates of instructional effectiveness able to differentiate PP	ANOVA, Means Testing	t, F tests, Post Hoc Tests of Group Mean Differences

* The district decided not to implement the information system supporting this data collection activity

Each supporting research question under RQ1A (Criterion Evidence) will be discussed in sequential order.

RQ1A (a). To what degree do value-added measures of instructional effectiveness correlate with measures of professional practice (PP)? The approach is correlation. The measures are r and r^2.

This research question investigates the correlation between the two primary components of the evaluation measure: PP and VAM. It is reasoned that stronger correlations indicate support of the posited construct, that each component contributes information to assessing the instructional competency of classroom teachers. Table 16 provides a descriptive summary of the two measures.

Table 16

Descriptive Statistics – Professional Practice and VAM Scores

Component	Metric	Mean	Std. Deviation	*N*
PP Score	Percent	.724361	.1298280	238
VAM Score	Median Percentile	49.657617	7.4592983	238

As shown, the mean PP percent score for all 238 teachers was approximately 72% (out of a total possible 66 rating points). This equates to approximately 47.52 points on the Danielson FFT rating scale. The mean of the median percentile VAM score for all teachers was just under 50.

Table 17 reports the Pearson product-moment correlation coefficient between the two measures.

Table 17

Correlation Between PP and VAM Scores

		PP Score	VAM Score
PP Score	Pearson Correlation	1	.254**
	Sig. (2-tailed)		.000
	N	238	238
VAM Score	Pearson Correlation	.254**	1
	Sig. (2-tailed)	.000	
	N	238	238

** Correlation is significant at the 0.01 level (2-tailed)

Using a significance criteria of $p < .05$, the data show a significant, positive, correlation between the two measures ($r = .254$, $n = 238$, $p < .001$). Squaring the correlation coefficient indicates the proportion of common variance existing between the measures. Here, $r^2 = .06$, indicating 6% of common variation. In addition, Pearson's correlation coefficient may be interpreted as an effect size (strength of the association). Referencing Cohen and Cohen (1983), an effect size of .254 is interpreted to represent a weak-to-moderate association.

RQ1A (b). To what degree do measures of PP assigned by qualified evaluators correlate with teacher's self-assessment of PP? The approach is correlation. The measures are r and r^2.

At the start of the study, the district planned on implementing an online data system to capture teacher's self-perception of their instructional competencies. Unfortunately, this system was not implemented and no data were available to investigate

this research question. Instead, teachers were instructed to discuss their self-perceptions with their local school administrator(s) during pre-conferences held at the start of the school year.

RQ1A (c). To what degree do VAM estimates of instructional effectiveness in reading and mathematics correlate? The approach is correlation. The measures are *r* and r^2.

Measures of student academic growth (VAM) are a primary component of the policy construct for assessing Teacher Instructional Quality (TIQ). As operationalized by the district, the academic growth component for Group A elementary teachers is constructed as a combination of math, reading, and science (grade 4) scores measured by the state's standardized achievement test across two academic years: SY2011-12 & SY2012-13.[8]

To construct the measures, growth (VAM) scores were computed for each student in the classroom in each subject area and transformed onto percentile scales based on the distribution of same-subject, same-grade scores for all students in the district. Classroom level aggregates were computed as median subject percentiles. The final within-year VAM score for a teacher is computed as the average median percentile across the three subject areas. At minimum, all teachers in the data had VAM scores for the 2012-13 school year. If two years of VAM scores were available, the teacher's final growth measure was represented as a simple average of the two years of aggregated information (i.e., VAM 2012 & VAM 2013).

---

[8] AIMS Science is only administered in Grade 4. Thus, VAM measures for grade 4 teachers are computed as the average median residual scores for math, reading, and science. VAM for non-grade 4 teachers are limited to math and reading.

The policy construct, together with the operationalized approach for assembling classroom growth measures, assumes that each component subject score contributes to the overall assessment of TIQ. That is, effective teaching is not subject-dependent and higher competency levels would be apparent (represented) in all achievement measures. If so, the associations between the component growth measures should be high, indicating that the measures equally capture the construct of instructional competency. If the associations between subjects and across years are not high, then the lack of common variance in the measures detracts from the reliability of the indicator and negatively impacts the inferences attributed to the metric.

Table 18 reports the descriptive statistics for the component VAM measures across the two years used to compute teacher's overall academic growth scores.

Table 18

Descriptive Statistics for Component VAM Measures

		SY2012-13			SY2011-12		
		Math	Read	Science	Math	Read	Science
N	Valid	233	233	67	181	184	48
	Missing	5	5	171	57	54	190
Mean		49.920	50.169	49.507	49.998	50.017	49.976
Std. Error of Mean		.887	.765	1.449	.885	.743	1.587
Median		50.897	50.396	49.786	49.297	49.765	50.063
Std. Deviation		13.543	11.670	11.860	11.908	10.076	10.992
Skewness		-.195	-.145	-.119	.041	.231	-.470
Std. Error of Skewness		.159	.159	.293	.181	.179	.343
Kurtosis		.035	.315	-.236	.297	.002	1.174
Std. Error of Kurtosis		.318	.318	.578	.359	.356	.674
Range		79.362	75.284	51.498	68.968	56.877	57.350
	Minimum	10.833	13.791	21.730	11.765	23.921	16.151
	Maximum	90.194	89.075	73.227	80.733	80.798	73.502
Percentiles	25	41.597	42.090	41.687	43.512	43.652	44.054
	50	50.897	50.396	49.786	49.297	49.765	50.063
	75	59.781	58.722	57.946	57.442	55.950	56.286

Statistics: All data – average median residual percentile

The Pearson Product-Moment Correlation is a measure of linear association between two variables (Ferguson & Takane, 1989; Klugh, 1986). It assumes that each measure is continuous and approximately normally distributed. Substantive deviations from normality distort the interpretation of the measure. The data presented in Table 18 provide indications of each variable's distributional characteristics. Specifically, the mean and median values for each variable are similar, suggesting that the impact of any existing outliers is minimal. The skewness indicators are all substantive less than +/- 2.0 and the kurtosis indicators are substantively less than 7.0. Chou and Bentler (1995) and

Curran, West, and Finch (1996) use these criteria for assessing problematic deviations from normality that can lead to estimation issues in procedures such as Maximum Likelihood and for statistical tests that require the assumption of normality to hold. Finally, the quartile quantities reported by the percentile intervals are generally symmetric about the median value of 50.

Table 19 reports results of a Kolmogorov–Smirnov test, a statistical procedure for assessing the degree to which sample data deviates from a normal distribution (Green & Salkind, 2011). For this test, the null hypothesis assumes the data are normally distributed. Interpretation of the test statistic is based on a rejection criterion of $p < .05$.

Table 19

Kolmogorov–Smirnov Test for Normality

Subject	Year	Test Statistic	Significance (2-tail)	Inference on H_O
Math	2013	.511	.957	Do Not Reject
Reading	2013	.455	.989	Do Not Reject
Science	2013	.621	.836	Do Not Reject
Math	2012	.768	.597	Do Not Reject
Reading	2012	.650	.792	Do Not Reject
Science	2012	.656	.783	Do Not Reject

For the sample data, the test statistics fails to reject to null hypothesis, suggesting an approximately normal distribution.

Determining that the VAM variables are continuous normal distributed allows for interpretation of the relative associations between the variables. In this regard, Table 20

reports the Pearson Product-Moment correlations between each of the component VAM measures.

Table 20

VAM Correlations at the Classroom Level for Group A Teachers

Subject/Year		Math 2013	Read 2013	Science 2013	Math 2012	Read 2012	Science 2012
	r	1	.259**	.308*	.163*	.074	-.031
Math_2013	Sig.		.000	.011	.031	.328	.838
	N	233	233	67	176	179	47
	r	.259**	1	.278*	.044	.038	-.220
Read_2013	Sig.	.000		.022	.564	.614	.138
	N	233	233	67	176	179	47
	r	.308*	.278*	1	.104	-.085	.097
Science_2013	Sig.	.011	.022		.494	.566	.557
	N	67	67	67	46	48	39
	r	.163*	.044	.104	1	.341**	.457**
Math_2012	Sig.	.031	.564	.494		.000	.001
	N	176	176	46	181	180	48
	r	.074	.038	-.085	.341**	1	.271
Read_2012	Sig.	.328	.614	.566	.000		.063
	N	179	179	48	180	184	48
	r	-.031	-.220	.097	.457**	.271	1
Science_2012	Sig.	.838	.138	.557	.001	.063	
	N	47	47	39	48	48	48

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

As shown, the component correlations between the 2013 subject areas are all positively significant at the $p < .05$ level, ranging in values between .259 to .278. The correlations between the 2012 variables were positive significant ($p < .05$) for all except the association between Science and Reading ($p = .063$). The sample size for Science in 2013 was 67 classrooms while the sample size in 2012 was 48. This reduction in sample size may have been sufficient to inflate the standard error term in the test statistic enough to render the association statistically insignificant at $p < .05$ level of criteria. Regardless, the within year component correlations reported for 2012 seem to exceed those for 2013

and generally fall into the moderate range of effect size (Cohen & Cohen, 1988). Interestingly, only one of the cross-year subject associations was found to be positive significant ($p < .05$). This was between Math 2012 and Math 2013. However, the effect size of the association is weak at .163.

The presence of insignificant associations among VAM components compared across years questions whether the overall VAM measures for each year are correlated. This is explored by assessing the correlation of the combined 2012 and 2013 VAM measures. Table 21 reports this correlation.

Table 21

VAM Correlations Across Years

		VAM_2013	VAM_2012
VAM_2013	Pearson Correlation	1	.115
	Sig. (2-tailed)		.124
	N	233	180
VAM_2012	Pearson Correlation	.115	1
	Sig. (2-tailed)	.124	
	N	180	185

As shown, the correlation of VAM measures across 2012 and 2013 is not significant ($p < .05$).

RQ1A (d). Do PP sub-scale scores display similar degrees of correlation with VAM measures? The approach is correlation. The measures are r, r^2, and χ^2.

Recall that the PP measure for a classroom teacher is based on the sum of ratings assigned to each of 22 behavioral elements. These elements are grouped within four

behavioral domains: Domain 1 – Planning and Preparation (six elements), Domain 2: Classroom Environment (five elements), Domain 3: Instruction (five elements), and Domain 4: Professional Responsibilities (six elements). Table 22 reports the underlying descriptive summaries for the four PP behavioral domain scales.

Table 22

Descriptive Statistics – PP Domains

		FFT Total	Domain 1	Domain 2	Domain 3	Domain 4
N	Valid	238	238	238	238	238
	Missing	0	0	0	0	0
Mean		47.866	12.782	11.286	10.634	13.164
Std. Error of Mean		.5331	.1591	.1298	.1340	.1600
Median		48.000	12.000	11.000	11.000	13.000
Std. Deviation		8.2241	2.4552	2.0027	2.0676	2.4689
Skewness		-.817	-.509	-.605	-.694	-.786
Std. Error of Skewness		.158	.158	.158	.158	.158
Kurtosis		3.129	1.621	3.262	1.850	2.791
Std. Error of Kurtosis		.314	.314	.314	.314	.314
Range		55.0	14.0	14.0	13.0	16.0
	Minimum	11.0	4.0	1.0	2.0	2.0
	Maximum	66.0	18.0	15.0	15.0	18.0
Percentiles	25	44.000	12.000	10.000	10.000	12.000
	50	48.000	12.000	11.000	11.000	13.000
	75	52.000	14.000	13.000	12.000	15.000

Under the hypothesized construct it is argued that each of these behavioral domains contributes differing, but distinct, information on TIQ. It is reasoned that if each of these domains contribute to the assessment of instructional quality, and the academic growth scores also contribute information to this construct, then the associations between

the behavioral sub-domains and student academic growth measures should also be positive significant at a level sufficient to provide meaningful information.

It is noted that while the individual PP elements ratings are ordinal measures and cannot be considered distributed normally, the summated scores may be assessed for adherence to normality. Table 23 reports the results of the Kolmogorov–Smirnov test for Normality on each of the PP domain scales (Green & Salkind, 2011).

Table 23

Kolmogorov-Smirnov Test for Normality: FFT Behavioral Domains

Domain	*N*	Test Statistic	Significance (2-tail)	Inference on H_O
D1: Planning	238	.228	.000	Reject H_O
D2: Class Environment	238	.172	.000	Reject H_O
D3: Instruction	238	.174	.000	Reject H_O
D4: Professional Resp.	238	.205	.000	Reject H_O
PP Total Score	238	.122	.000	Reject H_O

As shown, the PP sub-scale measures are not distributed normally. For this reason, assessing the degree of association requires methods that account for the non-normal attributes of the information. The Spearman's rank correlation coefficient assess statistical associations based on the ranks of the data and does not impose a normality assumption on the data (Klugh, 1986). Using this method, Table 24 reports the Spearman rank order (Rho) correlation between the VAM measures and each of the PP sub-domains.

Table 24

Spearman's rho Correlations of PP Sub-Domains with VAM

			FFT Total Score (Pct.)	VAM	D1	D2	D3	D4
Spearman's rho	FFT Total Score (Pct.)	r	1.000	.255**	.900**	.871**	.911**	.805**
		Sig.	.	.000	.000	.000	.000	.000
		N	238	238	238	238	238	238
	VAM	r	.254**	1.000	.185**	.272**	.241**	.197**
		Sig.	.000	.	.004	.000	.000	.002
		N	238	238	238	238	238	238
	Domain 1	r	.900**	.185**	1.000	.707**	.817**	.706**
		Sig.	.000	.004	.	.000	.000	.000
		N	238	238	238	238	238	238
	Domain 2	r	.871**	.272**	.707**	1.000	.748**	.614**
		Sig.	.000	.000	.000	.	.000	.000
		N	238	238	238	238	238	238
	Domain 3	r	.911**	.241**	.817**	.748**	1.000	.621**
		Sig.	.000	.000	.000	.000	.	.000
		N	238	238	238	238	238	238
	Domain 4	r	.805**	.197**	.706**	.614**	.621**	1.000
		Sig.	.000	.002	.000	.000	.000	.
		N	238	238	238	238	238	238

**Correlation is significant at the 0.01 level (2-tailed).

Predictably, the rank correlations between each of the PP sub-domains and the total PP scores are positive significant ($p < .05$) with relatively strong values ranging from .805 to .911. In addition, the inter-correlations between PP domains is moderate to strong, ranging from a low of .614 (D4: Professionalism and D2: Classroom Environment) to .817 (D3: Instruction and D1: Planning). The correlation between the PP domain scores and the VAM measures is positive significant but uniformly weak in effect, ranging from .185 (D1: Planning) to .272 (D2: Classroom Environment).

RQ1A (e). To what degree are high, middle, and low VAM estimates of instructional effectiveness able to differentiate PP? The approach is ANOVA, means testing. The measures are t, F tests, and post hoc tests of group mean differences.

Arguably, the intent of the hypothesized evaluation construct is to distinguish and differentiate levels of TIQ. That is, the framework is specifically intended to rank order the instructional competency of classroom teachers across a continuum of performance. By utilizing quantitative measurement scales, the assumption is that teachers placed higher/lower on the scale possess higher/lower levels of instructional competence as informed by the evaluation system's two main components: ratings of PP and the academic growth of students (VAM). If each component contributes information to this inference, teachers placing higher on either scale are inferred to be better at their craft than those attaining at lower levels. Thus, teachers with higher academic growth (VAM) scores should also have attained higher PP scores.

However, it might be argued that the measurement precision of each component is not perfect. That is, the statistical models used to estimate student growth contain sampling error, specification error, and the achievement data itself contain error. In addition, each evaluator's knowledge and application of the PP rubrics is imperfect and inconsistent. Finally, it might be argued that numerous other factors influence the evaluation process, creating construct-irrelevant variability in the observed data.

Within this context, an alternative approach for exploring the criterion associations was undertaken that still attempted to evaluate the viability of the hypothesized construct. The question posed was as follows: Do teachers with relatively high/low VAM scores also display relatively high/low PP ratings? To assess this, the distribution of average median classroom VAM scores was placed onto a percentile distribution. From this distribution, teachers in the top 90^{th} percentile, the 45^{th} to 55^{th} percentile, and those in the bottom 10^{th} percentile categories were identified. Then for

each of the three growth-performance groups, the average PP rating was computed. Identification of the three groups was accomplished based on the combined VAM field using the Rank Cases module in SPSS (Version 21) applying the Fractional Rank method with ties assigned to the "Low" value.[9]

A one-way analysis of variance (ANOVA) approached was applied to test whether the mean PP scores differed across the three VAM performance categories. To the extent that statistically significant differences in mean PP scores are evident between the three growth-performance groups, the premise that the two components contribute similar information toward the underlying TIQ construct is supported.

This represents a less precise approach to assessing the association between the component measures because it does not utilize the continuous variability of the underlying scales. Rather, it groups the teachers into more generalized growth-performance levels and asks whether the PP ratings align with these designations.

Table 25 reports the distribution of teachers placed into the three growth-performance groups.

[9] VAM_Percentile <= 10.4 assigned to Low Group; VAM_Percentile >= 44.5 & VAM_Percentile < 55.5 assigned to Middle Group; VAM_Percentile >= 89.5 assigned to High Group.

Table 25

Teachers Assigned to VAM Percentile Group

Growth Performance		Frequency	Percent	Valid Percent
Valid	Low	24	10.1	31.6
	Mid	27	11.3	35.5
	High	25	10.5	32.9
	Total	76	31.9	100.0
Excluded		162	68.1	
Total		238	100.0	

As shown, 76 teachers were identified across the three VAM growth performance groups with 24 to 27 situated within any particular group. Overall, these members represent approximately 32% of the total population of Group A teachers (n = 238). Approximately 10% (n = 49) of all teachers located at the two extremes (High/Low) and another 10% in the middle group (Mid). Table 26 reports the descriptive statistics associated with the VAM and PP measures across the three growth-performance groupings.

Table 26

Descriptive Statistics by Growth-Performance Group

	Growth-Performance Group	PP Score	VAM Percentile
Count	Low	24	24
	Mid	27	27
	High	25	25
Mean	Low	.641901	36.323212
	Mid	.737581	49.823735
	High	.782810	61.790671
Median	Low	.673735	38.155992
	Mid	.747348	49.791045
	High	.766818	61.348927
Std. Deviation	Low	.179553	4.524455
	Mid	.104485	0.610142
	High	.092252	2.092618

The data indicate a steadily increasing mean PP score across growth-performance groups: .64 to .74 to .78 for the Low, Middle, and High groups, respectively. It is noted that there is a 10 point different between the Low and Mid group means and a 14 point difference between the Low and High group means. In contrast, there is only a four point difference between the Mid and High group means.

For the Mid and High groups, the reported mean and median values are generally close, suggesting a relatively balanced within-group distribution. However, the median for the Low group is three points above the mean, suggesting the influence of some outlier values. Finally, the standard deviations of the PP scores appear to decrease across groups and vary substantively in magnitude.

The presence of unequal variances (heteroscedasticity) would constitute a violation of a basic distributional assumption required for estimating unbiased f-test

statistics when employing ANOVA techniques (Ferguson & Takane, 1989). To examine this possibility, the Levene's test of equal variances was applied to the data: $L(2, 73) = 2.642$, $p = .078$. The size of the p-value ($p > .05$) results in a failure to reject the null hypothesis of equal variances. Based on this finding, use of a standard ANOVA procedure is warranted and post hoc test procedures assuming homogeneity of variance may be applied.

The Null hypothesis being examined by the ANOVA procedure assumes no significant difference ($p < .05$) between each of the mean PP scores across the three growth-performance groups. Table 27 reports the results of analysis.

Table 27

ANOVA Tests of Mean Differences: FFT Total Weighted Score (Pct.)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.253	2	.126	7.509	.001
Within Groups	1.230	73	.017		
Total	1.483	75			

The ANOVA results indicate a significant difference ($p < .05$) between mean PP scores across the three growth-performance groups, $F(2, 73) = 7.509$, $p = .001$. To examine the location of the differences, post hoc tests were conducted using the Tukey HSD (honestly significant difference) procedure (Stevens, 1996). The procedure compares all possible combinations of mean differences in an attempt to ascertain which combination is statistically different. The Tukey procedure uses a t-statistic and adjusts

for the error rate associated with making multiple comparisons. Table 28 reports the results from the Tukey post hoc tests of mean differences.

Table 28

Tukey HSD Multiple Comparisons Test of Mean Differences

(I) VAM Percentile Group ID	(J) VAM Percentile Group ID	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Low (m=.64)	Mid (M =.74)	-.0956805*	0.0364097	0.028	-0.182788	-0.008573
	High (M =.78)	-.1409088*	0.0370888	0.001	-0.229641	-0.052176
Mid (M=.74)	Low (m =.64)	.0956805*	0.0364097	0.028	0.008573	0.182788
	High (M =.78)	-0.0452283	0.0360222	0.425	-0.131409	0.040952
High (M=.78)	Low (m =.64)	.1409088*	0.0370888	0.001	0.052176	0.229641
	Mid (M =.74)	0.0452283	0.0360222	0.425	-0.040952	0.131409

*. The mean difference is significant at the 0.05 level.
Dependent Variable: FFT Total Weighted Score (Pct.)

Tukey post-hoc comparisons indicate that the mean PP score for the Low VAM-Performance group were significantly ($p < .05$) below that of the Mid and High groups. In contrast, the difference in mean PP scores between the Mid and High groups was not found to differ significantly. This suggests that VAM and PP indicators discriminate TIQ only at the lower range of the measurement scales.

Research question 1B: Content evidence (RQ1B). Research Question 1B Content Evidence, examines the suitability of VAM and PP components to adequately represent the hypothesized TIQ construct. For the PP component this involves examining the associations between the 22 elements comprising the Danielson Framework for

Teaching (FFT) and the structural consistency of the four specified domains. These attributes are evaluated using both quantitative and qualitative approaches including exploratory and confirmatory factor analytic techniques, use of the Lawshe Construct Validity Index, and reflections from stakeholders as to the framework's capacity for adequately identifying and distinguishing levels of instructional competency. The capacity of the VAM component as a measured of instructional quality is also examined based on stakeholder feedback. An outline of the supporting research questions associated with RQ1B is provided in Table 29.

Table 29

Supporting Questions for Research Question 1B

Item	Description	Approach	Measure
RQ1B (a):	To what degree do empirical ratings of PP correspond with the theoretical FFT construct?	Exploratory & Confirmatory Factor analysis	Factor Extractions; Factor Loadings; χ^2; AIC; α
RQ1B (b):	To what degree does the factor analytic structure of empirically-based PP scores differ between less experienced and more experienced teachers?	Exploratory & Confirmatory Factor analysis	Factor Extractions; Factor Loadings; χ^2; AIC; α
RQ1B (c):	To what degree do the 22 elements contained within the theoretical FFT framework adequately represent the latent TIQ construct?	Lawshe CVI Questionnaire, Stakeholder Interviews	CVI; Coded Interview Responses
RQ1B (d):	Do perspectives differ among stakeholders regarding the capacity for VAM and PP measures to adequately represent and differentiate the instructional quality of classroom teacher?	Stakeholder Interviews; Stakeholder Questionnaire	Coded Interview Responses; Questionnaire Item Responses

The Danielson PP scale. As mentioned, qualified evaluators assign a performance rating to each behavioral professional practice (PP) component within the Danielson Framework for Teaching (FFT). Ratings are based on observations of classroom instruction and other artifacts/documents associated with the instructional process. Each element identifies a specific behavioral characteristic associated with instructional practice. Under the evaluation system, the 22 element ratings (with possible values between zero and three) are summed to form a total PP score ranging from zero to 66. The TIQ construct posits that higher individual element ratings, and higher total PP scores, represent greater degrees of instructional competency.

To investigate these assertions, four supporting research questions are posited:

- *To what degree do empirical ratings of PP correspond with the theoretical FFT construct?*
- *To what degree does the factor analytic structure of empirically-based PP scores differ between less experienced and more experienced teachers?*
- *To what degree do the 22 elements contained within the theoretical FFT framework adequately represent the latent TIQ construct?*
- *Do perspectives differ among stakeholders regarding the capacity for VAM and PP measures to adequately represent and differentiate the instructional quality of classroom teacher?*

Examination of each supporting research question is provided below.

RQ1B (a). To what degree do empirical ratings of PP correspond with the theoretical FFT construct? The approach is exploratory and confirmatory factor analysis. The measures are factor extractions, factor loadings, χ^2, AIC, α.

This research question utilizes both confirmatory and exploratory factor analytic techniques to assess the latent factor structure posited by the Danielson Framework for Teaching (FFT). The intent is to investigate the degree to which the measured data conform to the posited four-domain factor structure and the independence (discrimination) existing between the behavioral constructs.

The initial analytic approach involves constructing a pre-specified latent factor model and assessing the degree to which the data fit that model (a confirmatory factor analytic approach). A well-fitting model would support the premise of a four-factor structure which provides information on the unique characteristics of teaching. In contrast, a poor fitting model would raise fidelity question regarding the structural framework and the suitability of the component scores to assess attributes of teaching.

The second analytic approach does not impose an a priori latent structure on the measured data. Rather the presence of one or more latent factors is freely revealed based on the strength of the inter-element correlations (an exploratory factor analytic approach). Here, the alignment between the data-exposed latent factors with the hypothesized framework is examined. Conformity between the two provides support for the posited constructs while lack of conformity raises inferential questions.

Confirmatory factor analysis. To begin, a confirmatory factor analysis approach was used to assess the degree to which the measured PP ratings conformed to a pre-specified four-domain latent model. The theoretical structure of the FFT is represented by the diagram presented in Figure 29.

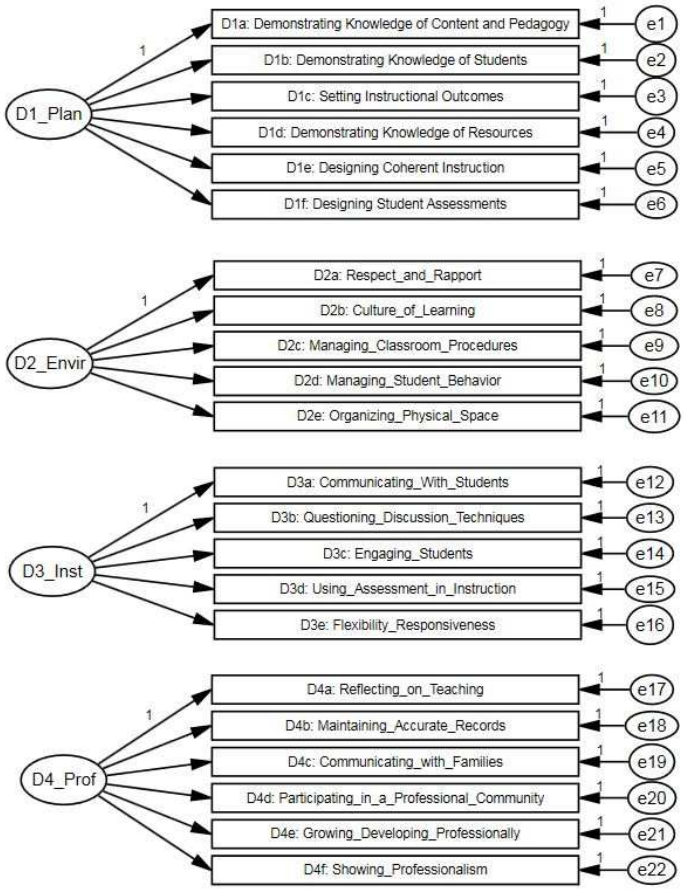


Figure 29. Theoretical structure of the Danielson Framework for Teaching (uncorrelated model).

Figure 29 specifies four latent (unobserved) factors: *D1_Plan* = Domain 1: Planning and Preparation; *D2_Envir* = Domain 2: Classroom Environment; *D3_Inst* = Domain 3: Instruction; and *D4_Prof* = Domain 4: Professional Responsibilities. Each of these factors affect between five and six measured (observed) variables as posited by the FFT framework. Errors terms are shown for each variable to reflect the amount of variance left unexplained by the associated latent factor. The design depicted in Figure 29 (above) represents the most restrictive representation of the FFT framework in that each of the four PP domains are independent: the covariance between each factor is explicitly

set to zero. This is "most restrictive" because it faithfully operationalizes the claim made by the framework's author, Charlotte Danielson (Danielson, 2007), that

> …each of the framework's four domains refer to a distinct aspect of teaching … the components within each domain form a coherent body of knowledge and skill that can be the subject of focus independent of the other domains. (p. 26)

A less restrictive interpretation of the framework is presented in Figure 30.

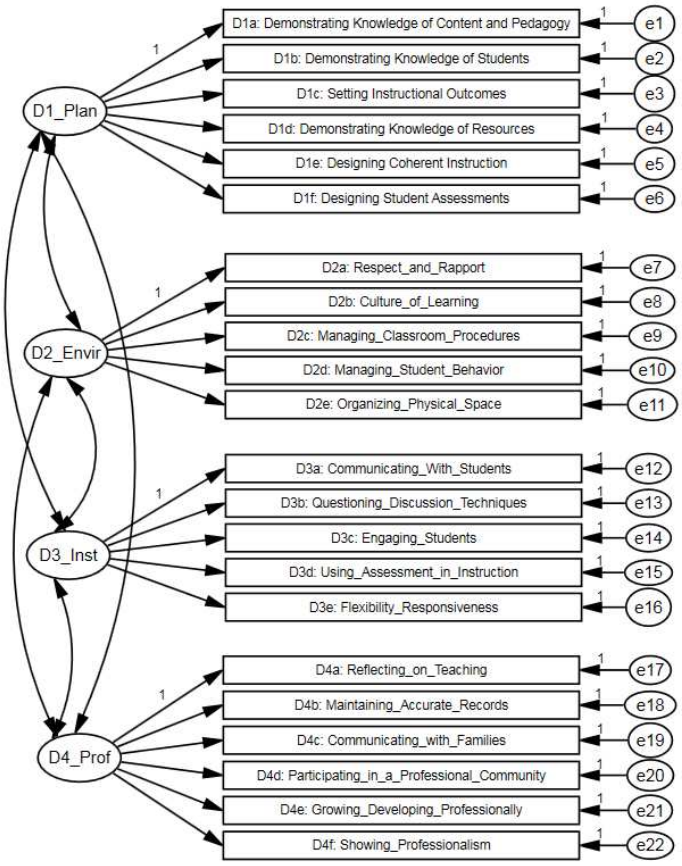


Figure 30. Theoretical structure of the Danielson Framework for Teaching (correlated model).

Arguably, this represents a more realistic representation of instructional activities by allowing for co-variation to exist between each of the four domains. With this view,

the instructional activities of a classroom teacher are not completely independent (uncorrelated) with other behaviors (for example, classroom environment and instructional preparation). Allowing the domains to co-vary acknowledges that facets of instructional behavior may be inter-related notwithstanding Danielson's claim of independence. It also provides a means to test the structural assumptions of the FFT framework between a very restrictive and less restrictive context.

Element correlations. The Mplus software program (version 7.11) was used to estimate CFA parameter values presented in this study. The command syntax associated with uncorrelated and correlated factor models is presented in Appendix N.

By default, Mplus computes polychoric correlations when ordered categorical data and WLSMV estimators are declared (Muthen & Muthen, 2012).[10] **Appendix I through K** provides the FFT element polychoric correlations, standard errors, and z-values for evaluating correlation significance among the data used in the Mplus CFA analysis. Overall, the average correlation among the 231 elements was .64 (standard deviation = .087) ranging from a low of .37 to a high of .85. All correlations were significant at a $p < .05$ level (Uebersax, 2006).[11]

Uncorrelated, four factor, CFA model. An uncorrelated, four factor, CFA model was estimated using Mplus Version 7.11. Because of the ordinal nature of the FFT rating scale, a weighted least squares (WLSMV) estimator was applied during the estimation process to correct for bias in estimating the standard errors of the factor loadings

[10] Information provided L. K. Muthen on the Mplus discussion board provided the following clarification: "The WLSMV estimator first computes a sample correlation matrix (tetrachoric, polychoric) and then fits the model to that, thereby estimating the model parameters. So the fitting of the model is similar to what is done if the outcomes had been continuous. No factor score estimation is involved in this, but the parameters are estimated directly." See http://www.statmodel.com/discussion/messages/23/646.html.
[11] The significance decisions for polychoric correlations are based on a z-value transformation outlined in Uebersax (2006) using p = 1.96 criteria. Mplus produces the correlation and associated standard errors from which to compute the z-ratio. The complete set of correlation, standard error, and z-value tables are presented as an appendix to this report.

(Muthen, 1991, 1994; Muthen & Muthen, 2012). The variances for each of the latent factors were constrained to one in order to set the measurement scale. All error variances were freely estimated as were each of the dependent variable factor loadings. No error covariances were presumed.

Under CFA, model fit indices convey the degree to which the pre-specified latent factor structure is able to reproduce the variance-covariance patterns observed in the measured data. By comparing the model's parameter-estimated variance-covariance matrix with the observed population variance covariance information, a variety of test statistics and similarity measures may be constructed. Table 30 report selected model fit statistics for the uncorrelated CFA model.

Table 30

CFA Model Fit Statistics for Uncorrelated Model

Statistic	Value	Degrees of Freedom	*P*-Value	Lower CI	Upper CI	Standard	Comment
Chi-Square	8109.019	209	0.000			P <= .05	Poor Fit
RMSEA*	0.399		0.000	0.391	0.406	<= .05	Poor Fit
CFI**	0.155					> .95	Poor Fit
WRMR	8.603					< 1.0	Poor Fit

Note. Number of Free Parameters: 82
* Root Mean Square Error Approximation: *p*-Value tests probability <= .05;
** Comparative Fit Index; *** Weighted Root Mean Square Residual

For the uncorrelated model, the model fit indices indicated a poor correspondence between the hypothesized latent factor structure and the measured data. The chi-square value, χ^2 (209, N = 238) = 8109.019, $p < .001$ is significant at a $p = .05$ level. Here, the

null hypothesis (H_O) assumes that the reproduced and population variance-covariance matrixes are the same. Thus, a non-significant *p*-value (i.e., $p > .05$) is desirable so as not to reject the null hypothesis (H_O) that there is no difference between the measured data and the hypothesized structural model. In this case, $p < .05$ and H_O is rejected suggesting that the structural (uncorrelated) model specification is not represented well by the data.

Because the value of the chi-square statistic may be influenced by large sample sizes, a number of additional model fits statistics are referenced (Brown, 2006; Browne & Cudeck, 1992). The RMSEA (root mean square error approximation) statistic assess model fit after adjusting for sample size effects (degrees of freedom). A RMSEA value less than or equal to .05 along with a significant *p*-value (*p*-value < .05) is desirable to indicate a good fit between the hypothesized factor structure and the measured data (Brown, 2006; Browne & Cudeck, 1992). For the uncorrelated model, the reported RMSEA is significantly above the desired threshold (RMSEA = .399, $p < .001$) suggesting that the reproduced variance-covariance deviates substantively from population values (rejecting H_O). In addition, the 95% confidence interval surrounding the RMSEA values does not include the .05 criteria and thus does not support the premise that the data fit the specified model.

The CFI (comparative fit index) statistic compares the chi-square of the estimated CFA model with that of a base-level (null) model (Brown, 2006; Hu & Bentler, 1999). Here, the base model assumes zero covariance between the measured variables. In this way, the CFI is an incremental index that assesses the degree to which the a priori structural model improves the fit over the null model. Generally, the CFI index should exceed .95 to provide support of good model fit (Brown, 2006; Hu & Bentler, 1999;

Marsh, Hau, & Wen, 2004). For the uncorrelated model, the CFI is reported to be .155 indicating poor fit between the measured data and the hypothesized factor structure.[12]

The factor loadings reported for the uncorrelated CFA model represent the correlation between the latent factors and each of its associated measured variables. The r-square indicates the amount of common variance explained by the factor. Table 31 reports the standardized factor loadings for the estimated uncorrelated four-factor model.

[12] It is noted that Mplus does not produce a SRMR statistic for models using estimators based on mean structures such as with WLSMV (See discussion posted at the Mplus web site: http://www.statmodel.com/discussion/messages/9/5810.html?1288140177).

Table 31

Standardized Factor Loadings for the Uncorrelated Four-Factor Model

DOMAIN1 By	Estimate (r)	Standard Error	Est./S.E.	Two-Tailed P-Value	R-Square
D1A	0.867	0.032	27.161	.000	0.752
D1B	0.844	0.035	24.445	.000	0.712
D1C	0.851	0.034	25.054	.000	0.724
D1D	0.838	0.04	20.96	.000	0.702
D1E	0.875	0.032	27.397	.000	0.766
D1F	0.870	0.034	25.945	.000	0.757
DOMAIN2 By					
D2A	0.841	0.036	23.276	.000	0.707
D2B	0.872	0.036	24.166	.000	0.760
D2C	0.844	0.037	22.845	.000	0.712
D2D	0.766	0.046	16.843	.000	0.587
D2E	0.722	0.059	12.212	.000	0.521
DOMAIN3 By					
D3A	0.736	0.049	14.96	.000	0.542
D3B	0.884	0.034	25.967	.000	0.781
D3C	0.869	0.036	24.166	.000	0.755
D3D	0.808	0.044	18.444	.000	0.653
D3E	0.752	0.043	17.667	.000	0.566
DOMAIN4 By					
D4A	0.803	0.04	19.889	.000	0.645
D4B	0.831	0.039	21.137	.000	0.691
D4C	0.803	0.038	20.875	.000	0.645
D4D	0.779	0.042	18.702	.000	0.607
D4E	0.885	0.038	23.500	.000	0.783
D4F	0.895	0.029	30.450	.000	0.801

As shown, all of the individual factor loadings (correlations) are significant and range in value between .722 and .895 with a mean of .829. The range of R^2 = .521 - .801 with a mean = .689. Thirteen (59%) out of the 22 R^2 values exceed .70 indicating strong associations between the majority of the elements and their respective factors.

Review of residual variance-covariance information provides a sense of how well the model parameter estimates are able to reproduce the correlations present in the population data. Lower residual correlations reflect better fit while higher values represent poorer fit. There is no set interpretation for the size of residual correlations. However, values close to zero are desired for supporting the proposition that the data fits the model well.

Appendix L presents the residual correlation matrix for the uncorrelated four-factor model. The data report a mean residual correlation of .45, ranging from a low of -.12 to a high of .85. One hundred and sixty-eight (73%) of the 231 residual correlations exceed .50, 111 (48%) exceed .60, and 39 (17%) exceed .70. This suggests that the uncorrelated four factor model is a poor fit to the data and is consistent with the findings from the model fit test statistics discussed above.

The CFA process also produces estimated factor scores for all individuals in the data base. The correlation between the factors provides an indication of construct discrimination. Large correlations suggest a lack of discriminant inference. That is, if you know an individual's score for one factor, you are able to derive the score on the second with reasonable accuracy. Thus, high correlations obfuscate discrete interpretation of the component factors. Table 32 reports the correlation matrix for the factor scores generated from the four-factor uncorrelated model.

Table 32

Uncorrelated Four Factor Model - Correlation of Factor Scores

	DOMAIN 1	DOMAIN 2	DOMAIN 3	DOMAIN 4
DOMAIN 1	1.000			
DOMAIN 2	0.768	1.000		
DOMAIN 3	0.855	0.794	1.000	
DOMAIN 4	0.779	0.697	0.697	1.000

The data indicate substantive correlations (.70 to .86) between the factor scores generated by the model. This suggests that the conceptual discrimination between factors may be limited.

Correlated, four-factor, CFA model. A second CFA model was estimated allowing for co-variation to exist between the four hypothesized factors. As mentioned, this is a less restrictive form of the first structural model because it does not impose a priori constraints on the association between the factors. Table 33 reports the model fit statistics for the correlated model.

Table 33

CFA Model Fit Statistics for Correlated Model

Statistic	Value	Degrees of Freedom	P-Value	Lower CI	Upper CI	Standard	Comment
Chi-Square	369.691	204	0.000			P <= .05	Poor Fit
RMSEA*	0.058		0.073	0.049	0.068	<= .05	Close Fit
CFI**	0.982					> .95	Good Fit
WRMR	0.972					< 1.0	Good Fit

Note. Number of free parameters: 87
* Root Mean Square Error Approximation: P-Value tests probability <= .05; **Comparative Fit Index; *** Weighted Root Mean Square Residual

The fit statistics for the correlated factor model are improved over those reported for the uncorrelated model. The chi-square value is substantively lower, $\chi^2(204, N = 231) = 369.691$, $p < .001$, even though it remains statistically significant at the $p = .05$ level (again rejecting H_O that the data adequately fit the hypothesized factor structure). The RMSEA value of .058 remains significant ($p = .073$) but is substantively below that reported for the uncorrelated model (0.399) suggesting improvement in model specification. Finally, the CFI value of .982 exceeds the recommended criteria of .95 as an indicator of good fit and the WRMR value of .972 also provides an indication of suitable fit (Hu & Bentler, 1999).

With regard to the RMSEA statistic, Browne and Cudeck (1992) note that the restrictive criteria imposed by the null hypothesis tests whether or not the RMSEA = 0. They suggest that the criteria of exact model fit (RMSEA = 0) is too restrictive and untenable in any condition other than theoretical (Browne & Cudeck, 1992, p. 231). Rather, the authors proposed a less restrictive interpretation. They propose that the desired RMSEA be approximately .05 and values close to that threshold be interpreted as evidence of "Close Model Fit." For the correlated CFA model, RMSEA = .058 is complemented by a parameter confidence interval of .049 to .068, the lower bound of the confidence interval inclusive of the .05 criteria of Close Model Fit.

From these values, it is evidenced that by relaxing the condition of independence between the latent factors (FFT behavioral domains), the fit of the model to the measured data improves. Table 34 reports the individual standardized loadings between each factor and their component variables.

Table 34

Standardized Factor Loadings for the Correlated Four-Factor Model

	Estimate (r)	Standard Error	Est. / S.E.	Two-Tailed P-Value	R-Square
DOMAIN1 By					
D1A	0.891	0.028	32.093	.000	0.794
D1B	0.870	0.029	30.156	.000	0.757
D1C	0.893	0.027	32.867	.000	0.797
D1D	0.871	0.037	23.851	.000	0.759
D1E	0.874	0.029	30.259	.000	0.764
D1F	0.892	0.030	29.781	.000	0.796
DOMAIN2 By					
D2A	0.818	0.034	24.033	.000	0.669
D2B	0.903	0.028	32.808	.000	0.815
D2C	0.815	0.034	24.220	.000	0.664
D2D	0.762	0.043	17.646	.000	0.581
D2E	0.756	0.046	16.602	.000	0.572
DOMAIN3 By					
D3A	0.849	0.035	24.457	.000	0.721
D3B	0.865	0.031	28.059	.000	0.748
D3C	0.891	0.030	29.267	.000	0.794
D3D	0.900	0.033	27.280	.000	0.810
D3E	0.805	0.040	20.239	.000	0.648
DOMAIN4 By					
D4A	0.830	0.036	22.978	.000	0.689
D4B	0.825	0.035	23.406	.000	0.681
D4C	0.789	0.036	22.226	.000	0.623
D4D	0.760	0.042	18.243	.000	0.578
D4E	0.917	0.030	30.338	.000	0.841
D4F	0.870	0.027	31.796	.000	0.757

For the correlated model, all of the individual factor loadings (correlations) are significant and range in value between .756 and .917 with a mean of .848. The R^2 values range from .572 to .841 with a mean = .721. As before, thirteen (59%) of the 22 R^2 values

exceed .70 indicating strong associations between the majority of the elements and their respective factors.

Appendix M presents the residual correlation matrix for the correlated four-factor model. The data suggest a much improved fit specification than reported for the uncorrelated model. Table 35 summarizes the comparative information for the two (uncorrelated and correlated) residual correlations matrices.

Table 35

Comparative Summary Residual Correlation Statistics

Residual Matrix Statistic	Uncorrelated Factor Model (#)	Correlated Factor Model (#)
Number of r	231	231
r-Mean	.45	-.01
r-Minimum	-.12	-.23
r-Maximum	.85	.14
r-Range	.97	.37
r: < 0	26	121
r: Exceed .1	181	13
r: Exceed .2	181	0
r: Exceed .3	181	0
r: Exceed .4	180	0
r: Exceed .5	168	0
r: Exceed .6	111	0
r: Exceed .7	39	0
r: Exceed .8	5	0
r: Exceed .9	0	0

For the uncorrelated model, the residual correlations were predominantly positive and moderate-to-large in magnitude (M = .45, 73% above 0.50) with factor score

correlations between .70 and .86. In contrast, the mean residual correlations for the correlated model approached zero (M = -.01) with a predominance (n = 121, 52%) less than zero. However, the magnitude of the correlated model's negative correlations were relatively weak with 104 (86%) less than -.10: 56 (46%) of the values were between zero and -.05, and another 48 (40%) between -.50 and -1.0. Only 17 of the negative values were in excess of -.10. These data indicate that the correlated four-factor CFA model reproduces the covariances of the measured data more accurately than the uncorrelated specification.

In addition to the individual variable factor loadings, the correlated CFA model was freed to estimate covariances (correlations) between each of the four hypothesized factors. As mentioned, larger covariances indicate greater degrees of association and reduced inferential discrimination. That is, it is presumed that each of the latent factors represent independent behavioral (instructional) domains that may be discretely measured and interpreted. Evidence of this would manifest in relatively low factor covariances, weak associations between factors. To the extent that the factor covariances are large, the scores received by a teacher on one factor provide information on behavioral scores received on the corresponding factor. In this way, large covariances would suggest a lack of inferential discrimination.

It must be noted that during initial estimation, Mplus produced an estimated factor covariance between factors one and three (r-estimated = 1.043) that exceeded the correlation statistic's theoretical upper bound of 1.0. This resulted in a non-positive definite covariance matrix condition (Muthen & Muthen, 2012). A positive definite criterion requires that every principal sub-matrix have a positive determinant and all

matrix eigenvalues be greater than zero (Brown, 2006). Presence of a non-positive definite condition prevents inversion of the data matrix necessary when executing the Mplus CFA weighted least squares (WLS) estimation procedure (Wothke, 1993; Muthen & Muthen, 2012; Rigdon, & Ferguson, 1991). Muthen (2004) notes that for this type of non-positive condition

> … in most cases, you will find a negative residual variance or a correlation greater than one. This can also be caused by dependencies in your data. You can ask for TECH4 in the [Mplus] OUTPUT command to see the model estimated correlations among the latent variables.

In essence, the correlations between the measured variables in domains one and three are large enough to create a statistical possibility that the upper bound of the model's reproduced (estimated) correlation exceeds 1.0.

Subsequent review of the Tech4 output in Mplus confirmed the presence of a reproduced correlation greater than one (r-estimate = 1.043) between the first and third factors. Muthen (2002) comments that

> …the [non-positive definite] message means that the full weight matrix is not positive definite which it should be when the necessary inversion of the matrix takes place for WLS. [However] WLSMV uses only the diagonal of this matrix and therefore does not run into this problem.

For this study, the four-factor correlated model utilizes a WLSMV estimator to properly account for the categorical (non-multivariate normal) nature of the rating scales. To properly represent the strong correlation underlying the data and permit completion of the Mplus estimation process, the covariance between the first and third factors was set, a priori, to a value of .90. Under this restriction, Table 36 reports the covariances (correlations) estimated between each of the four latent factors for the correlated CFA model.

Table 36

Correlated Model Factor Correlations

Factor	Estimate (r)	Standard Error	Est. / S.E.	Two-Tailed P-Value	R-Square
DOMAIN 1 With					
DOMAIN 3	0.900*	na	na	na	0.810*
DOMAIN 2 With					
DOMAIN 1	0.904	0.025	36.568	.000	0.817
DOMAIN 3 With					
DOMAIN 2	0.919	0.02	44.885	.000	0.845
DOMAIN 4 With					
DOMAIN 1	0.891	0.024	37.162	.000	0.794
DOMAIN 2	0.845	0.031	27.262	.000	0.714
DOMAIN 3	0.822	0.028	29.724	.000	0.676

* Covariance restricted at .90

The data indicate strong correlations (r = [.82 to .92]) between the latent factors. All estimated covariances are significant at the p = .05 level. The squared correlations indicate between 68% to 85% common variance. As with the factor score correlations revealed by the uncorrelated model, the data for the correlated model again raise questions on the ability of the individual domain scores to inform on independent aspects of instructional practice. Figure 31 displays a graphical representation of the correlated CFA model with restricted covariance imposed between factors one and three.

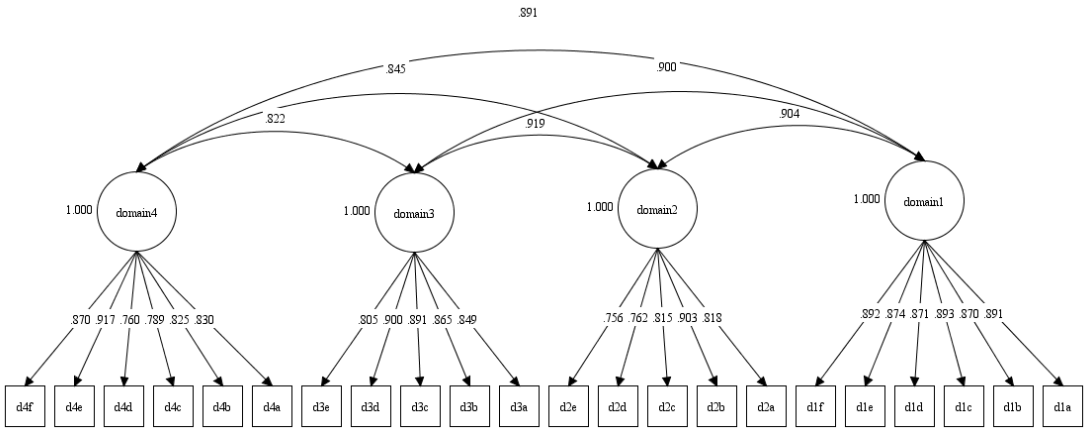


Figure 31. Parameter estimates for the Correlated CFA Model.

Results from the CFA examination suggest that a fully restricted (uncorrelated factor) model does not correspond well with the measured data. Relaxing the assumption of inferentially independent behavioral domains greatly improves model fit. However, both specifications suggest an inability to inferentially discriminate between the individual domain constructs.

Exploratory factor analysis. To further investigate the nature of the hypothesized FFT four-factor framework, the CFA analysis was followed up with an exploratory factor analysis (EFA). EFA differs in intent from the CFA in that the latter pre-specifies a latent factor structure along with the explicit alignments of each measured variable (Browne, 2006). Here, the intent is to validate the degree to which the hypothesized structure is supported by empirical data.

No such a priori assumptions are made under EFA. Under EFA, the measured data are allowed to *reveal* the latent factor structure (Pett, Lackey, & Sullivan, 2003). Assigning interpretive meaning to each *revealed* factor is dependent on the meanings

ascribe to the measured variables based on the strength of correlation. Thus, EFA is inherently an exploratory process while CFA is inherently confirmatory. As applied herein, the EFA approach asks whether the measured data, if allowed to associate without restriction, supports both the presence of a hypothesized four-factor structure and the associated inferential interpretations posited under the Danielson FFT framework. This would be evidenced by the following observations: First, the EFA process reveals four distinct latent factors; Second, the alignment of measured variables to latent factors permits interpretations consistent with the domain representations posited by the Danielson Framework.

To explore this, an EFA was conducted using both SPSS version 21 and Mplus version 7.11. Each software application provides particular strengths related to data processing, data exploration, and estimation under conditions of non-multivariate normal data. SPSS was primarily used to examine factor extraction criteria while Mplus was employed to compute the final EFA parameter estimates.

Factor extraction. The first step in the EFA approach is to determine the number of latent factors to extract from the data set. There are two main EFA approaches for examining factor extraction. The first is principal components analysis (PCA). PCA is primarily a data reduction technique, concerned with reducing a large number of correlated variables into a smaller set of uncorrelated measures for use in subsequent analytic routines (Pett et al., 2003). By design, PCA is a total variance technique that does not parcel out common and error variance. It is therefore less appropriate for applications focused on factor interpretation and/or reformulation. The second approach, common factor analysis (CFA – not to be confused with confirmatory factor analysis),

explicitly parcels total variance into common and unique (error and specification) components. The subsequent identification of latent factors is inferred based on the amount of shared (common) variance. It, therefore, permits more construct-informed interpretations. For this study, computational approaches related to common factor analysis are employed. Specifically, the method of principle axis factoring (PAF) is primarily used to deconstruct the variance components within the data set (Pett et al., 2003).

To investigate the potential number of factors to extract, the following analytic approaches and identification criteria were applied to the data (Table 37):

Table 37

Factor Extraction Criteria

Analytic Method	Factor Extraction Criterion
Guttman-Kaiser Eigenvalue Criterion	Eigenvalues > = 1.0
Scree Plot	Δ Slope (Elbow)
Percent Variance Explained	Common variance accounted > 50 %
Minimum Average Partial (MAP) Test	Maximized Proportional Common Variance
Parallel Analysis	Non-Random Eigenvalues
Chi-Square (χ^2)Test	P-value not sig

Each if these approaches provide indications on the number of dominant latent factors underlying the data. By applying different analytic approaches, the researcher is afforded more information from which to make the extraction decision.

To begin the EFA process, Pett et al. (2003) note that "… it is important to determine if there are sufficient numbers of significant correlations among the items to justify undertaking a factor analysis…" (p. 74). They caution that the presence of excessively high item-correlations ($r > .8$) create issues related to multi-collinearity that limit the ability to inferentially differentiate between the dominant latent factors. In contrast, a large number of excessively low correlations ($r < .3$) limit the ability to clearly discern dominant latent factors. That is, the EFA analysis may identify as many factors as there are variables in the data set.

Recall that the data's polychoric correlation matrix was computed as part of the initial CFA analysis. Appendix I through K provided the FFT element polychoric correlations, standard errors, and *z*-values for evaluating correlation significance. Review of the information indicated the average correlation among the 231 elements was .64 (standard deviation = .087) ranging from a low of .37 to a high of .85. All correlations were significant at a $p < .05$ level. Examination of the individual correlation values reveal six (3%) above the threshold of .80 and no values less than .30.

Three additional methods for testing the suitability of the data to support EFA analysis include computing: (1) the value of the data matrix' determinant, (2) the Bartlett's test of sphericity (BTS), and (3) the Kaiser-Meyer-Olkin test (KMO; Pett et al., 2003). To conduct an EFA using PAF (a regression-based approach), the determinant of the data matrix cannot be negative or zero (i.e., non-positive definite). This condition

would prohibit inversion of the data matrix, a computation required by many statistical and regression-based procedures.[13]

The BTS statistic tests the null hypothesis (H_o) that the data's correlation matrix is an identity matrix (i.e., zero correlations between the measured variables). This condition would negate the need to undertake an EFA which is inherently intended to evaluate the size and clustering of inter-item correlations. The BTS statistic is distributed as a chi-square. Rejection of H_o would indicate that the data's correlation matrix is not an identify matrix.

Finally, the KMO assesses the degree to which inter-item correlations are composed of common (shared) variance. Larger values reflect stronger common-variance association between variables. To the degree that inter-item covariances are predominantly not due to error, or influence by other items, the EFA process is free to generate inferentially meaningful factor (clusters of items that share common variance). A desirable value is the KMO statistic is equal or above .90 (Pett et al., 2003, p. 78).

Table 38 report the results for each of the three EFA suitability tests discussed above. Each test was generated using SPSS version 21 on the original evaluation rating data matrix.

[13] From Muthen & Muthen, http://www.statmodel.com/discussion/messages/8/289.html?1336168023 : "The default for EFA in Mplus is the unweighted least squares estimator for which positive definiteness is not required or checked. If you use the ML estimator in Mplus, it will check for positive definiteness of the sample correlation matrix."

Table 38

EFA Tests of Sample Adequacy

Test Name	Value/Test Statistic	Criterion	Finding
Determinant	3.021E-006	Not Zero	Very Close to Zero
BTS	χ^2 (231, N=238) = 2908.490, $p < .001$	Reject H_o	Reject H_o
KMO	.960	> .900	Very Good

Both the BTS and KMO values support application of the EFA analysis. However, the determinant of the correlation matrix is very close to zero indicating the presence of high inter-item collinearity among selected variables.

The initial EFA was conducted on the individual FFT element rating data applying a principal axis (common) factoring method using SPSS version 21 and a augmented weighted least squares (WLSMV) estimator in Mplus version 7.11. For the SPSS PAF extraction, all possible factors (21 out of 22) were extracted in order to generate a complete descriptive summary. Table 39 reports the factor eigenvalues for the total (initial eigenvalues) and common PAF (extracted sums of squared loadings) variance components.

Table 39

Component Variance (Eigenvalues) for Full Factor Extraction: Total Variance Explained

	(SPSS EFA - Continuous)						(Mplus EFA WLSMV – Ordinal)	
	Initial Eigenvalues			PAF Extraction Sums of Squared Loadings				
Factor	Eigenvalue	% of Variance	Cumulative %	Eigenvalue	% of Variance	Cumulative %	Mplus Eigenvalues: Total	% of Variance (Est.)
1	10.719	48.722	48.722	10.436	47.438	47.438	14.548	66.127%
2	1.256	5.709	54.432	.964	4.384	51.822	1.333	6.059%
3	.954	4.334	58.766	.687	3.122	54.945	.824	3.745%
4	.877	3.989	62.755	.568	2.580	57.525	.808	3.673%
5	.809	3.676	66.431	.474	2.152	59.677	.687	3.123%
6	.709	3.224	69.654	.393	1.786	61.463	.549	2.495%
7	.640	2.910	72.564	.319	1.452	62.915	.475	2.159%
8	.569	2.584	75.148	.284	1.291	64.206	.384	1.745%
9	.534	2.428	77.576	.238	1.084	65.290	.355	1.614%
10	.514	2.336	79.913	.225	1.021	66.310	.327	1.486%
11	.493	2.239	82.152	.204	.929	67.240	.309	1.405%
12	.454	2.065	84.217	.181	.823	68.062	.285	1.295%
13	.450	2.047	86.264	.171	.776	68.838	.243	1.105%
14	.418	1.900	88.164	.134	.608	69.445	.216	0.982%
15	.401	1.822	89.986	.115	.525	69.970	.187	0.850%
16	.396	1.802	91.788	.103	.469	70.439	.157	0.714%
17	.371	1.685	93.473	.046	.210	70.649	.130	0.591%
18	.326	1.482	94.955	.034	.155	70.804	.087	0.395%
19	.316	1.435	96.390	.019	.086	70.890	.060	0.273%
20	.293	1.331	97.721	.015	.067	70.958	.041	0.186%
21	.268	1.219	98.940	.002	.008	70.966	.024	0.109%
22	.233	1.060	100.000	na	na	na	-.031[14]	na

Note. Extraction Method: Principal Axis Factoring.

The SPSS initial and PAF-extracted variance measures suggest the presence of a single dominate factor (Factor 1) accounting for 49% of total variance and 47% of the common variance among the measured variables. Similarly, the Mplus WLSMV estimates indicate a single dominant factor (66% explained variance). All remaining factors individually contribute an additional 6% or less of explained variance.

[14] Since the data matrix is composed of ordinal scale ratings, Mplus computes polychoric correlations from which the EFA analysis is based. Large polychoric correlations may result in a non-positive definite condition which in turn may lead to negative eigenvalues. Responding to this condition, Muthen and Muthen (http://www.statmodel.com/discussion/messages/8/205.html?1348513453) comment "… I think this is ignorable. With categorical variables and WLSMV, you work with tetrachoric and polychoric correlations which are computed for pairs of variables at a time and therefore can produce a non-positive definite sample correlation matrix - which has some negative eigenvalues. You can still get a pos-def model-estimated correlation matrix. If the model fits well to this sample correlation matrix, you can view the situation as the non-pos def sample correlation matrix was not "significantly non-pos def." There have been ideas in the literature about deleting the eigenvalues and eigenvectors for the negative eigenvalues and recreating the sample correlation matrix this way, smoothing it, and then fitting the model, but I am not sure that is an important improvement. If you use ML instead, this issue does not come up because ML does not fit the model to those sample correlations…" (http://www.statmodel.com/discussion/messages/8/205.html?1348513453).

Extraction criteria - Guttman-Kaiser (eigenvalues > = 1.0). The Kaiser-Guttman extraction criteria suggest selection of factors based on the size of their eigenvalues. Eigenvalues represent the proportional amount of variance explained by the particular factor (Yeomans & Golder, 1982; Brown, 2006; Pett et al., 2003)). Eigenvalues greater than one indicate that proportionally more of the total variance is explained by that particular factor. Eigenvalues less than one reflect a greater degree of unexplained variance associated with that factor. Yeomans and Golder (1982) note that

> …the technique is justified in the original Guttman article in terms of it providing a lower bound for the number of common factors underlying a correlation matrix of observed variates having unities in the main diagonal. More intuitively the argument has been advanced that no component "explaining" less than the variance of an original variate can be deemed to represent a significant source dimension… (p. 222)

This comment is informed by noting that EFA procedures transform the original units of measure into standard scores whose variance is equal to 1.0 for any given variable and the sum of these standardized variances (total variance) is equal to the total number of variables. If each variable's standardized variance contributes a unit value of 1 to total variance, then a factor whose eigenvalue is 1.0 is accounting for the same amount of total variance as the original item. Brown (2006) explains "… because the goal of EFA is to reduce a set of input indicators, … if an eigenvalue is less than 1.0, then the corresponding factor accounts for less variance than the indicator (whose variance equals 1.0)…" (p. 26). Similarly, an eigenvalue greater than 1 indicates that the factor explains more proportional variance than any given item. The rational for the Guttman-Kaiser criteria is to select factors that explain more of the variance than that contributed by any single item (an eigenvalue > 1.0). A criticism of this approach is that it is somewhat

arbitrary and does not involve any inferential considerations related to the aligned variables.

Referring to Table 39 above, the eigenvalues reported for the SPSS initial extraction and the Mplus extraction identify two factors meeting the Guttman-Kaiser criteria. In addition, the SPSS PAF eigenvalue for the second factor (.964) is very close to the criterion threshold. This information suggests that two factors be extracted from the FFT rating data.

Extraction criteria - scree plot (Δ slope (elbow). Another approach to identifying the number of factors to extract is to examine a graphical depiction of eigenvalues. This is known as a scree plot. Figure 32 displays the (SPSS) scree plots generated by SPSS and Mplus.

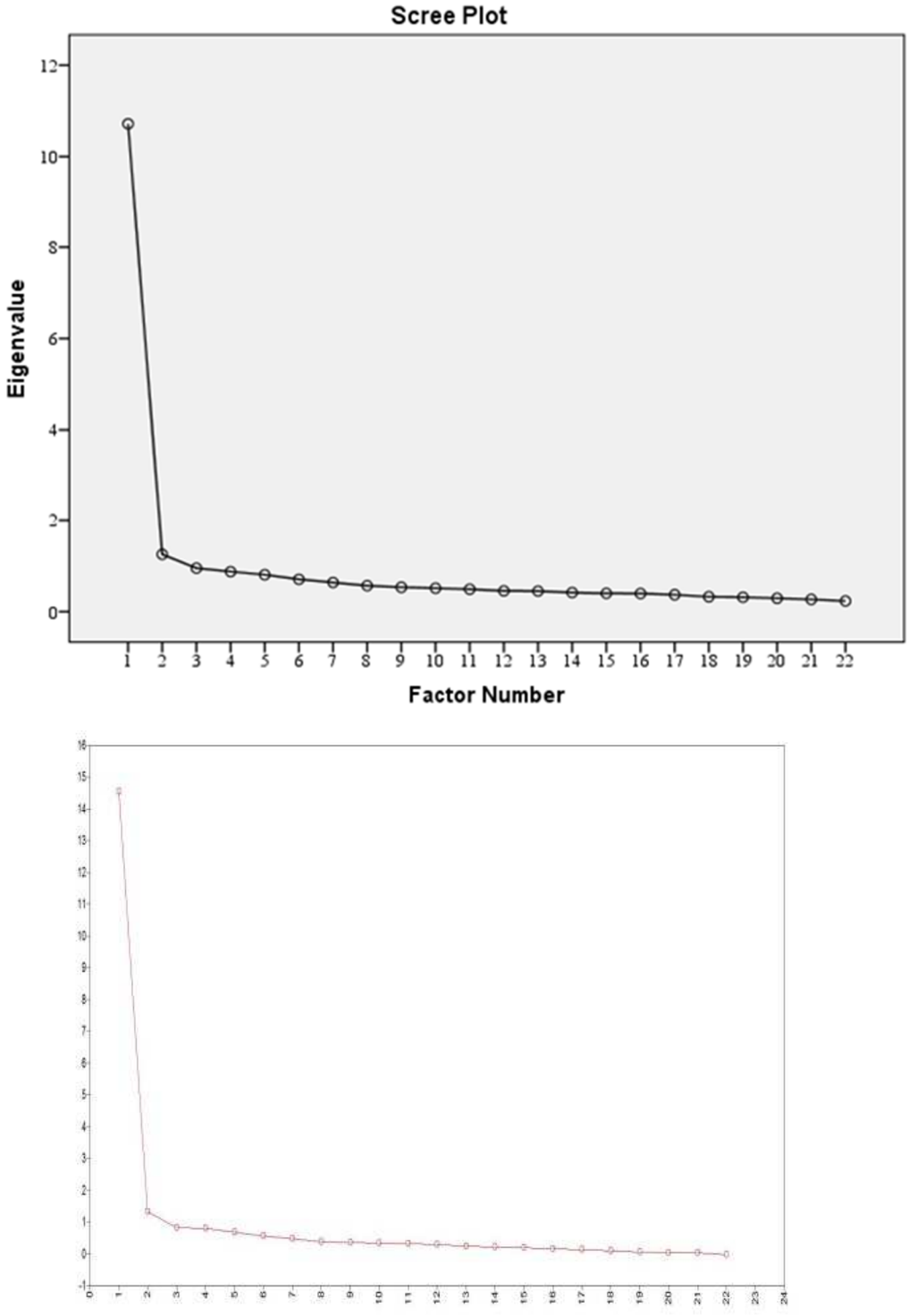


Figure 32. SPSS/Mplus factor scree plots. Top figure = Eigenvalue Scree Plot (SPSS); Bottom figure = Eigenvalue Scree Plot (Mplus).

Scree plot factor selection focuses on the change in slope occurring between successive eigenvalues. The slope represents the marginal (additive) proportion of explained variance contributed by the next factor. The elbow selection rule suggests selecting factors up to the point where the change in slope is essentially zero, "… the last substantial decline in the magnitude of the eigenvalue…" (Brown, 2006, p. 27). From the scree plots provided above, one dominant factor is evident with consideration of a possible second factor.

A criticism of this approach is that it is somewhat arbitrary and open to the interpretation of the observer. In addition, it is entirely mechanical and does not involve inferential considerations of the aligned variable. Brown (2006) notes that "… the Scree test performs reasonably well under conditions such as when the sample size is large and when well-defined factors are present in the data …" (p. 27).

Extraction criteria - percent variance explained (Common variance accounted > 50 %). Another criteria for factors extraction focuses on the total proportion of explained variance. Here, the researcher is concerned with accounting for as much of the explained variance in the data as possible. Pett et al. (2003) note "… the researcher terminates the factor extraction process when a threshold for the maximum variance extracted (e.g., 75-80%) has been achieved…" (p. 116). Here, the researcher sets the maximum desired variance threshold under the premise that a factor model lacking sufficient explanatory power also lacks inferential value (i.e., it fails to account for the nature of the underlying, unobserved, latent constructs). The problem with this approach is that it does not consider inferential aspects of the factors extracted, the factors extracted may not be clearly interpretable. This is especially problematic for factors contributing minimal marginal

variance which are identified simply to reach the a priori maximum variance threshold. Finally, the maximum variance threshold is, in itself, an arbitrary criterion.

Again referring to Table 39 (above), the first factor in each of the extraction approaches account for (approximately) just over 50% of total variance in the data while each successive factor contributes (approximately) less than 6% of incremental variance. Arguably, setting a maximum variance threshold that forces inclusion of factors that do not provide substantive contributions is questionable. Regardless, the data suggest the presence of one-to-two dominant latent factors underling the rating data that explain approximately half of the total/common variance.

A criticism associated with use of the Guttman-Kaiser, scree plot, and variance explained decision rules are that they are not statistically based and require substantial application of judgment on the part of the researcher. However, two alternative procedures address these concerns by incorporating more rigorous statistical properties into the decision analysis. These are the minimum average partial (MAP) test and parallel analysis. For each, O'Connor (2000) notes that

> …these procedures are statistically based, rather than being mechanical rules of thumb. In parallel analysis, the focus is on the number of components [factors] that account for more variance than the components derived from random data. In the MAP test, the focus is on the relative amounts of systematic and unsystematic variance remaining in a correlation matrix after extractions of increasing numbers of components…" (p. 396)

Application of each approach to the FFT rating data is discussed below.

Extraction criteria - Velicer's minimum average partial (MAP) test (maximized proportional common variance). The MAP test attempts to identify the ratio of common to unique (error) variance attributable to each extracted factor. It is an iterative approach that first identifies the total amount of common (shared) variance accounted for by the

first factor. This common variance is then partialed out from the remaining variance available to the second factor. The process is repeated for each successive factor, partialing the incremental common variance left over from the previous factor's allocation. For each extracted factor, the ratio of common to unique (error) variance is computed and reported in the form of both a squared partial correlation coefficient and a 4^{th}-power correlation coefficient. The extraction process stops when the amount of unique (error) variance exceeds that of common variance (O'Connor, 2000, p. 397). O'Connor notes that "…components are no longer retained when there is proportionately more unsystematic variance than systematic variance…" (p. 397).

For the FFT rating data, a Velicer's MAP test procedure was conducted, generating a vector of squared partial correlation coefficients for each factor. Table 40 reports these values.

Table 40

Velicer's Minimum Average Partial (MAP) Test

	Average Partial Correlations	
Factor	Squared	Power 4
0	.2171	.0514
1	.0127	.0004
2	.0118	.0004
3	.0134	.0005
4	.0157	.0007
5	.0190	.0012
6	.0227	.0016
7	.0282	.0028
8	.0345	.0046
9	.0410	.0066
10	.0479	.0090
11	.0554	.0129
12	.0622	.0148
13	.0727	.0188
14	.0913	.0242
15	.1109	.0327
16	.1371	.0468
17	.1730	.0666
18	.2297	.1100
19	.3211	.1894
21	1.0000	1.0000

The information in Table 40 indicates that the smallest average squared partial correlation is .0118 associated with the second factor and the smallest average 4^{th}-power power partial correlation is .0004 associated with the first factor. From these results, Velicer's minimum average partial (MAP) test identifies between 1 and 2 influential factors.

Extraction criteria - parallel analysis (non-random eigenvalues). Parallel analysis attempts to identify latent factors that account for more variance in the data than occurs by random chance. To distinguish this, a large number (e.g., 100) of randomly generated data sets are constructed consistent with the number of variables in the originating sample. For each randomly generated data set, an EFA is completed generating a vector of eigenvalues. Then, the mean and 95th percentile of the eigenvalue vectors are computed and plotted against the eigenvalues of the original data. The inference is that eigenvalues observed in the original data set that exceed those produced by the randomly generated data reflect true latent components existing outside the bounds of random chance.

For this study, 100 randomly generated data sets were estimated using the PAF-parallel analysis process in SPSS (O'Connor, 2000). Table 41 reports the results, comparing eigenvalues derived from the original FFT data set with the mean and 95th percentile randomly generated values.

Table 41

Comparative Eigenvalues: PAF versus Parallel Analysis

Factor	Eigenvalues Raw Data	Eigenvalues Means	Eigenvalues 95th Percentile
1	10.252382	.681712	.780199
2	.783903	.577060	.642289
3	.485432	.503313	.566715
4	.383194	.435171	.489736
5	.288429	.370215	.410388
6	.209346	.319645	.370685
7	.123582	.264198	.313992
8	.092491	.211936	.258094
9	.050307	.164364	.207330
10	.037954	.120295	.162481
11	.027653	.078887	.118560
12	.009833	.030295	.071547
13	-.012358	-.007120	.025541
14	-.049134	-.044227	-.015784
15	-.066971	-.084061	-.049207
16	-.071885	-.122041	-.094110
17	-.128053	-.160550	-.130053
18	-.141838	-.199019	-.171805
19	-.160837	-.237681	-.208848
20	-.167950	-.275541	-.246734
21	-.189847	-.316748	-.278678
22	-.202571	-.367533	-.331901

Specifications for this Run: (Method) PAF/Common Factor Analysis & Random Normal Data Generation; (Number of Cases) 238; (Number of Variables) 22; (Number of Datasets) 100; (Percentile Threshold) 95th

The data identify two factors whose eigenvalues exceed the random values generated through parallel analysis. Thus, parallel analysis suggests two primary factors underlying the measured FFT rating data.

Extraction criteria –chi-square (χ^2) test (P-value not sig). Factor analytic techniques attempt to model (e.g., reproduce, account for, represent…) the variance-covariance structures observed in a set of measured data. If the parameters estimated by the factor analytic model accurately reproduce these structures, then the revealed factors provide an evidence-based representation of the latent (unobserved) constructs inherent in the data. Comparing the observed and reproduced variance-covariance structures

provides opportunity to construct statistical tests of correspondence (model fit) and to assess the appropriateness of the factor extraction decision. The null hypothesis (H_o) for this type of comparison is that the factors extracted exactly reproduce the variance-covariance structure among the measured variable. Rejection of H_o suggests the model does not fit the data well. Thus, in this formulation, it is desirable to not reject H_o, suggesting the model estimates adequately account for the variance structures.

A chi-square (χ^2) goodness of fit test is often employed to test the adequacy of the reproduced factor structure. Two prevalent estimation approaches for constructing the EFA-based χ^2 test are maximum likelihood (ML) and generalized least squares (GLS). These estimation approaches differ in the method used to derive the reproduced (common) variance estimates, referred to as communalities, that are initially substituted along the diagonal of the measured variable's correlation matrix (necessary for the computational procedures directing the factor extraction process). However, the distributional assumptions of the two approaches differ: ML requires data to be multivariate normal while selected GLS methods may be applied to measures based on ordinal scaling (Muthen & Muthen, 2012; Brown, 2006; Pett et al., 2003). Since the FFT rating data is based a non-multivariate normal (ordinal) scaling, the weighted least squares (WLSMV) estimator available in Mplus was used to assess the EFA model fit based on a χ^2 goodness of fit test. For a more complete understanding of the analytic foundations and related computational procedures to derive communalities estimates, the reader is referred to Brown (2006), Pett et al. (2003), and Muthen and Muthen (2012).

The procedure for applying the χ^2 test involves sequentially extracting an ever increasing number of factors, each time evaluating the significance of the χ^2 statistic. The

point at which χ^2 becomes no longer significant identifies the number of extracted factors that adequately reproduce the variance-covariance structure of the measured data. Pett et al. (2003) suggest that "… the researcher starts with a one-factor model and adds additional factors until the significant test indicates that a given factor model does not significantly deviate from the observed data…" (p. 121). This process was conducted in Mplus version 7.11 using the WLSMV estimator. Table 42 reports the resulting χ^2 model fit statistics (results for the first six extractions are presented).

Table 42

Summary of Chi-Square Model Fit Information

Model	Number of Parameters	Chi-Square	Degrees of Freedom	P-Value	Comment
1-factor	22	400.940	209	0.0000	Reject H_o
2-factor	43	249.164	188	0.0019	Reject H_o
3-factor	63	208.172	168	0.0191	Reject H_o
4-factor	82	162.196	149	0.2173	Failure to Reject H_o
5-factor	100	128.593	131	0.5431	Failure to Reject H_o
6-factor	117	102.067	114	0.7809	Failure to Reject H_o

Applying a $p = .05$ significance criterion, the χ^2 test of model fit suggest that a four factor extraction model adequately reconstructs the variance-covariance structure of the measured data. However, the χ^2 statistic is known to be sensitive to large sample sizes (n > 100) possibly leading to identification of extraneous factors that have little inferential

interpretation. Pett et al. (2003) note that some researchers suggest increasing the p-value criteria to protect against type II error rates due to reliance on accepting rather than rejecting the null hypothesis. However, the authors caution that this approach may also result in accepting more, possibly improper, number of factors.

Summary of extraction routines. Table 43 summarizes the number of factors indicated by each of the extraction criterion discussed above.

Table 43

Summary of Extraction Criterion

Test Method	Criteria	Number of Suggest Factors
Guttman-Kaiser Criterion	Eigenvalue >= 1	2 Factors
Variance Explained	>50 % explained	1-2 Factors
Scree Plot	Δ Slope	1-2 Factors
Parallel Analysis	Non-Random Eigenvalues	2 Factor
MAP Test	Prop. Common Variance	1-2 Factors
χ^2 Test	P-value not sig.	4 Factors

Note: (MAP) Velicer's Minimum Average Partial (MAP) Test; (χ^2) Chi-Square

The information presented in Table 43 suggests that one to two factors be extracted for further analysis and interpretation. Because the intent of this research question is to investigate the hypothesized factor structured of the Danielson FFT framework, two EFA extraction models were examined: a four-factor structure consistent with posited framework and a second specifying two-factors based on the extraction analysis provided above. The two models are compared and discussed within the context

of the research question. Mplus version 7.11 was used for estimating both models applying a WLSMV estimator to account for the data's ordinal rating scale followed by oblique (Oblimin) rotation using Kaiser row standardization (Pett et al., 2003; Muthen & Muthen, 2012).

The Oblimin (oblique) factor rotation procedure was applied to each model to enhance the interpretability of the solutions. Pett et al. (2003) note that "… rotation is the process of turning the reference axes of the factors about their origin to achieve a *simple structure* and theoretically more meaningful factor solution…" (p. 132). Oblique rotations do not impose covariance restrictions between the extracted factors (as is the case for orthogonal rotations). Oblique rotations are appropriate where it is reasoned (a priori) that the constructs underlying the data are (to some degree) interrelated (Pett et al., 2003, p. 149). Allowing for this in the modeling framework (hopefully) provides a more realistic depiction of the true nature of the latent structure.

Use of Oblique rotations produces two representations (matrices) of the factor-variable relationships. The first is the factor-pattern matrix. This matrix reports values that are interpreted like partial standardized regression coefficients, the influence of each measured variable has after holding all other variables constant. The larger the value, the greater influence the variable has in determining explained (common) variance. Pett et al. (2003) argue that it is the information contained in the factor pattern matrix that informs on the simple structure after rotation.[15]

---

[15] Pett et al. (2003) note that "…it is the factor pattern matrix in an oblique rotation that is used to determine the extent to which simple structure has been achieved…" (p. 143). The coefficients reported in factor pattern matrices are interpreted like standardized regression weights, indicating the degree of influence a variable has on explaining common variance within a factor after holding the contributions of all other variables constant.

The second matrix is the factor structure matrix. This matrix represents simple variable-factor correlations based on total explained variance. However, because these correlations are based on total (common plus unique) variance, the coefficients become less interpretable as the degree of factor covariance increases – as error variances become increasingly shared across factors it becomes more difficult to isolate a given variable's contribution within a specific factor (Pett et al., 2003, pp. 150-151). Pett et al. (2003) note that "…if the factors are too highly correlated, the factor structure matrix may not share the same simple structure as the pattern matrix…" (p. 163). For this reason emphasis is placed on presenting and interpreting the factor pattern matrices produced by the four-factor and two-factor EFA models.

Pattern matrix loadings for the two-factor EFA model. Table 44 presents the oblique pattern matrix loadings for the two-factor EFA model. All significant ($p < .05$) factor loadings above .600 have been bolded to aid in interpretation.[16] In addition, the variables have been organized by their representation within each of the four Danielson FFT domains.

[16] Cohen (1988) provides a rule of thumb suggesting that effects sizes of .2, .5. and .8 as relatively "weak", "Moderate", and "strong", respectively.

Table 44

EFA Two-Factor OBLIMIN Rotated Pattern Matrix

Variable	Factor 1	Factor 2
D1a Demonstrating Knowledge of Content and Pedagogy	**0.601***	0.317*
D1b Demonstrating Knowledge of Students	0.198	**0.687***
D1c Setting Instructional Outcomes	**0.829***	0.084
D1d Demonstrating Knowledge of Resources	0.174	**0.717***
D1e Designing Coherent Instruction	**0.893***	-0.004
D1f Designing Student Assessments	0.273*	**0.645***
D2a Respect and Rapport	0.490*	0.338*
D2b Culture of Learning	**0.642***	0.266*
D2c Managing Classroom Procedures	**0.849***	-0.026
D2d Managing Student Behavior	**0.717***	0.057
D2e Organizing Physical Space	0.142	**0.618***
D3a Communicating With Students	**0.608***	0.243*
D3b Questioning Discussion Techniques	**0.891***	-0.023
D3c Engaging Students	**1.011***[17]	-0.132
D3d Using Assessment in Instruction	0.429*	0.471*
D3e Flexibility Responsiveness	0.146	**0.666***
D4a Reflecting on Teaching	0.078	**0.740***
D4b Maintaining Accurate Records	-0.016	**0.822***
D4c Communicating with Families	-0.088	**0.864***
D4d Participating in a Professional Community	-0.074	**0.824***
D4e Growing Developing Professionally	0.041	**0.857***
D4f Showing Professionalism	-0.113	**0.970***

* Significant at the 5% level

The size of the individual variable-loadings indicates unique alignments to each of the two factors, that is, variables do not strongly load simultaneously on both factors. Additionally, it is noted the pattern of loadings span multiple FFT domains for each factor. However, it does appear that the first factor is differentiated by components from

[17] Recall that loadings reported by the pattern matrix under oblique rotation are interpreted as standardized regression coefficients and are not bounded between +/- 1.0.

Domain 2 (Classroom Environment) and Domain 3 (Classroom Instruction) while the second factor aligns to all components in Domain 4 (Professional Responsibilities). Variables from Domain 1 (Planning and Preparation) are split even between both factors

Pattern matrix loadings for the four-factor EFA model. Table 45 presents the oblique pattern matrix loadings for the four-factor EFA model. All significant ($p < .05$) factor loadings above .600 have been bolded to aid in interpretation. In addition, the variables have been organized by their representation within the four Danielson FFT domains.

Table 45

EFA Four-Factor OBLIMIN Rotated Pattern Matrix

Variable	Factor 1	Factor 2	Factor 3	Factor 4
D1a Demonstrating Knowledge of Content and Pedagogy	0.**661***	0.135	0.19	0.237
D1b Demonstrating Knowledge of Students	0.231*	0.**600***	0.125	0.072
D1c Setting Instructional Outcomes	**0.796***	0.183	-0.133	0.022
D1d Demonstrating Knowledge of Resources	0.219	0.**664***	-0.004	0.114
D1e Designing Coherent Instruction	**0.866***	0.001	-0.004	0.158
D1f Designing Student Assessments	0.265	0.**767***	-0.217	-0.014
D2a Respect and Rapport	0.494	0.142	0.477*	-0.045
D2b Culture of Learning	0.**618***	0.213*	0.248	-0.071
D2c Managing Classroom Procedures	**0.799***	0.071	0.112	-0.364*
D2d Managing Student Behavior	0.647*	0.154	0.037	-0.147
D2e Organizing Physical Space	0.124	0.**627***	0.112	-0.106
D3a Communicating With Students	0.561	0.213*	0.259	-0.133
D3b Questioning Discussion Techniques	**0.862***	0.041	-0.081	0.056
D3c Engaging Students	**0.979***	-0.117	0.001	0.127
D3d Using Assessment in Instruction	0.42	0.594*	-0.213	-0.037
D3e Flexibility Responsiveness	0.216	0.564*	-0.061	0.344
D4a Reflecting on Teaching	0.165	0.468	0.305	0.278
D4b Maintaining Accurate Records	0.008	0.**781***	0.082	-0.022
D4c Communicating with Families	-0.069	0.**873***	-0.005	-0.051
D4d Participating in a Professional Community	-0.078	0.**877***	-0.045	-0.062
D4e Growing Developing Professionally	0.139	0.**653***	0.118	0.327*
D4f Showing Professionalism	-0.059	0.**764***	0.324	0.118

* Significant at the 5% level

The data indicate numerous strong, significant, loadings within the first two factors and only one-to-two significant, albeit weak, loadings present in factors three and four. This pattern underscores information presented by the model fit statistics indicating that the four-factor model does not adequately reproduce the variance-covariance relationships in the measured data. By forcing extraction of four factors, the loading patterns and magnitudes of the latter two factors limit their interpretability. In addition, the pattern loadings for factors one and two seem consistent with those reported under the two factor EFA model.

Factor correlations for the two- and four-factor EFA models. Table 46 reports the factor correlations for the two- and four-factor EFA models.

Table 46

Oblimin Factor Correlations for the Two- and Four-Factor EFA Models

Two-Factor EFA Model

	Factor 1	Factor 2
1	1.000	
2	0.797*	1.000

Four-Factor EFA Model

	Factor 1	Factor 2	Factor 3	Factor 4
1	1.000			
2	0.756*	1.000		
3	0.285	0.352	1.000	
4	0.135	0.155	0.049	1.000

*Significant at the 5% level

The factor correlations for the two- and four-factor EFA models indicate a strong correlation between factors 1 and 2 and weak associations attribute to factors three and four. The magnitude of the correlation (r = .76) between factors one and two again raises concerns on the inferential discrimination between the constructs. While the individual measured components load independently on their respective factors, the information becomes redundant when the factor correlations are strong.

RQ1B (b). To what degree does the factor analytic structure of empirically-based PP scores differ between less experienced and more experienced teachers? The approach is exploratory and confirmatory factor analysis. The measures are factor extractions, factor loadings, χ^2, AIC, α.

This question explores the stability of the factor structure between subgroups based on years' experience within the district. The posited theory underlying the Danielson FFT Framework is that the four-domain evaluation structure properly informs on instructional competence regardless of any mediating conditions (i.e., distinctions in teacher tenure and/or instructional experience). For the data set used in this study, 141 (59%) of the Group A teachers have been employed by the district for four or more years while 97 (41%) have been employed for three years or less. To assess (in part) the tenability of the framework's stability, factor structures for teachers with more/less instructional experience are explored using EFA. [18] All models were estimated using Mplus version 7.11 applying WLSMV estimators for correction of ordinal scaling and

[18] It is recognized that district employment is an imperfect measure of actual instructional experience. It is possible that some teachers that have been employed by the district for three or less years have many more years of actual teaching experience in other educational agencies. However, this information was not available in the district's database.

oblique (Oblimin) rotation with Kaiser row standardization (Pett et al., 2003; Muthen & Muthen, 2012).

Summary of model fit information – instructional experience. For this research question, two factor extraction criteria were utilized: χ^2 Model fit statistic ($p > .05$, failure to reject H_o) and the Guttman-Kaiser eigenvalue criterion (eigenvalues > = 1). Table 47 summarizes the χ^2 fit information for models sequentially extracting 1-to-4 factors.

Table 47

Summary of Model Fit Information by Instructional Experience

	Probationary (3 Yrs. Or Less)				Continuing (4+ Yrs.)			
Model	Number of Parameters	Chi-Square	Degrees of Freedom	P-Value	Number of Parameters	Chi-Square	Degrees of Freedom	P-Value
1-factor	22	294.658	209	0.0001	22	306.068	209	.0000
2-factor	43	197.143	188	0.3091	43	209.14	188	0.1388
3-factor	63	167.53	168	0.4957	63	182.714	168	0.2071
4-factor	82	143.906	149	0.6025	82	154.993	149	0.3516

As shown, the χ^2 model fit indices identifies a two-factor model as the most parsimonious non-significant ($p > .05$) specification for each teacher experience group. Similarly, the eigenvalue values for the probationary and continuing instructional groups are provided in Table 48.

Table 48

Eigenvalues for Probationary and Continuing Instructional Group EFA[19]

	Probationary (3 Yrs. or Less)			Continuing Teachers (4+ Yrs.)		
Factor	Eigenvalue	Percent Variance	Cumulative Percent	Eigenvalue	Percent Variance	Cumulative Percent
1	15.648	71.13%	71.13%	12.598	57.26%	57.26%
2	1.437	6.53%	77.66%	1.987	9.03%	66.29%
3	0.853	3.88%	81.54%	1.171	5.32%	71.61%
4	0.739	3.36%	84.90%	0.984	4.47%	76.09%
5	0.629	2.86%	87.75%	0.905	4.11%	80.20%
6	0.614	2.79%	90.55%	0.728	3.31%	83.51%
7	0.499	2.27%	92.81%	0.574	2.61%	86.12%
8	0.387	1.76%	94.57%	0.53	2.41%	88.53%
9	0.376	1.71%	96.28%	0.502	2.28%	90.81%
10	0.305	1.39%	97.67%	0.465	2.11%	92.92%
11	0.275	1.25%	98.92%	0.395	1.80%	94.72%
12	0.233	1.06%	99.98%	0.314	1.43%	96.15%
13	0.131	0.60%	100.57%	0.303	1.38%	97.52%
14	0.111	0.50%	101.08%	0.296	1.35%	98.87%
15	0.093	0.42%	101.50%	0.193	0.88%	99.75%
16	0.061	0.28%	101.78%	0.149	0.68%	100.42%
17	0.027	0.12%	101.90%	0.09	0.41%	100.83%
18	-0.007	-0.03%	101.87%	0.073	0.33%	101.16%
19	-0.033	-0.15%	101.72%	0.035	0.16%	101.32%
20	-0.096	-0.44%	101.28%	-0.006	-0.03%	101.30%
21	-0.117	-0.53%	100.75%	-0.109	-0.50%	100.80%
22	-0.163	-0.74%	100.01%	-0.176	-0.80%	100.00%
Sum	22.002	100.01%		22.001	100.00%	

As shown, the Guttman-Kaiser criterion (eigenvalue > 1.0) supports extraction of a two-factor model for Probationary Teachers, accounting for 78% of cumulative (total)

[19] Negative eigenvalues when using WLSMV for ordinal Scales - Muthen and Muthen (2012): "I think this is ignorable. With categorical variables and WLSMV, you work with tetrachoric and polychoric correlations which are computed for pairs of variables at a time and therefore can produce a non-positive definite sample correlation matrix - which has some negative eigenvalues. You can still get a pos-def model-estimated correlation matrix. If the model fits well to this sample correlation matrix, you can view the situation as the non-pos def sample correlation matrix was not "significantly non-pos def." There have been ideas in the literature about deleting the eigenvalues and eigenvectors for the negative eigenvalues and recreating the sample correlation matrix this way, smoothing it, and then fitting the model, but I am not sure that is an important improvement" (Muthen, 2011).

variance explained. For the Continuing group, the first three factors report eigenvalues above 1.0 for a (total) cumulative percent variance explained of 72%. Accordingly, in addition to exploring the hypothesized four-factor model for both groups, a two-factor model is added for the Probationary Group and both a two- and three-factor model is examined for the Continuing Group.

Factor pattern matrix for probationary (three years or less experience) teachers. A two- and four-factor model was specified for the Probationary teacher group. The variable loadings for the factor pattern matrices obtained from Oblimin (oblique) rotations are provided in Table 49. All significant ($p < .05$) factor loadings above .600 have been bolded to aid in interpretation. In addition, the variables have been organized by their representation within the four Danielson FFT domains.

Table 49

Pattern Matrix for a Four- and Two-Factor Model for Probationary Teachers

	Probationary Four-Factor Model				Probationary Two-Factor Model	
Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 1	Factor 2
D1A	**0.778***	0.295	-0.112	-0.164	**0.643**	0.331
D1B	0.257	**0.691***	0.065	-0.031	0.218	**0.755***
D1C	**0.885***	0.067	0.029	0.122	**0.927***	0.059
D1D	**0.648***	0.133	0.227	-0.107	0.508	0.437
D1E	**0.976***	0.048	-0.12	-0.049	**0.923***	0.016
D1F	0.458	0.343	0.268*	0.039	0.432	0.559
D2A	0.017	0.61	0.268	0.007	0.010	**0.793***
D2B	0.501	0.401	0.075	-0.028	0.457	0.488
D2C	0.185	**0.623***	0.084	0.277	0.410	0.460
D2D	0.254	0.436	0.071	**0.576***	**0.653***	0.116
D2E	0.091	0.05	**0.846***	0.043	-0.003	**0.769***
D3A	0.215	0.512	0.159	0.153	0.323	0.510
D3B	**0.982***	-0.081	-0.039	0.137	**1.041***	-0.146
D3C	**0.928***	-0.022	0.042	0.08	**0.937***	0.015
D3D	0.401	0.412	0.245	0.126	0.442	0.545
D3E	**0.774***	-0.123	0.358*	0.015	**0.661***	0.266
D4A	0.374	0.329	0.196	-0.323*	0.074	**0.746***
D4B	0.004	**1.060***	-0.171	-0.061	0.073	**0.825***
D4C	0.125	**0.589***	0.176	0.014	0.119	**0.704***
D4D	0.013	0.632	0.254	0.005	0.003	**0.805***
D4E	0.563	0.346	0.18	-0.375*	0.233	**0.766***
D4F	-0.061	0.56	0.571	-0.191	-0.277	**1.161***

Oblimin Rotated Loadings; (*) significant at 5% level; (Bold) sig, >= .6 rounded

Consistent with previous findings, the loadings associated with factors three and four for the theoretical four-factor extraction model are generally uninterpretable. Each produces only one significant loading with a value greater than .60. In contrast, the two-factor extraction model is more interpretable with well-aligned and significant loadings. Here, the Domain 4 (Professional Responsibilities) construct is clearly delineated while factor one is identified by a sampling of components from the remaining three domains.

Factor pattern matrix for continuing (four or more years of experience) teachers. A two-, three-, and four-factor model was specified for the Continuing teacher group. The variable loadings for the factor pattern matrices obtained from Oblimin (Oblique) rotations are provided in Table 50. All significant ($p < .05$) factor loadings above .600 have been bolded to aid in interpretation. In addition, the variables have been organized by their representation within the four Danielson FFT domains.

Table 50

Pattern Matrix for a Four-, Three-, and Two-Factor Model for Continuing Teachers

	4-Factor Cont. Model				3-Factor Cont. Model			2-Factor Cont. Model	
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2
D1A	0.626	-0.113	0.089	0.437	0.491*	0.417	0.006	**0.626***	0.249*
D1B	0.114	0.492	0.080	0.352	0.136	**0.698***	0.044	0.331*	0.510*
D1C	0.575	0.180	0.221	-0.015	**0.640***	0.092	0.216	**0.663***	0.195
D1D	-0.082	0.760	0.258	0.124	0.094	**0.604***	0.268*	0.166	**0.720***
D1E	0.790	-0.062	0.046	0.175	**0.732***	0.166	-0.006	**0.834***	0.012
D1F	0.243	0.302	0.555	-0.057	0.344	0.173	0.531*	0.255	**0.643***
D2A	0.582	0.332	-0.308	0.401	0.529	**0.647***	-0.356*	**0.753***	0.109
D2B	0.561	0.309	0.069	0.091	**0.612***	0.291	0.061	**0.715***	0.191
D2C	0.988	0.113	-0.011	-0.342	**1.080***	-0.284	0.025	**1.043***	-0.289*
D2D	0.490*	0.364	0.022	0.024	0.559*	0.264	0.02	**0.661***	0.139
D2E	0.129	0.723	0.113	-0.020	0.257	0.478*	0.158	0.334	0.497*
D3A	0.504	0.400	-0.041	0.145	0.550*	0.407	-0.047	**0.706***	0.173
D3B	0.771	-0.080	0.248	0.028	**0.760***	-0.006	0.211	**0.753***	0.116
D3C	0.761	-0.027	0.025	0.127	**0.745***	0.101	-0.01	**0.839***	-0.047
D3D	0.470	0.138	0.454	-0.062	**0.552***	0.023	0.446*	0.468*	0.419*
D3E	-0.082	0.194	0.342	0.528	-0.146	**0.706***	0.26	-0.069	**0.819***
D4A	0.140	0.228	0.158	0.531	0.065	**0.725***	0.087	0.237	**0.604***
D4B	0.108	-0.087	0.764	0.135	0.097	0.137	**0.688***	-0.055	**0.805***
D4C	0.051	0.062	0.691	0.120	0.07	0.202	**0.646***	-0.064	**0.824***
D4D	-0.029	0.188	0.675	-0.008	0.057	0.137	**0.649***	-0.085	**0.764***
D4E	0.040	0.219	0.257	0.484	-0.027	**0.689***	0.18	0.08	**0.709***
D4F	0.177	0.018	0.358	0.516	0.074	**0.597***	0.254	0.157	**0.692***

Oblimin Rotated Loadings; (*) significant at 5% level; (Bold) sig, >= .6 rounded.

The factor loading information for the four-factor model estimated from the Continuing instructional group data reports only a single significant ($p < .05$) value even though some of the loadings are in excess of .50. This contrasts sharply with the findings from the four-factor extraction models specified for the combined (all teachers) and probationary instructional groups. Review of the raw data revealed that none of the 141 teachers in the Continuing instructional group attained a zero (Unsatisfactory)

performance rating for any of the 22 FFT behavioral elements and few attained the second lowest rating of *1 (Basic)*. This concentrates the distribution of ratings for this group between a *2 (Proficient)* and *3 (Distinguished)*. The attenuation from a four-option to a binary-option scale may be increasing the standard errors of the loading estimates. In addition, if the four-factor model substantively misspecified for the Continuing instructional group, this may be why the substantive difference in identifying statistically significant loadings.

The loadings reported for the three- and two-factor Continuing teacher extraction models are better behaved, revealing more interpretable (plausible) solutions. However, the alignment between the components and the hypothetical behavioral domains is inconsistent. That is, the highest loading components do not readily cluster within the pre-established FFT constructs. Only loadings reported for factor two from the two-factor model align to the pre-specified Domain 4 (Professional Responsibilities). This is consistent with previous findings for the two factor model containing data for all the teachers.

Comparative factor correlations. Table 51 reports the factor correlations produced by the Probationary EFA analysis.

Table 51

Factor Correlations for the Probationary Two- and Four-Factor Models

	Probationary Four-Factor Correlations				Probationary Two-Factor Correlations	
	F1	F 2	F3	F4	F1	F2
F1	1.000				1.000	
F2	0.726*	1.000			.744*	1.000
F3	0.526*	0.576*	1.000			
F4	0.138	-0.027	-0.012	1.00		

(*) significant at 5% level. Oblimin Factor Correlations

The data indicate that both the two- and four-factor Probationary models report strong correlations between factors one and two. In addition, moderately strong correlations are reported between factors three and the first two factors. However, previous review of the pattern loadings (above) failed to provide any interpretable understanding of how factor three might be interpreted other than it represents a single overall latent construct, only a single loading exceeding .60 was significant. Finally, none of the correlations associated with factor four are significant ($p < .05$). Table 52 provides similar information for the Continuing EFA model analysis.

Table 52

Factor Correlations for the Continuing Four-, Three-, and Two-Factor Models

	Continuing Four-Factor Correlations				Continuing Three-Factor Correlations			Continuing Two-Factor Correlations	
	1	2	3	4	1	2	3	1	2
1	1.000				1.000			1.000	
2	0.541	1.000			0.599*	1.000		0.646*	1.000
3	0.419	0.488	1.000		0.365*	0.545*	1.000		
4	0.381	0.385	0.421	1.000					

(*) significant at 5% level. Oblimin Factor Correlations

The data in Table 52 indicate no significant ($p < .05$) correlations between any of the factors extracted under the Continuing four-factor model. In contrast, the three-factor specification indicates moderately strong correlations (.37 to .60) and the two-factor model reports a value of .65 existing between factors one and two. Again, this suggests that the four-factor specification may not be appropriate.

In summary, the 4-factor model does not appear to represent an appropriate structure for the empirical data. A plausible 3-factor model may be present for Continuing (experienced) teachers, but not for probationary teachers. Finally, the data reveal a plausible 2-factor structure, but with differing interpretative construct components.

RQ1B (c). *To what degree do the 22 elements contained within the theoretical FFT framework adequately represent the latent TIQ construct?* The approach utilizes the Lawshe Content Validity Index (CVI) questionnaire and stakeholder interviews. The measures are CVI, coded interview responses.

This research question explores the representational adequacy of the Danielson FFT Framework on the teacher instructional quality (TIQ) construct. Two methodological approaches were used to assess this context. First, a Content Validity Ratio (CVR) questionnaire was adapted for use in this study as a measure of FFT content adequacy (Lawshe, 1975; Wilson et al., 2012). A copy of the CVR questionnaire is provided in Appendix C. Second, as part of the stakeholder interview protocol, participants were asked to reflect on the content adequacy and representation of the Danielson framework as a measure of instructional quality. The analysis is presented below for each of these two data collections, beginning with the *Lawshe LCVR* followed by *Stakeholder Reflections*.

Construct Validity Index(CVR). The CVR was administered (online) to a purposeful sample of individuals (subject matter experts) based on their in-depth familiarity with the district's evaluation system. Lawshe (1975) used the term subject matter experts (SMEs) to characterize a person's capability for assessing the suitability of test items as empirical measures of the identified latent construct. For this study, members of the Teacher Evaluation Committee ($N = 12$) and teachers serving in the role of instructional growth coaches (IGTs; $N = 24$) served as representative SMEs (total $N = 33$). Each member possessed operational familiarity of the 22 FFT components, the FFT scoring rubrics, the evaluation/observation process, and previous classroom instructional experience.

Invitations to complete the CVR were provided by this researcher during regularly scheduled meetings of the two SME groups. The purpose of the survey (to provide feedback to district policy makers on the evaluation system) and its use in

academic (dissertation) research was discussed along with caveats of data confidentiality, voluntary participation, and desired timeline. A follow up email was sent to all members of the two groups containing a description of the activity, invitation to participate, and links to the online (Survey Monkey) questionnaire.

Prior to its use, drafts of the CVI were distributed to selected central office administrators for review and comment. This included two members of the Teacher Evaluation Committee group, two members of the district's research office staff, and a central office administrator not associated with the Teacher Evaluation Program activity. Feedback received from these individuals resulted in modification of the instrument prior to its implementation.

The CVR questionnaire contained two primary sections. Respondents were first asked to provide a brief description of what they believed defined a *good/effective* teacher:

> Please use the space below to briefly outline/describe what YOU believe it means to be a "Good/Effective" teacher. Feel free to insert sentences/paragraphs or simply enter key phrases, concepts, or words that reflect your thinking.

Following this, respondents were asked to rate the contribution each FFT element made to identifying a *good/effective* teacher:

> *On the table below, please indicate whether obtaining a rating on the specific element is Not Necessary, Useful, or Essential for identifying a good/effective teacher*

Finally, the CVR collected two compositional items: current employment position (classroom teacher, district/school administrator, and instructional growth coach) and years in current position.

Table 53 reports descriptive information for individuals completing the CVI.

Table 53

Descriptive Sampling Information for the CVR

Descriptive Measure	Amount
Number of SMEs Invited to Complete the CVI	33
Number of Completed Questionnaires	23
Completion Response Rate	70%
Number providing definition of a "Good/Effective" Teacher	20
Response Rate (Total Population)	61%
Response Rate (Returned Questionnaires)	83%
Number providing rating of FFT Element Importance	21
Response Rate (Total Population)	64%
Response Rate (Returned Questionnaires)	88%
Responses by Position	
Classroom Teachers (Evaluation Committee)	5 (100% of SMEs)
Instructional Growth Coaches:	14 (58% of SMEs)
District/School Administrators (Evaluation Committee)	4 (100% of SMEs)
Responses by Years Employed in District	
1-Year:	6
2-Years:	4
3-Years:	3
4-Years:	2
5 or More Years:	8
Total:	23

Lawshe (1975) describes content validity as "… the extent to which communality or overlap exists between (a) performance on the test under investigation and (b) ability to function in the defined job performance domain…" (p. 566). His particular context focused on job tasks and the need for employers to construct valid measures of task

performance. In this setting, content validation requires an empirical assessment of how well a measurement instrument used to assess job performance adequately represents the performance domain. Wilson et al.'s (2012) subsequent review of Lawshe's use of the CVR expressed this as

> Content validation rests on demonstration that the test's items are a representative sample of all items within the content domain of interest. Whether the researcher is evaluating the items on a test, questions in an interview, or elements of a set of accreditation standards, the items, questions, themes, or elements should all reflect the intended content of the evaluation tool…" (p. 197)

The CVR operationalizes this activity in the form of SME ratings on each FFT element using a three-point rating scale (*Not Necessary, Useful, or Essential*). Lawshe's content validity methodology "… [makes] a linear transformation of the ratio of the number of SMEs judging an item [element] to be "essential" to the total number of SMEs in the panel…" (Wilson et al., 2012, p. 198). The CVR ratio is expressed as:

$$CVR_i = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}$$

where n_e is the number SMEs indicating that element (i) is "essential" and N is the total number of SMEs providing ratings on element (i). Values of the CVR range between +/- 1.0. If all SMEs rate an element as "essential", the CVR = 1.0. When more than one-half of the SMEs rate an element "essential", the CVR will be greater than 0.0. When exactly one-half score the element "essential", the CVR = 0.0. Finally, if less than half rate the element "essential" the CVR will be negative. Using the Lawshe methodology, an item is seen as substantively contributing to the latent construct (with high content validity) when more than one-half of the SMEs rate the element as "essential" (i.e., when CVR > 0).

Most importantly, Lawshe (1975) provided a table of critical values from which to test the hypothesis of whether the observed proportion of SME "essential" ratings exceeds chance expectation (i.e., a CVR > 0.00). Subsequently, Wilson et al. (2012) recomputed Lawshe's table of critical values in an effort to correct for ambiguities/errors present in the originally published paper. Their analysis clarified the proper format for evaluating the one-tailed hypothesis. Following Wilson et al. (2012), the CVR null hypothesis (H_o) takes the form of a one-tail test: $H_o : n_e \leq \frac{N}{2}$ versus $H_a > \frac{N}{2}$. The standardized (z-normal) critical value of the test statistic becomes $\frac{n_{e.\alpha} - \frac{N}{2}}{\frac{\sqrt{N}}{2}} = Z_\alpha$, which is more simply expressed as $CVR_\alpha = \frac{n_{e.\alpha} - \frac{N}{2}}{\frac{N}{2}} = \frac{Z_\alpha}{\sqrt{N}}$. From the Wilson et al. (2012) reconstructed table of critical values, a CVR value ($N = 21$, $p = .05$, one tail test) = .359 is required to reject H_o that less than one-half of the SMEs rated a particular FFT element as "essential". Table 54 reports the CVR results for the SMEs on the Danielson FFT elements. The elements have been organized by their representation within the four behavioral domains.

Table 54

CVR Values for SMEs Rating of Danielson FFT Elements

Component Description	Count "Essential" Rating (n = 21)	Percent Very Important	Lawshe CVR	Reject Ho (N = 21, p = .05, one-tail) = .359
1a. Demonstrating Knowledge of Content and Pedagogy	15	71%	0.429	✓
1b. Demonstrating Knowledge of Students	21	100%	1.000	✓
1c. Setting Instructional Outcomes	17	81%	0.619	✓
1d. Demonstrating Knowledge of Resources	4	19%	-0.619	
1e. Designing Coherent Instruction	20	95%	0.905	✓
1f. Designing Student Assessments	16	76%	0.524	✓
2a. Creating an Environment of Respect and Rapport	20	95%	0.905	✓
2b. Establishing a Culture for Learning	21	100%	1.000	✓
2c. Managing Classroom Procedures	18	86%	0.714	✓
2d. Managing Student Behavior	18	86%	0.714	✓
2e. Organizing Physical Space	6	29%	-0.429	
3a. Communicating With Students	20	95%	0.905	✓
3b. Using Questioning and Discussion Techniques	19	90%	0.810	✓
3c. Engaging Students in Learning	21	100%	1.000	✓
3d. Using Assessment in Instruction	18	86%	0.714	✓
3e. Demonstrating Flexibility and Responsiveness	16	76%	0.524	✓
4a. Reflecting on Teaching	16	76%	0.524	✓
4b. Maintaining Accurate Records	8	38%	-0.238	
4c. Communicating with Families	10	48%	-0.048	
4d. Participating in a Professional Community	8	38%	-0.238	
4e. Growing and Developing Professionally	15	71%	0.429	✓
4f. Showing Professionalism	14	67%	0.333	

As reported, ratings by SMEs identify 16 out 22 (73%) elements as meeting the Lawshe criteria for construct validation. This suggests that six (27%) of the elements may not be viewed as contributing substantive evidence for identifying or distinguishing attributes of instructional competence. Interestingly, four of the six components within Domain 4 were found to be non-essential. It is noted that Domain 4 was identified as a defining construct in the factor analytic analysis (i.e., one of two dominant factors).

Lawshe (1975) argues that the CVR is "… an item statistic that is useful in the rejection or retention of specific items…" (p. 568). However, he extends the CVR by computing a more global construct validity index (CVI) on the entire test. The CVI is computed as the mean CVR across all items. Higher CVI values indicate greater degrees of content validation as expressed by the SMEs. Computation of the CVI is appropriate for an instrument measuring a single cohesive construct (such as instructional quality) or on component sub-scales that attempt to measure independent attributes within a larger construct. Table 55 reports the CVI for the FFT instrument as well as for each FFT behavioral sub-domain.

Table 55

CVI by Category

Category	CVI (All)	Number of Items	CVI (Retained)	Number of Items
Instrument	0.476	22	0.709	17
Domain 1: Planning & Prep.	0.476	6	0.695	5
Domain 2: Classroom Environment	0.581	5	0.833	4
Domain 3: Instruction	0.790	5	0.790	5
Domain 4: Professional Responsibilities	0.127	6	0.429	3

The CVI for the complete FFT evaluation instrument is .476. This includes items that were identified by SMEs as being non-essential (below the critical value) within the instructional quality construct. The instrument CVI after removing poorly aligned items increases to .709 (a 49% increase). Sub-scale CVI values (all items) range from .127 on Domain 4 (Professional Responsibilities) to .790 on Domain 3 (Classroom Instruction). If poorly-fitting items are removed from the sub-scales, domain CVIs increase substantively. However, Domain 4 remains substantively least influential.

Lawshe (1975) and Wilson et al. (2012) do not provide guidance on critical values for the CVI other than higher values indicate stronger alignment with the latent construct. For element CVR values, statistical significance identifies items for retention. On this basis, if the only items retained had a minimum CVR value of .359, then the instrument average CVI would be .359. Perhaps this provides some minimally acceptable CVI threshold from which to make assertions on the suitability of the instrument and its

subscales. Regardless, the combination of CVR and CVI calculations suggest that up to five items are not essential and Domain 4 reports the weakest alignment to the instructional quality construct as assessed by the SMEs.

Stakeholder reflections. As part of the content investigation, interview participants in each of four stakeholder groups (seven teachers, eight principals, four district policy, and three state policy: n = 22) were asked to reflect on the adequacy of the Danielson Framework for Teaching (FFT). The FFT-related discussions were embedded within a broader interview protocol regarding the overall adequacy of the evaluation system. The prompts related specifically to the Danielson framework were structured as follows:

- Do the Danielson FFT elements cover all aspects of what it means to be a "good teacher"?
- Are characteristics of "good teaching" missing in the FFT Framework?

Expectedly, members of the teacher, principal, and district interview groups were knowledgeable of the Danielson FFT since it served as the organization's adopted evaluation system for certified staff. Prior to conducting the interviews, these stakeholders had received substantive amounts of professional development, trainings, and information. Expectedly, members of the state-policy group had less knowledge of the district's evaluation process (i.e., the details, content, or structure). However, all state members were aware that districts throughout state were adopting nationally recognized standard-based evaluation programs (i.e., Danielson, Marzano, Strong, etc.). Thus, for the state members, interview prompt/probes were aligned more generically to the use of such standards-based systems.

Construct overview. Two dominant perspectives were codified from the stakeholder narratives: *Missing/Incomplete Attributes* characteristic of Teacher Instructional Quality (TIQ), and *Adequacy of Existing FFT Components*. The former highlights concerns over omission of important attributes/impacts credited to good/effective teaching while the latter reflects favorably on the components currently reified by the Danielson framework. Between the two, it is argued that stakeholders find the measured FFT components as necessary, but insufficient, for attaining an accurate representation of good/effective teaching.

A third, less evident, concept in the collective narratives concerns the idea of *Instructional Complexity*. That is, stakeholders view instructional practice as a complex, multi-faceted, and dynamic endeavor that is difficult to measure in quantifiable terms. Here, attempts to reduce, simplify, and/or categorize complex instructional behaviors/impacts are seen as inherently problematic leading to an incomplete, inaccurate, accounting of professional quality. In addition, the dynamic context of instruction makes categorization/standardization inappropriate for the purpose of evaluating all teachers.

Conceptual descriptions of the generalized *Missing/Incomplete Attributes* concept are provided below:

> Missing/Incomplete Attributes: Stakeholders express a general concern that, in its current form, the Danielson FFT fails to account for some important attributes of good/effective teaching. In this way, the FFT is seen as incomplete. The omitted attributes are predominantly affective in nature, relating to non-academic impacts on students and competency-related instructional practices. For the latter, affective dimensions of professional practice depict a teacher's connection/commitment to the profession and to each student's personal/emotional well-being. In addition, selected actions/activities related to day-to-day instructional duties are also believed to be underrepresented.

There is a sentiment that there is more to teaching than what is being measured. In this way, teaching is seen as a multi-faceted, complex, and dynamic process. In its current form, the system seems weighted toward a reduced set of easily measureable actions and outcomes while the more affective aspects of teaching are devalued and/or excluded. The unmeasured are the intangibles. Omission is partly due to the emphasis on the easily quantifiable.

Codes and identities delineating *Missing/Incomplete Attributes* are presented in Figure 33 for three sub-categorizations of the concept - *Missing Affective (Student), Missing Affective (Teacher), and Missing Professional Practice Indicators (Teacher)*.

Missing Affective (Student):

- Motivation, creativity, inspiration, engagement, motivation, perseverance, persistence, dedication, problem solving, change in attitude, enjoyment, connection with school, attendance; child's 'good experience' while at school; child experiences, emotion, beliefs; sense of community; technology, 21st century skills; student affective growth measures; [student] desire/willingness to learn.

Missing Affective (Teacher):

- More to teaching than what is measured; System weighted toward actions and outcomes; affective aspects of teaching have been devalued in TEval framework because of transition to quantified measures & narrowing of measured content; intangibles;
- Interaction/connection with students, parents, colleagues, (it is not just about actions);
- Sense of tone, feeling of the classroom; relationship between student and teacher; feeling of the classroom and learning environment ("next class is totally different");
- Cannot score teachers' commitment, devotion, attitude, emotional passion for students and teaching; character; compassion, caring; clear connection with teaching; "extra mile" effort;
- Teaching difficult to quantify, complex, dynamic;

Missing Professional Practice Indicators (Teacher):

- Points for continuing education & training (national board, master's degree, etc.); teachers' content knowledge (FFT weighted too much to activities); how well teachers knows content;
- Mentoring/counseling activities; interaction with students; what teachers do with students or professionally;
- Extra responsibilities; collegial member of faculty, collaborative; professional responsibilities/activities at school (rather than district) are not counted (i.e., RTI, meetings, scheduling, drama club, before after school tutoring); if activities not at district level it is hard to get distinguished rating; no technology use indicators;

Figure 33. Codes and identities delineating *Missing/Incomplete Attributes*.

Conceptual descriptions of the *Adequacy of Existing FFT Components* concept are provided below:

> Adequacy of Existing FFT Components: While highlighting numerous areas where the FFT may be incomplete, stakeholders regard the attributes already reified in the system as aligned with good/effective teaching. The FFT components are seen as emblematic of classroom activities. In addition, the scoring rubrics bring clarity and structure to the process of evaluating which helps ensure fairness and consistency. Stakeholders note that to conduct the FFT portion of the evaluation, evaluators must (physically) be in classrooms (observing, experiencing). This brings authenticity to the process. In addition, evaluators are required to be trained and certified in the evaluation process. This brings consistency and reliability. FFT scores are seen as evidence-based measures aligned to established best practice standards. For these reasons, FFT ratings are generally perceived as more representative of instructional competency than are test scores.

Codes and identities related to *Adequacy of Existing FFT Components* include the following as shown in Figure 34:

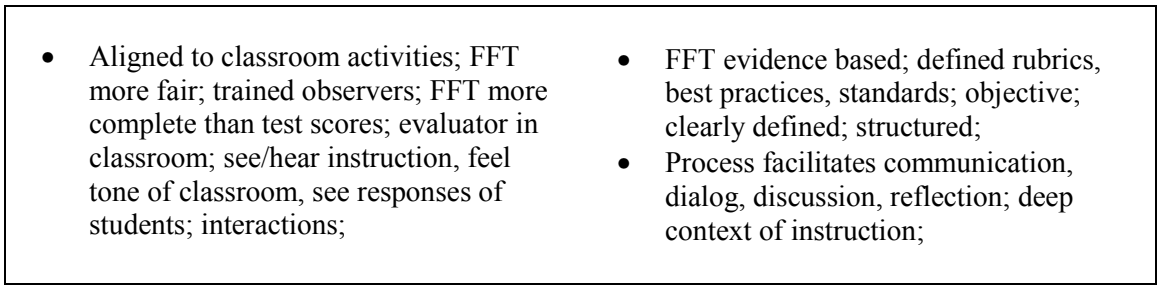


Figure 34. Codes and identities related to *Adequacy of Existing FFT Components*.

Conceptual descriptions of the supporting *Instructional Complexity* concept are provided below:

> Instructional Complexity: The stakeholder narrative underlying Instructional Complexity posits that teaching is complex, dynamic, and difficult to measure. Here, no single instrument or method is viewed as being sufficiently detailed to capture the many-faceted nature of classroom teaching, whether with regard to student-teacher relationships or the dynamic pedagogical intricacies of

instruction. Dynamic complexity leads to systems that omit important attributes that are difficult to categorize, standardized, and measure – related to affective impacts of instruction. As above, these attributes are predominantly affective in character.

Codes and identities delineating *Instructional Complexity* include the following as shown in Figure 35:

Changing, dynamic, situational; Attributes of quality are situational, relative to instructional context (students, subject, grades, schools, etc.); not subject to standardization;	Reduction, simplification, categorization;

Figure 35. Codes and identities delineating *Instructional Complexity.*

Incomplete/missing attributes – impact on students. As presented below, stakeholder narratives highlighted concerns over the (Danielson) framework's inability to fully account for important affective impacts afforded students by a good/effective teacher. These attributes are depicted as non-academic, concerning aspects of the personal, emotional, and self-perceptual. In this narrative, good/effective teachers actively and purposefully affect student engagement, motivation, desire and willingness to learn, as well as traits such as dedication, perseverance, and a student's overall enjoyment of the learning process. Here, good/effective teachers are characterized by their ability to cultivate student's sense of personal value, self-worth, and a belief/hope for the future as a primary objective of their instructional practice.

Participant reflections consistently identified these dimensions as an important attribute of good/effective teaching. To the extent that the Danielson framework fails to

fully incorporate these attributes, evaluation scores are perceived as being incomplete. In this regard, a teacher (below) comments:

> Interviewer: Okay. Do you think that the teacher evaluation process that we have now effectively or accurately distinguishes between the better teacher and less competent teachers? How well does our system do this now?
>
> Teacher 106: …I don't know. It's hard to say, because I think what makes a good teacher encompasses a lot more than what's on that [Danielson] rubric…

For this individual, the first reaction is to compare an idealized conceptualization of good/effective teaching with the components reified under the Danielson framework. The conclusion is that there is "…a lot more than what's on that [Danielson] rubric…" By starting the dialog with "…I don't know. It's hard to say…", suggests that instruction is complex and difficult to fully itemize. The same teacher further shares:

> We talk all the time here that, as a staff, that one of the most important things with us, with the kids, and the families, is that we feel like you have to build that relationship first, and then the trust comes, and [then] the by-in from them comes, and there's more output from them once that's established … I think that there could be teachers who do a really good job of that, and that may or not be factored in… (Teacher 106)

For this individual, good/effective teachers must first "… build that relationship [with students]…" because relationships are the foundation for building trust. In turn, trust enhances a student's "buy-in" (i.e., connection) to the learning process which then leads to greater learning (i.e., "…there's more output from them once that's established…") In the dialog, relationship is one of the "…most important things…" for a good/effective teacher to foster and is perceived as remaining unmeasured by the Danielson components. Regarding these types of affective impacts on students, the same teacher expands his/her description:

> … It would be nice if there were an added component. My kids are doing functional test writing, and they wrote demonstrative speeches. I have a little boy, who probably won't ever perform very highly on benchmarks or AIMS, and he was giving his speech today about he's a Boy Scout, and his speech was on how to start a survival fire. He did a video, and the way that he was talking and answering questions and explaining it, I got emotional because I don't get to see that side of him very often because traditional standards-based education, when it's pen and paper or pencil, is difficult for him. I wish that there was a way to incorporate experiences like that - that our kiddos have [those] big successes for some that don't always have those same successes… (Teacher 106)

This reflection is about recognizing the personal and emotional impacts teachers have on students (i.e., "… that side of him …"). The concern is that these attributes are not accounted for in the evaluation process (i.e., by "…traditional standards-based education… pen and paper…") and that "…It would be nice if there were an added component… a way to incorporate experiences like that…" The implication is that by excluding these types of affects, the evaluation results fail to capture a complete picture of instructional quality.

Discussing missing components in the Danielson framework, another teacher states "… mean, there's always the attendance. There's always the motivational factor…" (101) and another teacher reflects "… We are human beings dealing with small human beings. I think that is, I really think [the] trick would be to get that personality part into an evaluation…" (104).

While the majority of principals did not emphasize concern over missing student affective indicators, one principal does share the teacher's perspective:

> Interviewer: Would you be saying that there is something missing in the current structure that you wish, or would like, to be able to put into that evaluation context - that you're feeling that's not in there sufficiently?

Principal 206: Yes. … There is a teacher … where I know kids don't wanna be in her class anymore. Students don't wanna come to school. I had so many parent complaints about the environment of that classroom. That piece is not in there [in the evaluation system]…"

This principal is connecting classroom environment to student motivation, engagement, and enjoyment in the learning process and the difficulty in representing that connection in the evaluation system.

In a similar vein, a district-level participant reflected on teachers' impact on student motivation:

> … I think Danielson's model is probably one of the better ones that I've seen, but it's certainly not perfect. For example, it doesn't capture a teacher's ability to truly motivate their kids or to make learning come alive for their kids. While it's one of the better models, it still narrows the scope of what people are looking for… (District 303)

This respondent acknowledges the value of the Danielson framework as indicative of quality instruction, but it nevertheless omits important affective impacts on students such as motivation. Because of this, it is argued that the system "…narrows the scope…" of what is measured and negatively affects the interpretation of the evaluation results.

Another district-level colleague stated similarly:

> … We do not have a teacher evaluation system that addresses the other part of the purpose of education - that talks about a child's experience, on a daily basis, coming to school. The teacher evaluation system may say certain cultural or environmental parts of the classroom experience, but it really doesn't get into evaluating whether or not a student has a good experience that day… (District 302)

This person explicitly connects the evaluation process to a larger "… purpose of education …" inclusive of the "child's experience," extending the role of the teachers beyond academics and into the affective. Again, the perspective suggests that the

evaluation process does not adequately account for student affective (emotional, personal) attributes. The implication is that this exclusion detracts from evaluation outcomes by not providing a complete picture of instructional quality (i.e., "…it really doesn't get into evaluating…").

A third district participant noted that some 21st century (higher cognitive) standards are not adequately addressed by the FFT stating:

> … How are we measuring that [student] creativity? How are we measuring that [sense of] community? Right now that really isn't a piece in Danielson. It's not even tools that we have in our toolkit, but we're beginning to develop those. Our Ed Leadership 21 partnership is helping us work with other districts across the United States to look at that. What do those other things look like in the classroom and for a student? As we develop those tools, that should inform, and help us develop, the teacher evaluation process too… (District 304)

While this is not strictly affected in terms of the emotional or personal, it suggests a limitation of the evaluation metrics to address non-traditional (21st century) academic outcomes (i.e., "… How are we measuring that creativity … community? Right now that really isn't a piece in Danielson…"

One of the state-level participants describes a good/effective teacher as being transformative – the ability to transform students from a passive to a self-motivated and self-directed learner. An effective teacher, it is argued, "… just triggers things…" in a student. At the same time, the participant expresses skepticism that the (state) evaluation system is able to capture this impact by stating "…You don't pick that up in the raw data sometimes…":

> … I'd like to think that a teacher who has all the attributes of a good teacher, a teacher that's engaging, that's effective, that might lead that student to think, "I know the content area that they're teaching me, but I go home and I want to learn more," or, "I want to read about it," or, "I want to study about it. I want to go to the library" - that teacher just triggers things. You don't pick that up in the raw

> data sometimes. You just don't. I think all of us have experience with those kinds of teachers that were transformative for those students. I'd like to think that somehow we could capture that… (State 403)

The stakeholder narratives suggest the Danielson FFT omits (or undervalues) important attributes of good/effective teaching. These attributes include non-academic, affective, higher cognitive, influences teachers have on students. By implication, stakeholders believe the addition of these attributes would improve the assessment of Teacher Instructional Quality (TIQ).

Incomplete/missing attributes – instructional practice. While most stakeholder groups highlighted missing components related to student affective domains, principal and district participants tended to reference additional missing instructional, pedagogical, attributes. These attributes include measures of teacher's content knowledge, use of technology in the classroom, omitted/undervalued aspects of professionalism, professional commitment, and indicators of the classroom environment (tone, feeling, atmosphere, etc.). With regard to content knowledge, a district participant (District 301) plainly states, "… I also think that it [Danielson FFT] doesn't capture a lot about the content knowledge that teachers have…" and explains:

> … The tangible thing that is somewhat missing is the skill and knowledge you would need for different content [areas]. I think it [Danielson] addresses more specifically the standard—maybe reading, writing, math—but [not for] our RCTE teachers, our PE teachers, the music teachers, those kinds of things. There is a hole or a gap in Danielson's [FFT], so that it forces us not to focus on [those] specific pieces… (District 301)

The latter comment implicates an inability to differentiate attributes of TIQ across dissimilar instructional contexts (student populations, grade level, subjects, etc.). Implicit is a belief that attributions of a good/effective teaching differ across instructional settings,

rendering static, standards-based, definitions less contextually sensitive (i.e., less inferential validity placed on evaluation scores or measures).

Additional reflections on unmeasured content knowledge include use of technology in the classroom. A district participant comments:

> … I think [the] Danielson [framework] is wide enough that you can add things in. I mean, we talked about that when we were talking about technology. It may not be spelled out there, but it is a tool. How are people using it? … As we know, Danielson has revised her own evaluation [system] several times. I think that's what we need to keep doing, because we don't have all the tools in place yet…(District 304)

Interestingly, this reflection positions Danielson's framework as flexible and dynamic, capable of being modified to fit expanded conceptualizations of TIQ. This is both a positive affirmation and a statement of limitation in its (Danielson's) current implementation.

One principal expressed concern that selected measures of professional practice and engagement/leadership in school activities are not fully accounted for in the FFT rubric. He/she comments:

> … I think what we need to do is probably, and I'm thinking with domain one and four … but really examine … if there's anything else that could or should be included. Maybe for us to accurately label … our teachers, these types of extra responsibilities [should be examined] … in light of the extras that I know many of our teachers are doing and some teachers are not doing… (Principal 204)

The reader is reminded that Domains 1 and 4 in the Danielson framework relate to instructional planning and professional responsibilities, respectively. The participant is concerned that when teachers take on extra responsibilities/activities in these areas they do not get full credit under the existing scoring rubric. The same principal provides an example:

> … I've got a teacher who is in charge of our RTI process. She schedules meetings. She is our drama club sponsor. She does before and after school tutoring. She does nothing at district level, so it's very difficult for her to get to that distinguished, because in some of those components, whatever your extra responsibility, you have to be at the district [involvement level], or you publish something, or do something like that. This person, she's there 8:00 at night, she's just dedicated to the kids at school. She just wants to focus on school. She's not interested in district. She just lives and breathes for school…(Principal, 204)

The narrative suggests that if activities are not taking place at the district level, they are not being fully reflected in the evaluation rubric. For this individual, the comment also raises the issue of professional commitment and dedication as an important attribute of good/effective teachers. The exemplar above (Principal 204) makes special note of this dedication (in terms of hours worked) and the emotional connection a teacher has to the school. This principal continues the dialog:

> … You can't rate passion. The expectation is you open your door at 8:00 [a.m.] and you welcome them [students] in. We've got the teachers that do, "Hi, good morning," and then you've got the teachers who are truly passionate. The love of their job flows from their body. Everything they do, everything they say, and it's just so apparent and obvious. Then the teacher right next door, doing the same thing, I mean, they're physically doing the same thing, but emotionally they're very different… (Principal 204)

Similarly, a different principal reflects on this connection to the profession:

> … that whole aspect of showing that they [teachers] care, and I guess it would go under the respect and responsibilities piece … [but] it doesn't specifically say this is where this has to go. It is very hard to include that teacher's compassion [for students], or the extra, extra stuff that you know that they go into… (Principal 201)

The two stakeholders (above) seem to make an implicit connection between professional commitment/passion and the classroom environment (i.e., supportive learning environment). In this regard, a different principal explains:

> … What I don't think can be recorded … is the feeling of a classroom, positive or negative, because I will have a teacher who, according to the rubric, is doing everything, but when I go in their classroom, it just, I ask the teacher, "Are you

> happy this year, … You just don't seem as positive." When I go in her class, I can feel like, it feels like a dark gloom. If I'm feeling it, obviously, the kids she sees feel it as well…(Principal 206)

The comment identifies a link between students and the learning environment and the difficulty of recording (quantifying, rating) this connection within the FFT evaluation process. The implication is that a teacher's attitude may directly affect student learning by negatively or positively impacting the classroom's environment. The phrase "…is doing everything, but when I go in their classroom…" suggests that the measured FFT attributes are incomplete with respect to this student-environment connection, where a teacher may be rated highly in the context of a poor learning environment. The same principal goes on to explain:

> … Do they [students] feel like I'm [the teacher] excited to come here every day and want to be here and want to do anything for this teacher? Then right across the hall, I have a teacher who teaches at a very high level. Honestly, her kids would bend over backwards to do anything for her. I have another teacher who her kids start off really slow, but they love coming to school every day. They kiss their brains every day. They have a real sense of community. I can try to emulate that as much in the creating a positive environment, but some things can't be recorded, so it's hard…(Principal 207)

These "intangible" affective aspects of professionalism are also noted by both district- and state-level participants. A district policy maker comments:

> … I think it (Danielson FFT] covers some of them [components of effective teaching] … I don't know that there is an instrument out there that covers all of those things. Some of them are intangible. They're harder to define and determine. When you go into a classroom there's a sense. There's a tone. There's a feeling...(District 301)

And a state-policy participant reflects:

> … Do I think the [state] framework includes all of the necessary components? Again, this is me. If I were still a superintendent, the 50 percent [portion] would not be all devoted to classroom observations. The 50 percent - I would call it practices, teaching practices. Teaching practices include instruction, mentoring, counseling, being a member of the faculty, collaborate. Your interaction with

> students would be part of your planning process. You would take into consideration the needs of youngsters as you develop your lesson. You would be aware of their motivation to learn, and willingness to learn, and what they need to be ready to receive your instruction. I could see some of those things, more affective relationship between a student and a teacher, as part of the planning…(State 402)

Similar to the in-district narrative, this state participant seems to emphasize the affected dimensions of professional practice: mentoring, counseling, relationships (interactions) with students, faculty collaboration, and the "…needs of youngsters…" including fostering "… a motivation to learn…"

Finally, only one classroom teacher identified professional improvement and growth as an underrepresented (undervalued) attribute within the FFT. This teacher comments:

> … I do believe there is a, if I can remember right, there is small portion in there that talks about what are teachers doing to get better. I would think that becoming a master teacher, that would be worth some more points than just presenting at a school staff meeting… (Teacher 101)

From the above, the perspective of missing/incomplete is a dominant theme for stakeholders when reflecting on the content adequacy of the Danielson FFT and its suitability for identifying/differentiating instructional quality. However, it is noted that the concept of missing/incomplete does not specifically address the appropriateness of those FFT elements that are actually measured. Indeed, this consideration—*Reflection on Existing FFT Components*—characterizes the second primary narrative revealed within the data.

Instructional complexity. As mentioned, the concept of instructional complexity is implicit within many stakeholder comments. Complexity within a dynamic instructional

setting suggests that no single instrument or method is able to capture the many-faceted, unpredictable, nature of classroom teaching.

In this regard, a district level participant notes that there are factors impacting instructional effectiveness that remain outside the control of teachers. This lack of control makes it difficult to reduce professional practice down to a "…widget system of measurement…":

> …An effective teacher, or a good teacher, does their best with what they have, but there are so many variables that they cannot control, that to say, "This evaluation system is effective or it's a useful tool and it's valid," I don't think we could really ever get to it. We could get close, and we can keep working at it, but I don't think we could ever have the widget system of measurement applied to students…" and "… There's a lot of variability in the profession that cannot be measured or accounted for… (District 302)

In this sentiment, complexity is a central concern (i.e., "…so many variables…") where some factors are out of the teacher's control. Importantly, these outside factors are not taken into account and cannot ever be taken into account by an evaluation system saying "…I don't think we could really ever get to it [instructional complexity and quality]…" For this person, the inherent complexity of instruction means that instructional quality cannot be reduced to a "… widget system of measurement…" By implication, any system that attempts to do so will necessarily produce an inadequate representation of competency (i.e., there is futility in developing a comprehensive teacher evaluation system).

Narratives involving complexity also imply the need to develop a clearly defined construct of good/effective teaching. A state-level participant farthest removed from the details of the Danielson framework, expresses frustration with the lack of operational clarity in the evaluation framework, states:

> … Teaching performance? What performances? Is it just teaching? Is it just what happens in a classroom? Does that make the quality of the teacher? Is it just the principal's observation during three, or two, or whatever it turns out to be, and I guess now, one full lesson a year? … The [state] statute doesn't say that the districts have to define what good teaching is. Nobody says you have to define what Developing is, or what Effective is, or Highly Effective. You just have to define what Ineffective is from the statute's point of view… (State 402)

The implication is that without a well-articulated definition of instructional quality, one that brings in all facets representative of good/effective teaching, any evaluation framework will inevitably be incomplete and imperfect in its measure. In addition, this lack of uniformity/specificity gives districts leeway to develop differing definitions, rendering any single district's measure of instructional quality suspect. The same state participant responds:

Interviewer:	Would I be far off in saying that there are 220 different definitions? I know that number is a little off.
State 402:	If you want to throw in the charter schools that we have, we probably do have 220 different definitions.

Adequacy of existing FFT components. Participants did not single out specific components within the FFT for removal. Indeed, the second dominant theme in the stakeholder narrative (*Adequacy of Existing FFT Components*) was that of general support and approval. Members of all groups reflected positively on the FFT content, believing it incorporated many important attributes of good/effective teachers: specifically its clarity, structure, and specificity. As noted above, the caveat from many stakeholders is that additional instructional components need be added to improve its content representation.

As before, only the teacher, principal, and district stakeholder groups have an intimate understanding of the FFT as implemented by the district. The narratives from

these members highlight key benefits of the framework including facilitation of dialog and communication, content clarity/structure/specificity (i.e., components that reify specific attributes of good/effective teaching), and the use of an operational rubric to identify distinctions in professional practice. One teacher responds directly:

> Interviewer: Do you feel that the 22 elements in the Danielson Framework (as a collection of behaviors and things that teachers do) adequately represent what it means to be a good teacher?
>
> Teacher 103: … Yes, yes. I've looked through it thoroughly, … I think it does because it does go from, well, communicating with families, to student behavior, procedures, academics, professional development, leading, following…

After responding, the teacher continues to highlight concerns related to rating bias and the lack of time evaluators spent in his/her classrooms, the latter being a big concern for many participants. However, for this teacher and others, the concerns are not with the existing FFT components but with missing/undervalued attributes (discussed above) and issues related to basic measurement quality such as rater bias, reliability, time/frequency of observation, and the impacts of non-instructional factors. These types of measurement-related issues are discussed in more detail under Research Question RQ1B (d).

In the reflection below, another teacher extols on the opportunity to communicate and dialog with his/her evaluator about the FFT components:

> … one of the things that I really like about the [Danielson portion of the] evaluation process right now is that the principal and the teacher can sit down and say, "This is what I gave you. This is what I saw. This is what I've given you: Proficiency." Then it gives that interaction where the teacher can say, "I understand that you think that, but I have this document, this document, and this research, and I think that I should be distinguished." Our principal here is very open to that. As long as all principals are like that, then I don't think there's a problem because the trust is there… (Teacher 101)

Arguably, for this person, the opportunity to have "interaction," dialog, and critical reflection seems to reinforce higher levels of trust in the evaluation results (i.e. "…the things that I really like … because the trust is there…"). The caveat is the concern over whether all principals (i.e. evaluators) actually facilitate/support this process of critical reflection and openness (i.e., "… as long as all principals are like that…").

Expressing "…I have this document, this document, and this research…" speaks to the value placed on the FFT's reliance on evidence and use of a rubric-based criterion for scoring instructional performance. After review of the evidence, trust originates from the evaluator's adherence/application of the rubric to arrive at an appropriate score.

A different teacher also expresses support for the FFT but laments the lack of time allocated to dialog and reflection:

Interviewer:	How well do you think the current implementation of the teacher evaluation process identifies and distinguishes good/effective teaching?
Teacher 104:	Probably about 70 percent. My opinion is about 70 percent. Only from the things I experience and see and my cohorts that I know and talk to.
Interviewer:	It has some good parts, but it's not quite there yet?
Teacher 104:	Oh yeah, absolutely, yes. Definitely has some good parts, without a doubt. It helped guide me. I went through it. I knew the tools that I needed. I am a checklist type of person. I am super checklister, so having the evaluation, and looking at things I needed to have, helps me with my organization, without a doubt. I know the tools I need and the tools that are available.
Interviewer:	How would you articulate what is missing in the capturing of information for evaluating a good teacher?
Teacher 104:	More one on one time, face to face, behind closed doors, with a teacher and the administrator.

Both teachers praise the opportunity to discuss their performance with evaluators. The difference between the two is time allocation. Regardless, for these individuals, communication, dialog, and critical reflection are valued aspects of the evaluation process.

In the following exemplar, another teacher comments on how implementation of the new Danielson FFT framework added specificity and clarity and improved on prior approaches. However, the person still believes some content areas remain incomplete:

> Interviewer: Do you think that the teacher evaluation process effectively or accurately distinguishes between more/less competent teachers? How well does our system do this now?
>
> Teacher 106: I think it's better now than it was. I definitely think that the new rubric and everything is more specific, and, I think, just the different categories … I know when we were going through the training it was a lot more understandable to us - what the expectations were … Yes. It was definitely more defined …I think it hits a lot of the main components [of good/effective teaching].

This teacher was previously referenced as saying:

> … I'm trying to think what I would think would be lacking … I don't know. It's hard to say, because I think what makes a good teacher encompasses a lot more than what's on that rubric. (Teacher 106)

The dialog represents a view that the components currently present in the FFT reflect important attributes of good/effective teaching while at the same time omitting some important attributes. Arguably, this person is suggesting that the system can be made more (content) representative by modifying/adding missing elements into the evaluation process.

Still a different teacher was very direct in his/her views of the framework:

Interviewer: Do you feel that the twenty-two Danielson elements that you're evaluated on cover all aspects of what it means to be a good teacher?

Teacher 102: Yeah. I think the measure is - it's got what it needs to be there.

Arguably, the phrase "…it's got what it needs to be there…" either reflects a minimum expectation of content coverage or is a statement on its comprehensiveness.

Principal reflections are generally positive regarding the comprehensiveness of the FFT. That is, most are less likely than teachers to qualify their support of the FTT with secondary concerns or issues such as missing/underrepresented attributes. A praises the FFT's capability to assess different aspects of teaching using its multiple-domain structure:

> I think the actual Danielson framework, all of the different domains and everything, they definitely do [identify a good/effective teacher]. There are areas for us to document and to talk to, and bring up those discussions about "how do you keep yourself up to date on the education, the knowledge, you have of your curriculum, the knowledge you have of your students". All the different content together, the sections, they build us a very good picture. (Principal 201)

Another principal (below) shares a similar conviction:

Interviewer: Do you think that Danielson captures the essence of good teaching pretty well?

Principal 208: I think it does. I think the components that are laid out there definitely would be exactly what we'd be looking for. Going back through the components I couldn't think of anything else that I would add; that I felt like I would need to know.

Each of these principals (201, 208) expresses their strong support using phrases like "… they definitely do…," "… a very good picture…," and "…exactly what we'd be looking for…" Similarly, another principal states emphatically "…it's not a guessing game … we

know this is what you have to have…" and then reflects on the specificity and detail provided under the FFT's multi-faceted framework:

> But again, it's not a guessing game. Now we know this is what you have to have. We're all on the same playing field, the data is there to support it, and it's not just what your evaluator feels at the time, but it's more of you're informed as to what you're working towards … that rubric is our guide to determine what is it that we have to do and strive to do to be the best that we can, so it's not a guessing game anymore. It really comes down to what evidence we provide to ensure that we are working towards having our students master the content. (Principal 203)

Interviewer: If I paraphrase it right, the evaluation system adequately breaks out the components of good/effective teaching?

Principal 203: Yeah, yeah. … so the Proficient, Distinguished, with us using Charlotte Danielson, so bringing all of that together because in the past I don't think it was as clear. I don't think it was as clear as to what I needed to do as an educator to be a "good teacher, good educator" in whatever role I was fulfilling.

Interviewer: Right now you think it pretty well covers a lot of the things about what it means to be a good teacher.

Principal 203: Yeah, I really do, but like I said, I'm sure next week is gonna be like, "Oh, wow, I didn't think of that," but that's part of growing as an evaluator.

This narrative is grounded in the concepts of data, evidence, objectivity, and certainty. In turn, they authorize attributes of trust, reliability, accuracy, and precision onto the measure. In addition, the phrase "…that rubric is our guide to determine what is it that we have to do…" implies that the FFT specifies all the components of good/effecting teaching, rationalizing a causal performance pathway of *if you do these things; you will be a good teacher*.

Interestingly, one principal who previously expressed confidence in the FFT framework also noted that its proper application/use requires familiarity and ongoing practice with its implementation:

> … I think, yeah, it does. As we get to use it more and more, I'm getting more comfortable with being able to place things that may not be cut and dry. This is student engagement. This is that. The more we use it, the more we can be able to put those things in there … (Principal 201)

The "…may not be cut and dry…" narrative also suggests that some unrepresented (or underrepresented) attributes might be accommodated into the FFT after evaluators gain additional experience.

The sentiments of district-level stakeholders (presented below) are similar to other groups: that the elements contained within the Danielson framework reflect attributes associated with quality instruction. However, like teachers, the caveat remains that some important attributes remain unmeasured.

In the following exemplar (below) a district participant perceives the FFT as an effective tool for assessing teacher's professional practice. At the same time, he/she qualifies the perspective by citing the framework's inability to capture selected affective impacts such as student motivation (note: this is an expansion of previously referenced narrative for this individual). The comment reflects the nuance in the narratives where stakeholders generally value the FFT components but acknowledge they are incomplete. The individual comments:

> … On the professional practice piece, the Danielson model, I was fortunate to use the Danielson model before I even came to [district] as a principal in another district. I have a lot of experience with it. I believe that it's very effective in terms of an evaluation system for teachers. It provides a lot of opportunity for growth of teachers and to really look for ways of improvement. It provides a lot of opportunity to spark conversation about good instruction and about what's happening in classrooms… (District 303)

> Interviewer: Are there things missing? How adequate is the Framework to measuring a good teacher?

District 303: That's a good question. I think Danielson's model is probably one of the better ones that I've seen, but it's certainly not perfect. For example, it doesn't capture a teacher's ability to truly motivate their kids or to make learning come alive for their kids. … I would say Danielson, while it's one of the better models, it still narrows the scope of what people are looking for.

In this comment, the district participant notes the effectiveness of the FFT to evaluate teachers (i.e., "… it's very effective … it provides opportunity for growth…") and its ability to facilitate discussion/reflection on important aspects of teaching (i.e., "…spark conversation…").

At the same time, the perspective is qualified as "…it's certainly not perfect…" Two attributes believed to be missing are "… a teacher's ability to truly motivate…" students, and the ability to make learning "…come alive…" As a result, despite its positive attributes, the participant concludes that the framework "…still narrows the scope…" of assessing good/effective teaching.

Another district participant (below) echoes nuanced support of the FFT:

> The Danielson framework] is very much like a rubric. We put numbers to it. We put labels to it, but it does give us an opportunity to look at really those resources that are being utilized, how a teacher is interacting with their students, to support all those things that we talked about, that facilitating and mentoring, and all of those pieces. (District 304)

For this individual, the framework provides "…an opportunity to look at…" important aspects of instruction including "…interacting with their students… facilitating and mentoring… all of those pieces…" In addition, it is the framework's rubric that permits targeted reflection by structuring and clarifying these instructional components (i.e., "… we put numbers to it. We put labels to it…"). The same individual continues:

Interviewer: I'm hearing you be kind of positive about the Danielson framework and the process as being a strong part of our evaluation system?

District 304: Absolutely. Because, again, we can give that feedback I talked about earlier, to a teacher, about how they are doing, not just as a test giver, but really as that mentor and that designer. What are those interactions? Show me that evidence that you are personalizing instruction with that student. You can get down to that level if it's used correctly. … I think, again, Danielson is wide enough that you can add things in. I mean, we talked about that when we were talking about technology.

Here, a good/effective teacher is portrayed as "… that mentor and that designer…" The framework permits reflection on "…those interactions…" and the degree to which the teacher is "…personalizing instruction with that student…" Because of its structure, the framework allows evaluators to "…get down to that level …" so that "…we can give that feedback…" The participant then adds the caveat "…if it's used correctly…" supporting a sentiment offered by a few other stakeholders that proper utilization of the Danielson framework requires understanding, training, and experience.

Interestingly, this district member (above) reflects that "…Danielson is wide enough that you can add things in…" a reference to earlier comments regarding missing/omitted instructional attributes (i.e., "…when we were talking about technology…). In this perspective, the framework is adaptive and flexible, capable of being modified to improve overall content representation.

Finally, a different district participant reflects upon the work that the organization has undertaken to implement its new evaluation system. The narrative notes that the district has "… done the best we can…" implying concerns and/or shortcomings remain with the evaluation activity. Regardless, it is argued that any organization must adopt some type of suitable "tool." Here, the Danielson FFT is seen as one of many different

'tools' that could have been utilized, each composed of similar core components, all of which display shortcomings and limitations:

> … I think we've done the best that we can. What I do believe about what we've done within our district is we've adopted Danielson's work and are using that consistently … I think you have to have a defined tool that says, "These are characteristics that we look for in a good teacher. These are the things that we would expect to see that cause students to achieve academically and to be successful not just in content."
>
> There are several rubrics out there. There are consistencies amongst some of them. They all hone in on similar areas. They've labeled them a little bit differently. There are some very specific areas like environments, and structural strategies, management techniques, professional behavior, planning. Those are in any tool that you're going to look at for evaluating a teacher. Those characteristics are going to be there. (District 301)

Summary – RQ1B (c) stakeholder narratives. In summary, the narrative provided by stakeholders reveals a very positive perspective of the Danielson FFT. The framework was generally seen as embodying many characteristics of good/effective teaching. Comparatively, principals were less likely to qualify their support of the FFT with shortcomings or concerns compared to teacher or district representatives.

However, it is apparent throughout the stakeholder discussions that important attributes are believed to remain unmeasured or under-measured. Mostly, these problematic attributes are affective in nature, related to impacts on students or to teachers' professional practices. There is the view that any attempt to reduce the professional practice down to a small set of measureable variables is difficult. Even so, the FFT is seen by many stakeholder as flexible, able to accommodate modification and adaptation. In addition, the concept of good/effective instruction is perceived as complex, multi-faceted, and dynamic. This perspective permits stakeholders to assert satisfaction with the elements presently contained in the FFT model while at the same time believing

the FFT may be substantively improved by incorporating additional affective components.

Petite assertions - RQ1B (c) stakeholder narratives.

Petite assertion #1. The measured elements contained within the FFT represent a subset of important characteristics within the TIQ construct. Stakeholders believe the existing components should be retained within the evaluation system.

Positive attributions to the FFT system include:

- Facilitates dialog, communication, and critical reflection of professional practices
- Inclusive of attributes characteristic of good/effective teaching
- Organization: Clarity, structured, relies on evidence, objective, rubric-based
- Flexible, adaptive, ability to modify
- Improvement over past evaluation tools and methods
- Multi-faceted, multiple domains of practice

Petite assertion #2. Stakeholders perceive that the FFT omits important attributes defining the TIQ construct. These attributes included:

Affective (Personal/Emotional) Impacts on Students:

- … Motivation, engagement, inspiration, desire, dedication, perseverance, connection, self-perception, value, trust, experience;

Attributes of Professional Practice:

- … (Affective):Dedication, commitment, passion, classroom environment, professionalism, relationship/connection with students,

- … (Practice): Content knowledge, use of technology, FFT differentiation across instructional contexts/settings, school-based professional activities, collaboration, involvement, mentoring (students), counseling (students),

Petite assertion #3. Stakeholders believe that adherence to the structured FFT framework acts to narrow the scope of evaluation and reduce the inferential integrity of scores (results).

Petite assertion #4. Principals hold substantively more positive perceptions of the FFT's content efficacy than do teachers, district, or state participants.

RQ1B (d). Do perspectives differ among stakeholders regarding the capacity for VAM and PP measures to adequately represent and differentiate the instructional quality of classroom teacher? The approach is stakeholder interviews. The measures are coded interview responses.

Introduction. Research Question RQ1B(d) examines stakeholder perspectives regarding the efficacy of the evaluation system to properly identify and distinguish Teacher Instructional Quality (TIQ). It specifically examines whether these perspectives differ across group membership. To the extent that all member groups share the same perspective, the presence of a well-defined efficacy construct is supported. Here, a unified perspective reflecting generally positive sentiments would suggest the evaluation system does an adequate job of measuring TIQ. In contrast, a unified perspective reflecting generally negative sentiments would suggest the opposite. Finally, divergent perspectives across stakeholder groups would raise questions regarding alignment between the TIQ construct and what is actually being measured.

The reader is reminded that four stakeholder groups (n = 22) were identified in the research design: teachers (n = 7), principals (n = 8), district policy (n = 4), and state Policy (n = 3). A common interview protocol was administered to all groups. During the interview activity, all participants were invited to reflect upon the ability of the "system" to identify good/effective teachers based on the following general prompt: *How well does the district's teacher evaluation system identify, and distinguish between "good/effective teachers"?* Follow-up discussion probes included:

1. Do the FFT elements cover all aspects of what it means to be a "good/effective teacher"?
2. Are characteristics of "good/effective teaching" missing in the FFT Framework?
3. Are test scores a suitable indicator of "good/effective teaching"?
4. Which would you place more confidence in for identifying a "good/effective teacher": evaluator observations or test scores? Why?

Narratives from the first two probes (regarding the efficacy of the Danielson FFT) were discussed in detail under Research Question RQ1B(c). As such, much of the discussion detailed under the current question focuses on the efficacy of the overall evaluation results and the contributory role of test scores (probes 3 and 4).

Construct overview. Stakeholder narratives concerning the efficacy of the evaluation system are organized into three dominant components: (1) *General Efficacy*, (2) issues of *Time, Frequency, and Observation*, and (3) concerns related to *Test Scores/Measurement.* To clarify the analysis, *Test Scores/Measurement* is further segmented into two sub-concepts: *Test Score Adequacy* (TSA) and *Metric Preference*

(MP). For each component/sub-concept, the perspectives within individual stakeholder groups are examined. A summary of findings is then presented followed by a collective assessment of findings aligned to this research question.

The qualitative structures assembled for this research question is multi-dimensional and nuanced within/between each of the four stakeholder groups. Figure 36 is provided to aid in reader understanding of the generalized component structure extracted from the narrative.

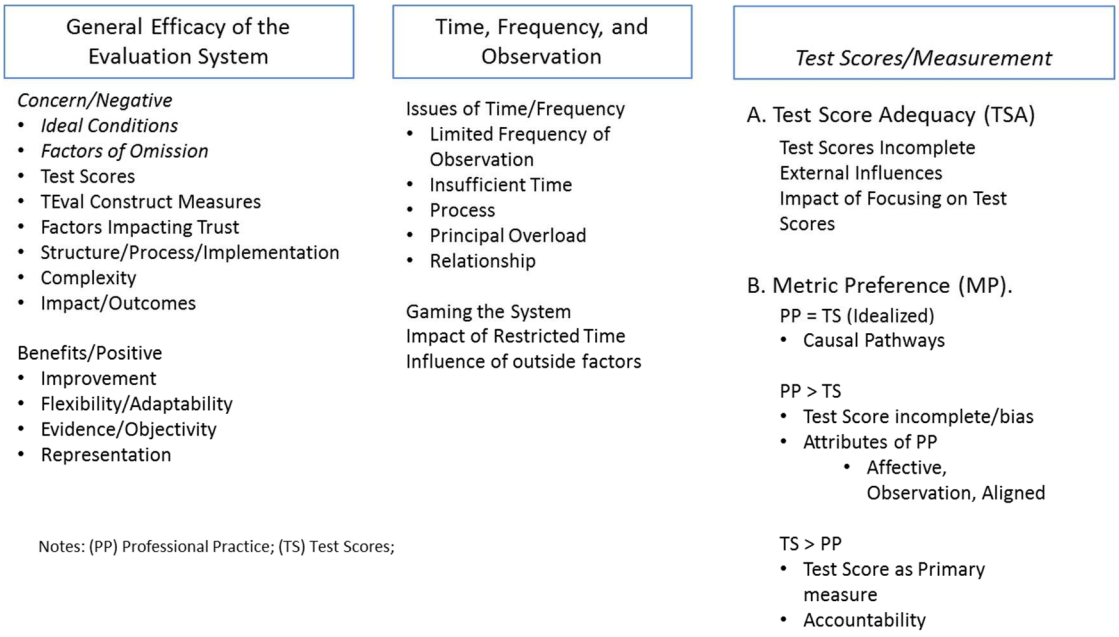


Figure 36. Component structure for Research Question RQ1B (d).

A review of the component conceptualizations with accompanying codes and identities are provided below.

Component 1 - General efficacy (of the evaluation system). General Efficacy (GE) refers to the overall ability of the evaluation system to adequately identify and distinguish between good/effective teachers. The reader is reminded that the evaluation system is comprised of two primary components: tests scores and ratings of professional practice. When combined, a generalized performance rating is assigned expressed in terms of one of four descriptive classifications: *Ineffective, Developing, Effective,* and *Highly Effective*. In this regard, *General Efficacy* is conceptualized as follows:

> General Efficacy: General Efficacy is an inclusive concept that combines all facets of the evaluation process (i.e., the Danielson professional practice ratings and achievement growth measures). It refers to the ability of the evaluation process to accurately identify, and distinguish between, good/effective teachers. In this way, it is concerned with the adequacy of the final evaluation outcome (total score and/or the overall performance classification). General Efficacy is derived from the collective narratives of the prompts/probes regarding each of the evaluation components (test scores, professional practice ratings, issues or measurement, content representation, etc.).

Codes and identities extracted from the stakeholder narratives delineating *General Efficacy* are included in Figure 37.

Issues/Concerns/Negative	
Ideal Conditions (TEval Appropriate Only Under Ideal Circumstances) • Requires no rater bias, everyone highly trained, consistency across schools, no rater personal/relationship bias; no external factors (i.e., principal not tired, not overloaded, and/or there is plenty time/frequency of observation); • Requires lots of time available and utilized for discussion and reflection; conference time; Factors of Omission: • Intangible Factors: There's a tone, there's a feeling (in the classroom); other part of education, a child's experience coming to school; not complete: missing, unmeasured attributes/factors (student and teacher); missing affective impacts; doesn't measure "feeling of classroom"; Test Scores: • Limited content (R/M); limited snapshot, incomplete, missing attributes of teaching; AIMS scores inadequate; AIMS does not measure student motivation (omitted attributes); test scores not suitable indicator; TEval Construct Measures: • Questions of rater reliability, validity; lack of consistent definition of quality teacher (i.e., differs district to district); the definition of quality teacher changes with context (i.e., CTE, arts, traditional schools, alternative schools, etc.)	Factors Impacting Trust: • Rater reliability, low time/frequency of interaction/observation (inadequate personal, get-to-know, etc.), relationship bias (not objective), uncontrolled student factors, influence of non-instructional factors; missing attributes on both teacher PP and student impacts; not representative of whole teacher or whole student; Evaluators can't distinguish between *Distinguished* and *Effective* teachers; Evaluators not in classrooms enough (i.e., lack of time/frequency – relationships, observation); Structure/Process/Implementation: • Evaluation model too simplified; single evaluator model is a poor implementation approach because evaluator is also the supervisor; • Statewide inconsistency, lack of uniformity; some districts do a better job of implementing the system than others; Complexity: • Teaching complex, "too many unknowns"; no single instrument (i.e., FFT) can cover all aspects of teaching; So many variables that cannot be controlled; too difficult to accurately measure; Impact/Outcomes • Narrows instructional focus; teaching restricted to measured FFT components and tested content; test scores force reduction, do only what is tested; • Negative effects on organization; stress, competition, isolation; missing grade levels (K-2) because of no standardized (state) test;
Benefits/Positive	
Improvement • Evaluation methods getting better over time; improving, but not there yet; doing best we can; have some foundational pieces; much better than previous system; Flexibility/Adaptability • Can add missing measures/attributes; adaptive, flexible, dynamic;	Evidence-Based/Objective • Evidenced-based, not opinion, objective, "what I saw"; structure removes "subjective stuff"; "not a guessing game"; rubric is a guide for objective ratings; provides structure/clarity/focus; Representation • FFT and TEval process builds a good picture of instructional quality; accounts for many important attributes; combination of Test Score and FFT provides suitable quality rating; (principals);

Figure 37. Codes and identities extracted from the stakeholder narratives delineating *General Efficacy* (of the evaluation system).

Component 2 - Time, frequency, and observation. The *Time, Frequency, and Observation* component was emphasized primarily by the teacher stakeholder group. However, some members within each remaining group (principal, district, and state) incorporated this concept into their discussions. The *Time, Frequency, and Observation* component is conceptualized as follows:

> Time, Frequency, and Observation: This component focuses on time (i.e., total time required to evaluate classroom teachers) and frequency (number of observations or interactions with teachers). It is anchored in the basic structure and procedural requirements of the evaluation system (i.e., number of formal observations, time allocated to conduct evaluations, etc.). The concept of relationship is important to this component. The practice of evaluation requires interaction, observation, communication, and reflection of practice. This necessitates interpersonal relationships established between evaluator and teacher.

Codes and identities extracted from the stakeholder narratives delineating *Time, Frequency, and Observation* are included in Figure 38.

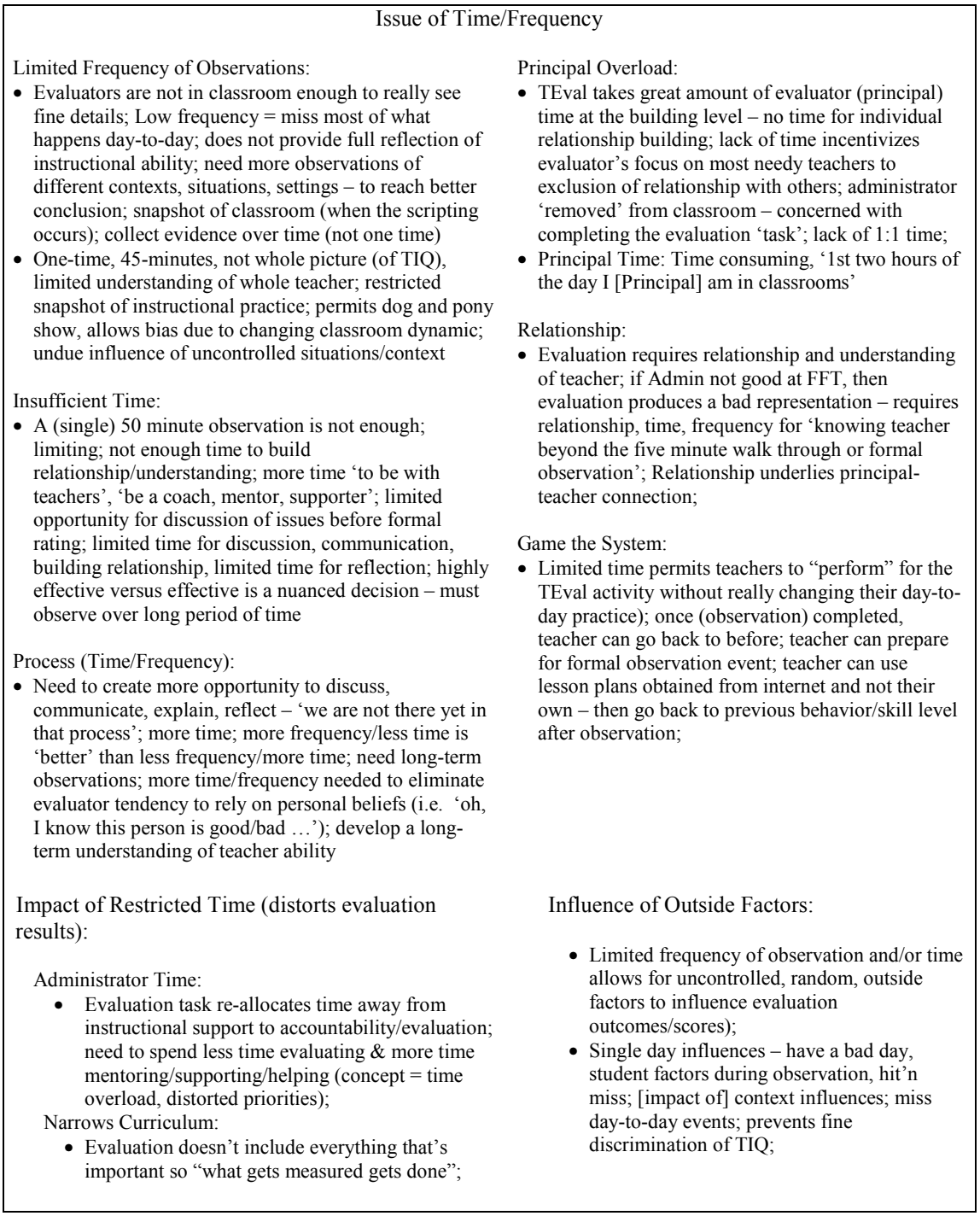


Figure 38. Codes and identities extracted from the stakeholder narratives delineating *Time, Frequency, and Observation*.

Component 3 - Test scores/measurement. The *Test Scores/Measurement* component represented a dominant theme throughout the narratives provided by all

stakeholder groups. It is broadly interpreted as the ability of evaluation metrics to adequately measure/represent the Teacher Instructional Quality (TIQ) construct. In this way, it concerns reflections on the measurement characteristics of both test scores and professional practice ratings. However, as discussed below, the measurement attributes of the (Danielson) professional practice scores were previously discussed under RQ1B(c). As such, much of the discussion presented for the *Test Scores/Measurement* section herein focuses on characteristics associated with (value-added) test scores as used in the evaluation process. To facilitate the analysis, the *Test Scores/Measurement* component is examined in terms of two sub-concepts evident in the original qualitative data: *Test Score Adequacy* (TSA) and *Metric Preference* (MP).

The stakeholder narratives concerning this research question were initiated by the following generalized interview protocol:

> *How well does the district's teacher evaluation identify, and distinguish between, 'good/effective teachers'?*
>
> - Probe 1: *Are test scores a suitable indicator of "good/effective teaching"?*
> - Probe 2: *If the two measures did not agree, which would you place more confidence in for identifying a "good/effective teacher," evaluator observations, or test scores? Why?*

For the purpose of presenting the analysis, the first probe is codified below in under the section header *Test Score Adequacy* (TSA) while the second under *Metric Preference* (MP).

The generalized *Test Scores/Measurement* component is conceptualized as follows:

Test Scores/Measurement: The Test Scores/Measurement component is conceptualized as the attributes/characteristics of the metrics used in the evaluation process to assess the overall instructional quality of classroom teachers. It is not restricted to statements of scale or statistical reliability. Rather, it also includes broader issues such as bias, the influence of non-instructional factors (construct irrelevant variance), and missing measures – any and all factors stakeholders believed influenced the quality of evaluation metrics.

The component is further broken into two sub-concepts: *Test Score Adequacy (TSA*) and *Metric Preference (MP).* See Figure 39. TSA concerns attributes specific to the use of standardized test scores in the evaluation process while MP concerns stakeholder perceptions of which metric, test scores or professional practice measures, reflect a more adequate representation of instructional competence.

Test Score Adequacy:	*Metric Preference:*
Test Score Adequacy examines stakeholder perspectives on the adequacy of standardized test scores to serve as an indicator of instructional quality. To the extent that stakeholders accept test scores as a measure of competence, their inclusion into the evaluation process is supported. Rejection of this premise raises questions regarding either construct adequacy and/or the need to utilize alternative indicators.	*Metric Preference* examines the relative importance of test scores and professional practice ratings to inform on good/effective teachers. The perspective is examined from two perspectives: (1) should the two independent measures of instructional quality align (i.e., should they correlate and have the same relative magnitude), and (2) if they do not align, which is a more accurate representation of true instructional quality?

Figure 39. Two sub-concepts of *Test Scores/Measurement.*

Codes and identities extracted from the narrative information related to the *Test Score Adequacy* (TSA) sub-concept are listed in Figure 40.

Test scores Incomplete:	External Influence:
• AIMS not representative of 21st century standards; test scores don't contain the educational outcomes we now value; not complete; AIMS not aligned with Career and College Ready (CCR) standards; Current AIMS is not proper measure (improper content, too narrow); Need to add other performance measures besides AIMS; Curriculum in transition to Common Core; Transition to PARCC	• Test scores influenced by student motivation, parent motivation to help students, things outside of control of teachers
• Not representative of non-tested instructional impacts on ability (speeches, projects, effort, reasoning, etc.); Can't reduce year of instruction to a test score	Impact of Focusing on Test Scores
• Group B has no aligned instructional measures; need to add KG, 1, & 2 – most important development years with no current impact data (not measured = not important = not attended to) [Reader's Note: "Group B" refers to teachers who lack state assessments of their instructed curriculum]	• Focus on test scores narrows curriculum to only what is tested, doesn't include everything that's important; 'what gets measured gets done'; instructional impacts on early grade levels not assessed by tests

Figure 40. Codes and identities extracted from the narrative information related to the *Test Score Adequacy*.

Codes and identities extracted from the narrative information related to the *Metric Preferences* (MP) sub-concept are provided based on the stakeholder perception of alignment/importance between test scores (TS) and professional practice (PP) ratings (i.e., three perspectives were expressed: PP = TS, PP > TS, and TS > PP). See Figure 41 and 42.

Codes/Identities for PP = TS (Aligned)	Codes/Identities for PP > TS (Not Aligned)
PP & TS should be aligned: Both needed for TIQ; high PP teachers should impact growth => good teachers should have high scores; *Causal Pathway – Idealized Alignment:* TS = PP if order of effects are (1st) improve affective environment (motivation & engagement, satisfaction, enjoyment of learning, teacher moral), which leads to (2nd) positive culture free of stress, lots of collegiality, support, communication, common goal, mutual respect, which leads to (3rd) increased achievement as measured on tests *TS Issues:* PP = TS in ideal setting but effects of poverty, economy, home issues, motivation, engagement, emotion, etc. means PP will not aligned to TS; TS <> FFT (correlate) due to many influencing variables on TS; *PP Issues:* PP = TS should occur if PP observations are unscheduled – but if PP observations are scheduled then formal evaluation will be biased (artificially high PP ratings) causing PP <> TS; *Learning is goal:* If teacher well-prepared, instructs well, is professional, this should lead to higher TS meaning TS & FFT should be aligned (especially using growth scores); *Implementation/System Issues:* TS = FFT if TEval system implemented to improve total system; wrong to focus on individual teacher in isolation; PP measures important for mentoring and improving TIQ in order to eventually increase TS;	*Omitted Attributes (Test Score):* TS does not include affective aspects of teaching, therefore can have high PP and low TS; TS exclude factors such as interactions, relationships, personal factors/contexts, parents, peers, community, etc.; TS excludes developmental time-pathways learning; TS limited perspective (outcome, not process); TS limited aspect of what it means to be a good teacher; too restrictive a view *Test Score Bias Issues:* TS snapshot in time, single-day; TS impacted by student motivation and other non-instructional factors; TS influenced by factors outside of teacher control (validity) but PP inside of control; TS impacted by human element - bad-day, lives of kids, factors outside of teacher control; random assignment issues; compositional issues on fairness, equity across instructional settings contexts; impact of student outliers on classroom aggregate scores *Research:* 'No evidence' that TS = good teacher; state requires (imposes) use of TS (not educators); Can have high TS and not be a good teacher and vice versa *Attributes of PP:* Evaluate over entire year; based on relationship & personal knowledge ('knows me very well'); based on discussions; evidence from observation (walk-through, over time) using rubric making it more valid than TS; PP is teacher-situational specific; based on established 'best' instructional practices; multiple observations/measures; lots of conversation, connection with principal; FFT promotes dialog, reflection, *PP Includes Affective:* TS can be high but still have students/parents not satisfied with education – low affective measures in the presence of high TS = TS are incomplete; FFT reflects relationship, connection with students; *Aligned to Classroom/Teaching:* FFT aligned to classroom activities (act of teaching), students while TS is snapshot; FFT = observation of classroom events (not reflected on TS); FFT more fair - evaluators are in classrooms, see teaching, PP, effort and strategies; FFT is classroom – able to 'see' other pieces; FFT captures PP; FFT more complete measure; FFT reflects actions in classroom, instructional activities, methods, all aspects of teaching; FFT observes classroom actions => 'see classroom', 'hear instruction', 'feel tone', 'see responses', 'see interaction between teacher and student'; FFT observation provides 'deep context of learning environment' – interaction, relationship, thinking processes, information processing *PP Trained Observers:* Trained observers; 'really know what they are looking for';

Figure 41. Codes/identities for PP = TS (Aligned) and PP > TS (Not Aligned).

Test Score Primary Measure:

- Education Goal: Purpose of education is learning 1st; if students not learning then teacher ineffective; It is outcome that matters, not the effort or affective; 'Kids here to learn'; PP doesn't measure if students have learned anything; academic growth more important than PP measure if bottom line concerned is content learning;
- TS Requirements: TS bottom line but TS must be unbiased, reliable, complete in content, not impacted by outside factors, reflect true learning - not snapshot;
- Causal Pathway: TS is best measure because learning leads to opportunity which leads to better life and success in future;
- Public Perception: If PP does not match TS, public will always value TS measure as indicator of quality;
- TS as Indicator: TS measure of progress & effectiveness of instruction; TS reflects specific content required to learn; if TS low then ask 'what am I doing wrong' (in instruction); TS validates/confirms PP; if TS<>PP then question PP ratings (biased); High TS means good/effective teacher;

Accountability:

State/Society: Require some basic measure of educational outcome; accountability for content learning; need to measure with standard approach (i.e. Standardized Test Scores), not teacher-made tests; need a comparative measure that is reliable/valid (i.e. Standardized Test Scores)

Test Score More Reliable Than PP:

If high VAM & low FFT -> would question the FFT (implies TS more reliable, stable, accurate); If TS <> FFT -> problem is with FFT - 'cause me to self-reflect' on my FFT ratings; TS more reliable, accurate, less prone to bias, subjectivity than FFT; TS has less error; FFT is (too?) subjective but TS is objective, reliable, more precise; qualifier = 'if tests are done well';

TS more valid if they trend over time (consistency over time); if FFT > Low TS over time, then begin to question FFT ratings; need multi-year TS to provide valid inference on outcomes

FFT Issues:

FFT biased by relationships, personal knowledge, history => TS > FFT; TS are created by experts -> removes questioning of quality of TS measure; TS is 'our purpose', 'most important to me' => TS > FFT; low growth -> teacher needs to improve PP; TS embodies instructional effectiveness; good teacher = high growth scores;

Figure 42. Codes/identities for *Test Score > Professional Practice.*

Stakeholder narratives.

Component 1 – General efficacy.

General efficacy - teachers. The narrative provided by teachers is generally negative with regard to the system's ability to derive a suitable assessment of TIQ. It is driven by three (sometimes overlapping) perspectives: a negative reaction to an increased emphasis on standardized test scores, the impact of important missing (unmeasured) attributes, and concerns about rater (evaluator) reliability. The combination of these

factors leads to a general mistrust in the evaluation outcomes. For example, after an extended discussion on concerns related to the evaluation system, one teacher simply responds:

> Interviewer: I'm interpreting this [discussion] as there are parts of the teacher evaluation system that raise some trust issues, issues that tends to bias or distort your sense of what teacher evaluation should be?
>
> Teacher 102: Yes. That is correct.[20]

Another teacher expresses a general mistrust of using test scores as part of the evaluation system as well as the validity of the composite (test plus PP) scores:

> No. I don't think it's [the evaluation system] complete. On the students' testing, it's not complete because most of the time, when you've got these kids in, there's not a true test of their knowledge … To put those two together [test scores and PP], I don't think that that is a true component of what a good teacher is. (Teacher 105)

The mistrust is based on the view that test scores fail to accurately measure what students have learned. Similarly, a third teacher acknowledges that the evaluation process is getting better, but is concerned about the increased reliance on test scores. He/she also shares the general view that important attributes remain missing on the FFT.

> Interviewer: Do you think that the teacher evaluation process effectively or accurately distinguishes between more/less competent teachers? How well does our system do this now?
>
> Teacher 106: I think it's better now than it was. I definitely think that the new rubric, and everything, is more specific … It [FFT rubric] was definitely more defined. [However] In my mind, it's so hard to distinguish who's the better teacher, or who's the best, because I feel like we just keep moving towards data [test scores] being the big deciding factor. I don't know, it's hard to

[20] Participant narratives are denoted by membership group (i.e. Teacher = Group 1, Principal = Group 2, District = Group 3, or State = Group 4) and an individual identification number. Thus 'Teacher 106' denotes individual '06' in the Teacher Group (1).

> say, because I think what makes a good teacher encompasses a lot more than what's on that rubric.

This teacher believes that the evaluation process has been improved, in part, from the specificity/clarity provided by the FFT portion of the system. However, it remains incomplete ("…what makes a good teacher encompasses a lot more than what's on that rubric…"). In contrast, less trust is placed on the test score measures ("…I feel like we just keep moving towards data [test scores] being the big deciding factor…").

A third teacher acknowledges test scores as a necessary, or perhaps an accepted, part of evaluation, while at the same time expresses concern that the system is not looking at "… the teacher as a whole…"

> Well, my perspective would be that they [policy makers] look at the teacher as a whole and not just data [test scores]. That would be my perspective, but I know that data is important. That's what makes it even, oh my goodness, harder to be at a struggling school that got a D [state accountability] label; that it's all about data, data, data, data, data. But yet they're not in the room with me when I'm having those amazing moments when kids are getting it. They are smiling and glowing, and they're [the students] going to remember that. (Teacher 104)

Here, the issue is that an excessive emphasis is placed on test scores while important attributes/impacts of teaching are neglected or are missing. This teacher also seems to use *data* in two different contexts. The first is in terms of state accountability "… a struggling school that got a D label…", exclaiming "… it's all about data, data, data, data, data…" where data refers to test scores. This teacher teaches at a school reporting low achievement scores and may be reacting to the heightened emphasis administration is placing on improving the accountability ratings. Nevertheless, increased reliance on these measures seems to be a concern. The second use of the term data relates to the FFT ratings assigned during classroom observations "…yet they're not in the room with me when I'm having those amazing moments when kids are getting it…" The

implication seems to be that the missing, unmeasured, attributes are affective in nature and the evaluation process fails to capture these effects.

Similarly, a fourth teacher acknowledges achievement as one aspect of instructional quality but questions the ability of any evaluation system to capture all of the important impacts teachers have on students.

> In my mind, teacher evaluation is moving more towards looking at data [test scores] and things like that, which is, obviously, very important because, unfortunately, there are teachers who aren't being effective [in raising academic achievement]. But it would be nice if there were an added component … I wish that there was a way to incorporate experiences like that that our kiddos have, that are big successes for some that don't always have those same successes. [the teacher is speaking about affective growth & impacts on students not referenced in the FFT and/or test scores] (Teacher 106)

The phrase "…I wish that there was a way to incorporate experiences like…" suggests a mistrust in the act of quantification, reduction, and/or simplification. It questions whether one can ever really measure the affective, the emotional, or the personal aspects of the teacher-student relationship.

The issue of rater reliability also seems to influence teacher's general perspective on system efficacy. The sentiment is that high rater reliability may be attained only under ideal conditions – raters that are highly trained, are able to apply the evaluation criteria objectively, are able to keep out personal/emotional relationships, and restrict the influence of other factors that might distort or bias an evaluator's judgment. One teacher reflects on some of these sentiments as:

> I would say the district has a lot of pressure on it for making sure that whoever they hire, principal-wise, that they really understand that they're affecting [the] lives of all of their teachers. As long as you have highly effective principals, and the principals are following the rubric, and they're being unbiased … I do think it's a little top-heavy principal-wise as far as the evaluation goes.

> The principal has to be very unbiased, and has to really stick to the rubric that they have, regardless if that teacher is doing a million things at school, or if they're friends, or not friends, or if they've had arguments before.
>
> If, again, going through that, the principal's tired, if this is his third evaluation of the week because there are these time limits that the principals have to do all these [evaluations], there are some teachers that could get shortchanged. (Teacher 101)

Skepticism is raised on a number of levels. First, the principal has to be highly effective. Second, the principal must adhere to the rubric and apply its criteria in an unbiased, objective fashion. Third, the principal has to ignore the teacher's "… doing a million things at school…" and "… if they're friends, or not friends, or if they've had arguments before…" Fourth, the principal has to guard against factors such as being tired and/or being pressured by lack of time. The result of any/all of these intrusions is that "… teachers … could get shortchanged…" The implication is that these ideal conditions are difficult to meet.

A different teacher suggests that evaluator ratings can be influenced (biased) by a teacher's effort to showcase their activities, even to the point of misrepresenting their day-to-day instructional practices.

> I do think that as fair and equitable as the evaluation has tried to be, I still think that there is the ability for admin to see things through rose-colored glasses, that they might see someone's effort as being much larger than maybe someone else who doesn't toot their own horn, who doesn't always go out and go, "Look what I did. Look what my kids did." Some of us don't do that. Then there are people who are great internet searchers and can find things on Teachers Pay Teachers [website] and create these amazing lessons that they didn't create. They got it from someplace else, and it looks really good. You would expect that the kids would learn, and sometimes, it's still a dog and pony show. (Teacher 108)

For this teacher, the teacher-administrator relationship is a source of evaluation bias. In addition, the opportunity for outright deception is considered. Presumably, this latter perspective is linked to the limited time and frequency available for evaluators to

adequately observe instructional practices. The phrase "…you would expect that the kids would learn…" suggests students are negatively impacted by these conditions.

The general perspective from the teacher group seems to be concern over whether the evaluation process adequately quantifies instructional performance. The perspective is co-constructed from concerns over a heightened emphasis on standardized test scores, missing/unmeasured instructional attributes, concerns regarding rater reliability/bias, and (arguably) hesitancy in attempting to reduce instructional quality to a simplistic, quantifiable, measure. Importantly, no teacher outright accused his/her evaluator of unethical or inappropriate evaluation/rating practices. Rather, there was a consistent concern that the issues identified above collectively lead to incorrect or inaccurate representations of instructional quality.

General efficacy - principals. In contrast to teachers, principals provide a much more positive narrative. Most display satisfaction with the evaluation system's ability to properly identify and distinguish the performance of classroom teachers. Rather than detract from accuracy, test scores were generally seen as a necessary component to the system. In addition, most (but not all) principals were less likely to identify substantive missing attributes in the FFT. Finally, a dominant concept present in the principal narrative is *evidence* and *objectivity*. These perspectives seem to provide authority that the system derives an accurate assessment of teacher performance.

One principal (below) expresses general acceptance of the evaluation process:

> I think we've gone too many years not truly evaluating our teachers. I think too many years we've gone by with teachers who are mediocre or worse, that are still in the system and are not doing their jobs. I think finally we've come up with something. Is it the best way to do it? I don't know what the best way to do it is … this is one way to do it … I like that there's components to it. It's not just

> what you do in the classroom. It's what you do in the planning of it and it's also what you do outside of the classroom. I feel like even if you have a weakness, there's other places to have strengths in that can build that score. The other part I like [are] the test scores. I do. I feel like I finally know who my teachers are and their abilities. I finally have an indication as to not just how the teachers do, but how are those students doing. (Principal 205)

This principal (above) believes in the evaluation system ("… finally we've come up with something…"). He/she values the multiple dimensions/measures and the inclusion of test scores. These facets provide important perspective on both teacher competency and student learning. The system provides opportunity for teachers to reveal strengths/weakness in a variety of performance contexts. Finally, the collective attributes of the system permits a perception of confidence and accuracy ("… I finally know … I finally have an indication…").

Another principal exposes support for this multiple component approach in the evaluation design:

> Interviewer: Do you think that the overall thing [evaluation system] captures the essence of good teaching pretty well?
>
> Principal 208: I think it does. I think the components that are laid out there definitely would be exactly what we'd be looking for. Going back through the components, I couldn't think of anything else that I would add; that I felt like I would need to know.

This participant is emphatic, confident, and assured. He argues without qualification. The key terms are "definitely" and "exactly." The comprehensiveness of the system provides an unassailable assessment of instructional quality. Still a third principal speaks specifically about the comprehensive nature of the Danielson FFT:

> I think the actual Danielson framework, all of the different domains and everything, they definitely do [identify good/effective teachers]. There are areas for us to document and to talk to, and bring up those discussions about how you keep yourself up to date on the education, the knowledge you have of your

curriculum, [and] the knowledge you have of your students. All the different content together, the sections, they build us a very good picture. (Principal 201)

A fourth colleague reflects on the combination of the main components (test scores and professional practice ratings):

> … those components of looking at the test scores in addition to the ratings - so the proficient, distinguished, with us using Charlotte Danielson - so bringing all of that together, because in the past I don't think it was as clear. I don't think it was as clear as to what I needed to do as an educator to be a "good teacher, good educator" in whatever role I was fulfilling.
>
> Interviewer: Okay, I'm going to infer that you feel this idea of a system that has student achievement and professional practice to be a good combination?
>
> Principal 203: Yes.

From these narratives, principals express satisfaction toward the evaluation system's ability to accurately inform on instructional competence. A substantive premise underlying their perspective is that of evidence and objectivity. These two concepts collectively permit school administrators to have confidence in both their ability to evaluate without bias and to trust the evaluation results. In addition, and importantly, evidence and objectivity also permit principals to perceive that teachers also share their perspective of fairness and objectivity (a view perhaps not strongly supported by data previously/subsequently presented). In this regard, reflections by three principals are as follows:

> I think it does. I like it. … When I look at the evidence and the data, then it takes all that subjective stuff out of the picture … The tool itself, the Danielson's rubric and the model that we're using online with the CES system, has made me much more aware of [being] evidence-based. I like that a lot, because then it's cut and dried. Here's the evidence. … Then the other piece, the student scores that are factored into the overall evaluations, as a leader, it has helped me. (Principal 202)
>
> It's not a guessing game. Now we know this is what you [the teacher] have to have. We're all on the same playing field, the data is there to support it, and it's

> not just what your evaluator feels at the time, but it's more of you are informed as to what you're working towards. [The] rubric is our guide to determine what is it that we have to do and strive to do to be the best that we can, so it's not a guessing game anymore. It really comes down to what evidence we provide to ensure that we are working towards having our students master the content. (Principal 203)

> The Danielson [evaluation] is very evidence-based, which is huge. It's not my opinion about what you should be doing here and there and this; it's just strictly this is what I saw, this is what I heard, this is what you were doing, this is what the students were doing. I think that was a very critical piece to come so that we could, the teachers could, feel like they're being consistently or fairly evaluated. (Principal 201)

Each appeal to an authority external to themselves: "…it takes all that subjective stuff out of the picture," "…It's not a guessing game," and "…It's not my opinion," The principal's perspective is that the Danielson FFT structure (components, rubric, and the process) combined with test scores yields an exacting measure on the instructional abilities of classroom teachers: the higher the evaluation score the higher the competence, the better the teacher. As previously mentioned, principals are more likely to offer this perspective without substantive qualification. In contrast, other stakeholder groups do not exhibit the same sense of confidence.

Finally, another principal sums up his/her feeling as:

> I think the measurement tool itself is doing a decent job. I think, especially since we've rolled out the rubric and we've really done a good job of training, and informing our teachers of what each category consists of.

Interviewer: Do you think that the system itself does a good job, in the end, of saying "you're a good teacher; you're not such a good teacher"?

Principal 204: I think it does. I do believe the criteria that is set helps take the subjectivity out of it. We are focusing more on evidence, and that has truly helped make this process more clear, not only to administrators, but most importantly to teachers.

For principals, the perception of system transparency, effective training, the reliance on evidence, and rating objectivity collectively supported a sense of competence and acceptability in the evaluation system. This perception permits this group to believe that teachers similarly trust and accept the evaluation results.

General efficacy - district. Similar to teachers, district-level participants generally reflected a negative view of the evaluation process. Members were more likely to reflect on the complexity of evaluating effective teaching, missing (but important) attributes, and limitations of standardized test scores as a state-policy required component of the system. District participants were less likely to comment on (or implicate) issues related to rater bias or reliability (consistency). This may be because they are more removed from (not responsible for) the day-to-day application of the FFT rubric scoring process.

One district participant clearly stated his/her perspective that the state-imposed evaluation framework will not effectively identify and/or distinguish teacher efficacy. The view is premised on the lack of a consistent operational definition of effective teaching. That is, districts throughout the state are free to develop systems based on their own local definition of teacher effectiveness. By extension, this raises questions on the validity of any local implementation, including that conducted by this organization.

> I don't think the framework—the state framework by itself, taken as a tool by itself—will effectively discern differences between good teachers and not good teachers. I don't. I think that it is so broad … there are too many differences. There are too many unknowns within that framework across districts. If we're going to say, "In Arizona, this framework is going to be able to tell us who our good teachers are and who our good teachers are not." I don't think it will do that. (District 301)

This same district participant raises concern over the use of test scores as a (partial) measure of effective teaching. He/she sees this as a source of tension with the

larger legislative requirement. Specifically, the state framework overvalues test scores as an indicator of instruction quality. The narrative suggests that if local districts were free to develop their own evaluation frameworks, the weights accorded to the measured components would be substantively different.

> My definition of good teachers doesn't incorporate some of the things that the state is incorporating as far as good teaching. The state's incorporated student achievement scores as part of what measures a teacher as being good or not good. I think that it's a piece … I think that it's certainly something that should be looked at. I think it can give you some information and a platform for questioning … and reflecting on their [teachers] practice. [But] Whether or not the test scores determine for sure that that teacher's a good teacher or not is - there's no evidence for me to support that it definitely says, "This is a good teacher or not a good teacher." I don't think student achievement necessarily gives us that direct information. (District 301)

Here, the participant is clearly devaluing the contribution test scores as an indicator of instructional quality.

Similarly, a second district participant agrees that test scores provide only limited perspective on instructional efficacy. However, the perspective (rationale) is different. For this individual, it is not that test scores are overvalued; it is that they are incomplete - they only include reading and mathematics.

> We have some foundational pieces. We just don't use everything that's available to us. We limit ourselves when it comes to assessments and evaluation pieces. When we don't look for the opportunity to expand what's already in place and look at other rubrics and other performances pieces … we're limiting ourselves … We can't define high standards as just reading and math, and we can't define success as just a higher score on a math and reading test. (District 304)

The implication is that test scores would become a more useful indicator if expanded to include more subject/content areas (i.e., "… We can't define high standards as just reading and math…"). At the same time, the criticism extends beyond test scores and onto the evaluation framework in general. The participant exclaims "…We just don't

use everything that's available to us. We limit ourselves …" In this way the evaluation system is perceived as adaptable, flexible, and capable of including currently unmeasured attributes and measures. Finally, like principals, this district participant (above) acknowledges positive aspects of the evaluation system ("…We have some foundational pieces…"). However, unlike principals, the system's shortcomings are interpreted as substantively harming the inferential integrity of the outcomes.

The third district participant continues the perspective that the system excludes important aspects of teaching: "…There are more variables in determining a student's learning than can be measured with our current system in Arizona…" However, the emphasis is on missing affective impacts of instruction:

> We do not have a teacher evaluation system that addresses the other part of the purpose of education, which talks about a child's experience, on a daily basis, coming to school. The teacher evaluation system may say certain cultural or environmental parts of the classroom experience, but it really doesn't get into evaluating whether or not a student has a good experience that day. There are more variables in determining a student's learning than can be measured with our current system in Arizona. (District 302)

The participant asserts that the affective impacts on students remain unmeasured and that these impacts directly affect learning outcomes. The view is inclusive of both the FFT and test scores, that both fail to account for "… a child's experience … coming to school..."

Finally, an extended narrative offered by a fourth district member (below) encapsulates the concerns raised by many stakeholders interviewed for this study. It highlights the disconnect between policy and practice, the complexity of measuring instructional practice, and the exclusion of attributes believed important to understanding instructional efficacy. As with many stakeholders, this district member ends his/her

comments in general support for the district's evaluation model saying "… I think that the [evaluation] model that we [the District] developed is as fair as can possibly be …" Arguably, the statement reveals the conflict existing between personal beliefs and the realities/compromises of implementing an organizational evaluation system.

> I think that teachers are a different breed of people than say, factory workers or those that build widgets or business-minded folks as well. Because I think that educators in general operate a lot on the feeling side, a lot on the emotional side, as opposed to sometimes the black and white side.
>
> I think that when you impose a system like the state did, that includes achievement as a part of the evaluation and at least a third of it, I think that you change teachers' mindsets in the way that they approach a school year. I think that they then focus solely on the outcome of one assessment, as opposed to the span of an entire school year with every curricular approach. I think that when you try to quantify 180 days of instruction to identify what a good teacher is based upon three days of an assessment … I think that's when you start to have a problem. You have people making legislative decisions that have never operated in the field of education. Legislators have really caved to trying to quantify something in a box, that's very difficult to quantify.
>
> I would say that we [the District] took the approach to attempt to minimize the achievement piece as much as possible, and yet it still accounts for a third of the evaluation. Truthfully, I think that that in general, has just freaked teachers out. Again, you're talking about folks that live in the gray; they live in a gray world and [its] forcing them to be black and white.
>
> I think Danielson's model is probably one of the better ones that I've seen, but it's certainly not perfect. For example, it doesn't capture a teacher's ability to truly motivate their kids or to make learning come alive for their kids. While it's one of the better models, it still narrows the scope of what people are looking for.
>
> I think that the [evaluation] model that we [the District] developed is as fair as can possibly be and I think that that's important. I think we've taken a really good approach. I know of other districts, what they're doing, and I think that our approach is a bit more sound. I think it's a bit more defendable down the road, and the reason I say that is it's consistent and it's fair. (District 303)

The perspectives expressed by this individual are nuanced, expansive, and revealing. The narrative seems to indicate the following:

- "… teachers are a different breed of people…": Teachers are not simply workers within a context of production ("…factory workers…widgets…") [Freire, 1970]. Their activity is not economic ("…business minded…"). Teaching is not mechanistic, highly structured, compartmentalized, or standardized ("…black or white…"). To be a teacher requires emotion ("…on the feeling side…") and emotion cannot be detached from instructional practice.
- "… when you impose a system like the state did…": Imposition and acceptance are not the same. Imposition requires subjugation, compliance, and an inability to resist. In the absence of a state-imposed framework, educators would have designed the system differently.
- "… people making legislative decisions that have never operated in the field of education…": Policy makers from outside the realm of education are uniformed and ignorant to education process and actions. The power-centered decision making process is influenced by non-educational factors and interests, resulting in poorly articulated policy that has substantive negative consequence on the profession and students.
- "…when you try to quantify 180 days of instruction [into] … three days of an assessment…": Instructional quality cannot be reduced, quantified, or simplified. Test scores are a poor proxy for instructional efficacy. Policy makers "…caved to trying to quantify something…" that cannot be quantified. Inclusion of test scores is convenient, simplistic, expedient, but inappropriate, incomplete, and insufficient.

- "… change teachers' mindsets…": Inclusion of test scores has a negative, unintended, impact on instruction. The consequence is to narrow, restrict, and/or limit instructional practice and student learning. It "… freaks teachers out," harming learning environments. This is atypical for teachers, "…they live in a gray world and [its] forcing them to be black and white…"
- The Danielson model is as "… fair as can possibly be…" It is a compromise, a "… good approach…", a place to start. It is imperfect because it lacks the affective (i.e., "…teacher's ability to truly motivate or to make learning come alive …"). As such, it "…narrows the scope…" of instruction. It is "… more defendable…" than other approaches – which are more imperfect. Credibility, in part, comes from the model's "consistent" application rather than its completeness.

In summary, district participants, unlike principals, generally held negative views on the suitability of the evaluation system to identify good/effective teachers. However, the sentiment is not an indictment of the district's system. Rather, the district group members see the organization's approach to evaluation as being restrained by state-imposed conditions. More specifically, district stakeholders believed the effect of emphasizing standardized test scores is negative and adversely impacts teacher behavior by restricting/limiting instructional adaptability. District participants also reflected on missing affective measures. Finally, district participants were generally supportive of efforts to design/implement the evaluation system, but there remains room to improve over time.

General efficacy - state. State participants necessarily did not have intimate knowledge of the district's evaluation system. However, they did express substantive perspectives on the legislative-imposed framework under which districts are expected to develop their systems. Overall, the sentiment was negative, citing poor implementation strategies at the district-level, the inherent difficulties of assessing teacher performance, the lack of a consistent definition of instructional quality, and the limitations of using AIMS as a primary metric of effectiveness.

In this regard, one state participant cited numerous concerns on the current evaluation framework.

> I think everybody has relatively simple models in mind in education, and I view those models have traditionally been, for the most part, fairly destructive. … The method by which you do it is critically important. … The worst managed ones, they're having the manager do all the evaluations, and if just my manager is evaluating me, now all of a sudden, it changes my relationship with my peers. To a certain extent, it creates negative interdependence, so it's unidirectional. I'm very concerned that almost all of our schools in Arizona have gone to unidirectional evaluation of teachers. It's overloading principals. (State 401)

Interviewer: Will it [the state policy framework for teacher evaluation] improve teaching?

State 401: I think it's going to have some negative effects, and I don't know how much, so my sense is that very few [districts] are getting it right, and it's taking teachers' energy away, it's discouraging them a little bit. It's leaving them a little less energized and a little bit less positive for their students. To me, it's a tragedy,

Accountability can actually damage a system because it's the old thing [of] what gets measured gets done. Well, if you have more than half of your pie not measured, it doesn't get done, and it actually will, the focus effects of focusing on what's measured relative to what's not measured, and what's not measured, kindergarten through 3rd grade, is enormously important.

The participant criticizes the supervisor-as-evaluator model of evaluation (the alignment currently used by the district). Here, the teacher's supervisor (principal) is also the evaluator of record. The individual (State 401) argues that this alignment harms the leader-mentoring relationship expected of school administrators, creating "… negative interdependencies…" and negatively impacting the school's collaborative culture. In addition, the participant sites omission of early-year (pre-k through Grade 2) academic outcome measures, seeing these as being "…enormously important…" Finally, he/she views the evaluation process as "… overloading principals…" due to the excessive time needed to evaluate all staff on a campus. This state member concludes that Arizona's evaluation system is going to have "… some negative effects…" and that "… accountability can actually damage a system…"

A second state participant focuses on a different set of concerns. First, the state framework is too broad and fails to provide for a clear definition of effective teaching: this results in different implementations (structures, metrics, and performance criteria) across districts, concluding "…we probably do have 220 different definitions …" Second, the participant criticizes not having direct measures of student academic outcomes for "… 70 percent of …" teachers (these are known as Group B teachers in the Arizona evaluation framework). The implication is that without an operational definition of good/effective teaching along with the lack of measures for Group B teachers, the system's credibility is circumspect. The participant reflects:

> I don't think the framework does it at all because the framework simply says that you need to have 33 percent of your evaluation based on quantitative data, and the other 67 and 50 percent based on performance, and the other 17 is up there as whatever. All it says is performance and test scores, and it's up to the school districts to determine [the details].

You [districts] have to come up with, and figure out, how you're going to use student quantitative data for that 70 percent of your [Group B] faculty. The [state evaluation] document says that if you don't have [classroom-specific] AIMS [test scores], then you're going to be evaluated on everybody else's AIMS scores. That's not sitting well today.

[In addition] the statute doesn't say that the districts have to define what good teaching is. They have to define what an incompetent teacher is, which again, goes to a mindset here on their [legislators] part. … If you want to throw in the 580 charter schools that we have, we probably do have 220 different definitions [of good/effective teaching]. (State 402)

Still a third state- participant questions the fundamental ability of the evaluation framework to measure teacher effectiveness.

Interviewer: How well does Arizona's evaluation framework position us to measure, identify, and perhaps more importantly distinguish on a continuum good/effective teaching?

State 403: Not well. I don't think we have a robust system as of right now just because I don't think we have a good assessment. The AIMS test is not, in my opinion, as effective an instrument that can tell with certainty how well a teacher is doing, because it's very haphazard.

I'm saying we don't, because we don't have a system that is able to gauge teacher effectiveness. I still think we have a ways to go, and what's hurting us is that … we're in transition right now. We're in transition between standards. We're going to be in transition between assessments.

Then added to that is that we have - and I think we're still struggling with this, I certainly know that schools are struggling with this - we have no way to determine whether or not the [Group A] A teachers are able to show performance versus [Group B] teachers that are not in the AIMS test.

[In addition] There are some school districts that I think do a much better job of it [developing an evaluation system] and have a better system of doing that. The state has imposed this criteria that we need to have a system, and the districts or charters need to start to do this. Some are very, very good, and some are very, very not good. It's all over the place …

Interviewer: Because there is no common definition to how that [evaluation system] gets implemented, you're saying that we need a better, more common system, more reliable and consistently implemented across all districts?

State 403: Yes.

This participant (above) responds directly to the question of system efficacy by stating "… Not well. I don't think we have a robust system as of right now …" The perspective is primarily based on transitions occurring in the state's adopted curriculum (standards) and the state's curriculum-aligned standardized test (AIMS). The implication is that metrics based on these resources are outdated, resulting in questionable inferences on teacher effectiveness.

Interestingly, this state participant notes the lack of assessment information directly attributable to Group B teachers citing "… we have no way to determine whether or not the [Group] A teachers are able to show performance versus [the Group B] teachers that are not in the AIMS test…" In addition, implementation variability across districts is a concern stating "…There are some school districts that I think do a much better job of it [developing an evaluation system] and have a better system of doing that…", implying that evaluation outcomes in some districts are more appropriate than others. Arguably, cross-district variability is due to the lack of a common construct definition of teacher effectiveness at the state policy level. Based on these concerns, the participant concludes "…we don't have a system that is able to gauge teacher effectiveness…"

Overall, state-level participants were generally negative with regard to the policy-directed evaluation framework's ability to identify, and distinguish between,

good/effective teachers. Some of their concerns corresponded with other stakeholder groups while some perspectives were unique to their membership.

General efficacy (summary). Teacher, district, and state participants provided generally negative reflections on the evaluation system's ability to accurately assess TIQ. In contrast, the principal group was more positive and held the sentiment that the evaluation system produced adequate representations. They based this perspective on the quality of training they (and teachers) received, clarity and specificity of the Danielson framework, and the use of objective evidence to assess instructional quality. Principals also believed that these factors lead to high levels of trust and acceptance in the evaluation results among teachers.

That said, perceptual distinctions do exist between group memberships. Teachers are more likely to include concerns over rater reliability and attempts to reduce teaching to a fixed set of measurable activities. State participants cite the large proportion of teachers lacking direct measures of the instructed content and the lack of early-grade indicators. District and state participants share their reticence regarding the lack of a common definition of instructional quality.

Both teacher and district participants comment on missing/unmeasured attributes of quality teaching as well as agreement that test scores were not good measures of instructional impact. District narratives convey a sense of ongoing improvement (i.e., getting better, doing the best we can) in spite of the state-policy imposed constraints. Finally, state participants reflect that the quality of evaluation systems vary across Arizona districts and that the single mentor-evaluator model could be detrimental to school climate and culture.

Component 2 - Time, Frequency, and Observation (Lack of…). Concerns related to *Time, Frequency, and Observation* (TFO) were raised by all four stakeholder groups, albeit, less so by most principals. Teachers repeatedly emphasized the need for more frequent dialog/communication with their evaluator and believed evaluators did not spend enough time in their classrooms, leading to inaccurate assessments of instructional quality. Principals reflected on time/frequency as being important for proper evaluation but did not raise it as an immediate threat to the integrity of evaluation results. Indeed, most principals felt that they were spending sufficient time to properly evaluate teachers. Both district and state participants expressed substantive concerns on the lack of time/frequency spent evaluating individual teachers. Finally, state participants noted the large amount of time required to evaluate building staff overall, resulting in the inability to spend sufficient time in any single classroom.

Interestingly, for some teachers (three out of the seven) and state (one out of three) participants, insufficient time/frequency was linked to a secondary concern: *Gaming the System*. The view implies that when the number/duration of evaluations is limited, it is easier for teachers to mask their day-to-day instructional practices. They can do this by preparing for the "official" observation event, showcasing instructional behaviors aligned to the evaluation rubric just for the period of observation and then reverting back to previous, non-aligned, practices. Arguably, this could lead to inflated ratings and an inaccurate assessment of instructional quality. Of note, principals and district-level participants did not offer this as a core concern.

Time, Frequency, and Observation - teachers. A general sentiment raised by teachers is that evaluators do not spend enough time *getting to know* them. As will be

developed throughout this section, this is a phrase of relationship: that to accurately assess, evaluators need to develop an intimate understanding of the purpose and motivations underlying particular instructional approaches. Embedded is an implied value ascribed with collegial dialog, critical reflection, and communication between teacher and evaluator. By implication, current observation protocols do not provide sufficient opportunity (Time/Frequency) to acquire this depth of understanding. The concern is expressed by multiple teachers:

> They're only seeing, let's say, 10 to 15 minutes, if even that. I don't think they were in here but five minutes when they were in here the last time. In five minutes, what can you really see? (Teacher 105)
>
> I could never do admins job because … they just don't have the time … I don't know, how often can they get into every classroom every day? (Teacher 106)
>
> There's a lot of pressure put onto the principals. The principals are coming in a couple times for short observations and then they have to fill them out on the computer. (Teacher 101)
>
> When you're being evaluated, and administration's only been in a room two to three times, and then they're basing all of your whole year, your evaluation, off of these one or two visits, and then that's it, it's kind of' nerve-racking. Being a veteran teacher, I don't get a lot of administrative visits and my evaluation was fine. (Teacher 102)
>
> Interviewer: You've come back to that idea here, that the Danielson components pretty well capture the activities of a quality teaching. If you score well then you're getting a fairly good representation of that?
>
> Teacher 102: Right.
>
> Interviewer: However, the frequency with which teachers are observed is limited and therefore may be biasing or distorting the end result?
>
> Teacher 102: Yes, absolutely.

In this last exemplar (above), the teacher (102) acknowledges the representation of the FFT components but questions the frequency of observations. This displays the tension between seeing the FFT as representing qualities of effective instruction and the structural limitations imposed by the implementation process. Importantly, the narrative implies that things are missed: key aspects of quality teaching as well as the impacts these activities have on students. A fifth teacher (below) reflects on this by stating:

> It's all about data, data, data, data, data. But yet they're [Principals/Evaluators] not in the room with me when I'm having those amazing moments when kids are getting it. They are smiling and glowing, and they're going to remember that. They may not show up on the data, but yet it's going to show up from that moment on. I wish more of that was seen.
>
> I feel like administrators are a little bit, they're kind of, removed from the classroom. I feel like they should be in the classrooms more. Not just with me, but with students, just so that they can get an idea of not just no news is good news. (Teacher 104)

The comment above links concerns of time/frequency with unobserved/unmeasured affective impacts of instruction (i.e., "…They may not show up on the data…" and "… administrators are …removed from the classroom…" By implication, more time needs to be allocated to observing and understanding instructional activities and the long-term impacts they have on students. This perspective, that extended/frequent observation is necessary, is also connected to a belief that instructional efficacy is a complex, multi-faceted, construct, if for no other reason that it involves human behavior. The same teacher goes on to express (below) concern that the evaluation system doesn't fully encapsulate "… the whole teacher..." due, in part, to limited observation time:

> What I think it [the evaluation system] should be is to really look at the whole teacher. But what I think it does is [that] it doesn't. If you just look at something, and go by data [test scores] or quick observations, there's no way you can possibly

> get the whole picture of a person. We are human beings dealing with small human beings. Now quick little GPSs that the evaluators do, oh my goodness, it's hit and miss and because we're alive and the classroom is alive. It's either 100 percent on task or a little mixed up in transition. The evaluator only gets a little glimpse. I see that as a problem. (Teacher 104)

Included here is the idea that limited observation/time raises the possibility that what is observed is not indicative of day-to-day classroom experiences or instruction. The comment "…The evaluator only gets a little glimpse…" implies that teaching is an expansive, dynamic, endeavor that changes day-to-day and by situational context. Limited observation time in the classroom creates a "…hit and miss…" approach to data collection. This person concludes his/her remarks by stating "…I see that as a problem…"

Teacher's sentiments within the *Time, Frequency, and Observation* concept are closely linked to the following perspective: accurately assessing instructional quality requires ongoing communication and dialog between evaluators and teachers. It places substantive importance on getting to know, of understanding, all of which seem central to the teacher's perspective. A sixth teacher expresses:

> I remember when I started teaching and I was at a different district, but even when I started teaching for [name], I felt like I had a lot more time with my administrators, … I felt like I had more time with them, whereas nowadays, I just don't feel like there's that part anymore.
>
> Interviewer: All right, I am hearing concepts of time (lack thereof), relationships, familiarity … that the relationship with the evaluator is important to you?
>
> Teacher 108: Yes. Because that way they at least know that if they came in, and it was five minutes of just craziness, that they knew, "Okay. That's not how she really is."

Here, the concern is that the "new" evaluation approach has led to a decrease in time spent with administrators, "… I just don't feel like there's that part anymore …" The

view is that the relationship, the communication, the dialog, all provides evaluators additional information from which to assess instructional quality. More relationship time is seen as a way to overcome potential observation bias created through limited time in classrooms.

One of the teachers previously referenced reacted to a question regarding what is missing from the evaluation process, directly responding "… more one-on-one time, face-to-face, behind closed doors, with a teacher and the administrator…" The exchange was as follows:

> Interviewer: How would you articulate what is missing in the capturing of information for evaluating a good teacher?
>
> Teacher 104: More one-on-one time, face-to-face, behind closed doors, with a teacher and the administrator … to find out when a teacher has concerns. And it's got to be safe. It's got to be absolutely, positively, a safe harbor situation, where they can vent, or ask thoughts, or ask from experienced administrators, or even just discuss feelings, concerns and options.
>
> … I've seen my administrator, what, five minutes this whole school year. Because I'm not one of those teachers that's on fire and needs a lot of help, I don't see him. But I want to see him. I deserve to see him as much as everybody else. Just because my grade level has their act together, so we don't get to see a lot of our administrators. I want to see him just as much as everyone else.
>
> … they [administrators] should never be too busy or too far removed to have ten minutes with each one of their teachers on their campus, to bridge that isolation and the communication and the collaboration. So that when a teacher is evaluated, they don't feel like it's a complete, just judgment.

The view throughout the narratives is that relationship, understanding, and time/frequency are linked, each an important and necessary characteristic of an evaluation process. The concern is that not enough time/frequency is made available.

These intricacies of these connections are as follows: To adequately evaluate one needs more information than can be obtained during the formal observation activity; the missing information is related to gaining a deeper or more complete understanding of instructional practices; deep understanding requires the formation of relationships; and fostering relationships requires more time/frequency, whether in the form of more observation time in the classroom or more one-on-one dialog/communication time between teacher and evaluator.

Finally, the comment above adds the importance of safety into the evaluation process: safety in terms of the teacher being able to dialog, vent, question, and/or openly express feelings without risk. By allocating time/frequency to collaborate, there is more trust given to the evaluation results, the scores are not "… just judgment…" (i.e., inaccurate, capricious, or otherwise exclusive of evidence or a deep understanding of the teacher's practices).

Teachers - game the system. Three of the seven teachers interviewed discussed the concern that limited time/frequency provided opportunities for teachers to prepare for the formal evaluation activity. The backdrop to this is, under the district's system, teachers and evaluators pre-arrange the dates, times, and lessons of the formal (full 45-minute) observation. However, it should be noted that the district's adopted implementation protocol requires evaluators to conduct up to five unannounced (approximately five-ten minute) classroom walk-through observations per year. Regardless, two of the teachers commented:

> All my evaluations here, I hafta' do this [Danielson FFT activities] instead of what I usually do, cuz I hafta' do this pony dance for my evaluation to make sure

> I get a good evaluation. When it's, in all actuality, many do something above and beyond what they actually really do on a regular basis. (Teacher 102)
>
> It is very easy for a teacher to just prepare for that one day and do phenomenal stuff that day just to get that principal to say, "Ooh, I like this, I like this, I like this." To score and to look at the rubric, it's very easy for a teacher to look at the rubric and say, "This is what I need to do to be distinguished, but I'm only getting evaluated once, or I'm only getting evaluated twice. Then as soon as that principal's outta there, I can go back and do whatever I want to do because I'm basically done." (Teacher 101)
>
> My one suggestion, or something probably for the eval, is that the eval happens whenever. The eval can happen at any day, because if the teacher knows that at any given moment, one of these formal evaluations could take place, I mean, every day you should be performing at that evaluation level that you have. (Teacher 101)

The first teacher (Teacher 102) is very direct, making a distinction between how he/she performs during the formal observation and day-to-day (non-evaluation) activities. The view is that the evaluation activity is a formality, something that needs to be completed. It is an ancillary event, outside, and external. To get the best rating, teachers perform for the event, a "…pony dance…" Interestingly, his/her view is that for many teachers, the evaluation activity does not reflect true professional practice. And possibly most important, the day-to-day professional practice of teachers is seen as superior to the practices required under the evaluation process.

The second teacher (Teacher 101) also recognized this performance potential, saying "… It is very easy for a teacher to just prepare…" The evaluation is a discrete event, happening infrequently. There is a sense of stoic perseverance: "… as soon as that principal's outta there, I can go back and do whatever I want to do because I'm basically done…"

This idea that the evaluation observation is an infrequent, distinct, event is echoed by a third teacher whereby lack of time prevents evaluators from conducting necessary follow-ups (i.e., checks and verification):

> I think it's a numbers game kind of a thing. I mean to me it's, "Okay, you have this, this, this, and this. Great, okay, thanks." You just sign off on it. All right, I'm done. Whew, I'm done for the year, kind of thing. … I just don't think there's enough time for the administration to even get around back to it. I just see that they're kind of running around [like] chickens with no heads on and trying to get it done. (Teacher 103)

In summary, lack of time/frequency is a concern shared by many teachers. It is seen as impacting the integrity of the evaluation process by allowing important instructional events and impacts to remain unobserved. It limits the ability to foster evaluator-teacher relationships necessary to gain deeper understanding of professional practices. Infrequent observation increases the potential for teachers to *game the system* by performing for the event.

Time, Frequency, and Observation - principals. Principals did not emphasize the topic of time/frequency to the same degree as teachers. As evaluators, they did not discuss it as an immediate concern or as a risk to the integrity of the evaluation results. Rather, they seemed to feel that sufficient time/frequency was being afforded to the process. One principal discussed the issue as an essential piece of the evaluation activity:

> To have an effective evaluation piece in place, there has to be time invested in getting to know the person as a teacher, how they work professionally, how they interact with students, how they interact with peers, how they interact with parents, to get to know them not just within the 50 minute block of instructional time, but it's getting to know them as they work with their peers in a planning group, or a PLC group, or in an SST, or an IEP meeting, to see what kind of conversations they have, to be part of those conversations so that you can best help them in those conversations. (Principal 204)

This principal (above) recognizes the need for sufficient time to develop relationships beyond that invested in the formal observation tasks. While this is consistent with the "getting to know" concept raised by teachers, it is not being discussed as an immediate concern or limitation. Indeed, it seems to be presented as a positive attribute in the current evaluation process, one that is actually taking place. This principal goes on to suggest that teachers also recognize and appreciate the time and feedback afforded by the evaluation system, stating:

> I do appreciate, although it's lots of time [conducting the evaluations], but I know that the teachers appreciate that specific feedback of what is exactly observed. Then what we can do is go ahead and connect it to what we need to in a rubric. (Principal 204)

Given the previously discussed feedback received from teachers, there seems to be a disconnect between the views of the two groups: principals see time/frequency as a positive while teachers see it as a concern.

Another principal also sees the current evaluation process as affording sufficient time/frequency by allowing for multiple, informal, walk-though observations:

> I think the most powerful thing that an administrator can do, or another teacher going in and observing a teacher, is the frequency; is being able to get in more frequently. Sometimes it's not so much the length of needing to be in there, but how frequent.
>
> When I was an instructional coach, being able to go in on a weekly basis and doing those smaller informals, just brought out … that opportunity to have a look at different strategies, and how the teacher is implementing certain methods that are methodologies. Then being able to follow-up with more frequent conferences and having those conversations.
>
> I really like the changes that were made by increasing informals or having those walk through opportunities … those five, ten minute [observations]; then being able to focus only on one formal for the teachers that are not on probation. … then letting the administrator continue with informals and being able to just constantly giving that immediate feedback and having that conversation. (Principal 208)

This principal (above) is direct with his/her statement that "…the most powerful thing that an administrator can do… is being able to get in more frequently… " Presumably it allows for a more comprehensive review of instructional activities over a longer period of time (i.e., a variety of context and classroom conditions) and reduces the chance of missing important instructional practices. Included is the opportunity for frequent dialog and discussions (i.e., "…being able to follow-up with more frequent conferences…"). Interestingly, the remarks are placed into the context of improvement (i.e., "…I really like the changes that were made…"), of getting better at the evaluation process. This last comment might also be interpreted to mean that as the evaluation process gets better, teacher's instructional competency increases (i.e., as a result of being evaluated by a good evaluation system/process). However, this connection is not explicitly made at this point on the narrative.

Finally, these comments suggest that principals recognize the contribution that multiple, informal, walk-through observations provide to arriving at a final assessment of instructional practice. In contrast, teachers seem to discount this aspect of the system. In their narratives, teachers focused almost exclusively on concerns related to the formal, 45-minute, classroom review. This is interesting because over the course of the school year the district's protocol requires administrators to conduct five (approximately 10 minute) informal classroom observations as well as one-to-two (45-minute, full lesson) observations. The question is why teachers fail to perceive these shorter reviews as legitimate contributions.

In contrast, one principal expressed a different perspective, raising concern that not enough time is afforded to the evaluation process:

> It [the Danielson portion] is very well put together and understandable to me, so that I can use it and be confident that what I'm giving as a rating is accurate. … I'm sticking to the rubric. Bang, bang, bang. I hope everyone else is.
>
> …. I will say this though; I will say probably the tough part for me is not having enough time to be in my classrooms to get a better picture. … It takes a lot of work, and I take it home. I don't have enough time to be with my teachers and really be more of a coach to them. That, I think, is probably my biggest beef - which has nothing to do with the rubric, but has everything to do with the time that I have to spend, which I don't think is enough. (Principal 207)

Here, the principal is qualifying the frequency concern by first praising the Danielson framework and its ability to provide objective, accurate, evaluations (i.e.,"…I can use it and be confident that what I'm giving as a rating is accurate. … I'm sticking to the rubric…"). However, later in the narrative are the statements of concern over time/frequency (i.e., "…the tough part for me is not having enough time to be in my classrooms to get a better picture…"). It is unclear whether this person is making a statement on the integrity of the ratings (i.e., without more time, the ratings are inaccurate) or simply, while the current system is sufficient, it could be improved.

Time, Frequency, and Observation - district. The reader is reminded that the district-level cohort (n = 4) included the organization's superintendent, two assistant superintendents, and a director-level position from human resources. Also as a reminder, district members are not responsible for actually evaluating classroom teachers. That is the responsibility of school-level administrators (principals and assistant principals).

Similar to principals, district level participants did not focus on time/frequency to the same degree as teachers. However, it was not altogether absent from the narratives.

Two out of the four individuals briefly discussed time/frequency, one reflecting in a positive frame while the other raised concern over the lack of time spent on discussion and/or critical reflection. The comments from the two district administrators are:

> I would tell my secretaries "Look, the first two hours of the day I'm in classrooms. Don't bother me unless the school's burning down." We trained our communities to know that if they had needs, don't come in the first two hours because I'm not coming back to the office, because I'm doing the more important thing that I need to do and that's I'm in classrooms so that I can have conversations with teachers about the things that we saw and instruction … Evaluation [also] occurred in the hallway when we would just strike up a conversation. I loved summative post-conferences because we got to just sit and talk about everything that we saw and those kinds of things. (District 303)

> To truly evaluate means you need to collect evidence over time. Then you need to have a discussion with the teacher about that evidence, so that they can contribute and explain why something happened or why this was that way or why that. Even reflect on those pieces, to ensure they we're getting the same kind of dialogue and the same kind of reflection from our teachers, between teachers and administrators. That's a big chunk. We're not there yet. I don't think we're there. (District 301)

The first district participant clearly valued the time allocated to discussing/reflecting with those he/she evaluated while serving as a principal. Here, the implication is that enough time is structurally available if one manages the process efficiently. While the person does not provide a direct reflection of current practices, not raising it as a serious concern suggests a view that time/frequency is currently being accorded to evaluation activities (i.e., no one has raised it as a problem to district leadership, and therefore it is not an issue). For the second participant it is unclear whether his/her comments reflect a structural limitation in the system or uncertainty about whether administrators are using time in a productive way. Regardless, the district narratives did not focus on time/frequency to the same degree as teachers.

Time, Frequency, and Observation - state. State participants are the most removed from evaluation activities taking place within any particular district. However, of the three state-level participants interviewed, two raised concerns related to *Time, Frequency, and Observation.* Their narratives reflect a tension between the substantive amounts of time it takes to evaluate all teachers on a campus versus the lack of time available to properly evaluate any individual teacher. Arguably, both concerns reflect on the same perspective, evaluators do not spend enough time in classrooms to get a proper assessment of instructional quality. For one of the interviewees, the time required to evaluate also takes away from other leadership responsibilities. The first state participant reflects:

> It [evaluating teachers] is overloading principals. In my view, the place I want my principal is not in his office doing paperwork, doing evaluations, I want him out in the parking lot in the morning, in the afternoon, greeting parents, talking to them, assessing the situation. Letting them know that they're creating that positive interdependence and welcoming of the parent, knowing that this school belongs to them because the principal's out front greeting people.
>
> … everybody is worried about "You'll know the truth about me, so I wanna set up a 'dog and pony show' so I can be at my best, so you can see me at my best on Tuesday afternoon at 3:30". I just think that's error - that we have to set up these elaborate observation systems. (District 401)

The individual (above) argues that the evaluation process distracts school leaders from more important activities: in this instance, activities aimed at improving school relationships with community/parents. He/she also raises the concern (as reflected in some teacher comments, above) that teachers are able to prepare for the observations (i.e., "…dog and pony show…"). By doing so, the evaluation activity is devalued, implicating the credibility of results. Like the teacher narrative, the context of preparing, or *gaming the system*, remains situated within the concept of limited time and frequency.

The same participant goes on to say:

> I'm probably pretty sure that most … districts are doing, a principal goes in and evaluates, maybe walks through three or four times, because they don't have time to do anything else. I think there's quite a few. (District 401)

In this reflection, the concept of time/frequency is viewed from the campus level. That is, for any single principal, having to single-handedly evaluate all of the teachers on a campus prevents spending sufficient time with any single teacher (i.e., "…they don't have time to do anything else…"). In this way, it is consistent with the teacher's perspective of limited time/frequency. It is also a criticism of the principal-as-evaluator model of evaluation.

The second state participant is more direct in raising concerns regarding limited classroom time/frequency:

> I don't think that the evaluators are sophisticated enough, nor will they ever be, [laughter] to distinguish between good questions in a classroom and questions that truly require the students to think at a higher level. I say this because, first of all, we're not in the classrooms enough to make those distinctions. (District 402)

Again, the view is that evaluators do not spend enough time in classrooms. Because of this, they are unable to observe important "…higher-level …" inquiry. Arguably, embedded in the comments is the view that evaluators are not sufficiently trained (i.e. not "…sophisticated enough…") and therefore require more time in any given classroom to properly assess the instructional quality of the teacher.

Time, Frequency, and Observation – (summary*).* Perspectives on *Time, Frequency, and Observation* (TFO) differed across stakeholder groups, in particular with principals. Generally, teachers and state participants held more negative views while district members did not raise it as being problematic. In contrast, principals were

generally positive, both about the need to allocate time/frequency to evaluation and their ability to do so. Only one out of eight principals discussed a desire to spend more time in classrooms. However, even here, there was no indication that the current amount of time seriously detracted from his/her assessment of instructional quality.

Teachers were most consistent in their perspective that evaluators did not spend enough time in their classrooms. For this group, relationship was an important component in their narratives. Here, relationship is being expressed as the need for more dialog, communication, and critical reflection with their evaluator. It links the importance of getting to know, understanding, and the whole teacher as key components for accurately assessing instructional competency. Teachers were also more explicit in the link between low time/frequency and the increased probability of missing important attributes of effective teaching including affective impacts on students and missed observations of good instructional practices. Imbedded in the teacher discussion is the idea of pedagogical complexity. Finally, teachers (along with some state participants) felt that low time/frequency increased the opportunity to game the system by preparing for the observation events.

Like teachers, members of the state group reflected concerns on insufficient time/frequency for classroom observations. Their perspective was nuanced; expressing concerns over the substantive time required to conduct campus-wide evaluations and not devoting sufficient time in any single classroom. Arguably, the two perspectives are the same, leading to inadequate assessments of instruction quality. Unlike other participant groups, some state members viewed evaluation as detracting from other important

leadership responsibilities resulting in negative impacts on school climate and community relationships.

Principals did not present time/frequency as a core concern. Rather, they recognized it as a necessary part of arriving at an accurate measure of teacher quality. Here, principals specifically acknowledged the importance of time for dialog, communication, and reflection. Principals were also more likely to believe (in contrast to teachers) that their current evaluation activities provided sufficient time/frequency, and that this view was also held by their classroom teachers.

Component 3 - Test scores/measurement. Test scores and measurement issues were a substantive concern for all stakeholder groups relating to system efficacy. The *Test Score/Measurement* category is broadly interpreted as the ability of evaluation metrics to adequately represent Teacher Instructional Quality (TIQ). It is not restricted to statements of scale or statistical reliability. Rather, it also includes broader conceptualizations such as bias, the influence of non-instructional factors, and missing measures. While the component conceptually includes issues specific to the FFT (Danielson) rubric, these aspects are not emphasized in the discussion below because they were previously discussed in detail under Research Question RQ1B(c). For this reason, reflections on *Test Score/Measurement* presented here relies more heavily on stakeholder views of standardized test scores as an appropriate component of the evaluation process. Included are some selected issues specific to the metrics used to quantify instructional performance.

As mentioned at the beginning of the section, the *Test Score/Measurement* narratives were initiated from two additional probes included in the interview protocol.

The main discussion prompt was initially presented as *How well does the district's teacher evaluation identify, and distinguish between, "good/effective teachers"?* Two follow-on probes were explored: (1) *Are test scores a suitable indicator of "good/effective teaching"?* and (2) *Which would you place more confidence in for identifying a "good/effective teacher": evaluator observations or test scores? Why?* In the analysis below, the first probe is referred to as *Test Score Adequacy* (TSA) and the second as *Metric Preference* (MP). Necessarily, each are large, multi-dimensional, concepts. Collectively they examine the suitability of the evaluation metrics to represent instructional quality by deconstructing the perspectives of each participant group.

Test score adequacy. *Test Score Adequacy* (TSA) examines the suitability of test scores to provide an adequate representation of instructional quality. The concept is nuanced because stakeholder narratives discuss this from many different perspectives. For example, test scores may be seen as inadequate because of their inability to assess important affective dimensions (impacts) of instruction; structural misalignments between tested and instructed content; biasing effects from non-instructional influences/factors; or the questionable suitability of inferring instructional competence from short-term, *snapshot* (single day, one time), measures.

Test score adequacy - teachers. Teachers expressed generally negative perspectives regarding the ability of test scores to adequately represent instructional quality. While a few teachers acknowledged test scores as a legitimate component, the overall sentiment was to minimize its weight in the evaluation process. Most focused on problematic issues believed to degrade the inferential integrity of the measures.

For teachers, test scores represent an incomplete and/or biased measure of instructional quality. A teacher comments:

> If you can get kids to come to school, that could be half the battle. Research will say if you can get kids to come to school a certain amount of time, then their success rate, their graduate rate, will go up. I think then, the test scores are probably going to take a backseat to that because they weren't even coming to school before. They were off the grid. Now you've motivated them, inspired them, you got them coming four days out of five or five days out of five. Then, once they buy into that, then you have to go into [getting them] to learn. (Teacher 101)

The implication is that there may be other short-term instructional goals and objectives than those assessed by traditional content-area tests. Here, the teacher argues that the focus may be to first change a student's affective (motivation, inspiration, engagement, etc.) relationship to the learning process. This type of short-term prioritization may (1) not be measured by standardized tests and (2) have a negative impact those types of scores.

Another implication (above) is that instructional quality might best be assessed using longitudinal measures, i.e., as a result of gradual changes in student motivation, near-term test scores may be low but increasing over time. In addition, there is an implied sentiment that measures of effectiveness may differ across students and context: some students may be currently ready for content learning while others require additional interventions and supports. In this regard, a different teacher responds:

> Interviewer: Taken on their own, are test scores a good or suitable measure of what it means to be a good teacher?
>
> Teacher 103: No. There are plenty of factors that affect test scores. I always tell even student teachers, they [students] are not widgets; they're people, so there's not a straight line trajectory [to learning].

A third teacher responds:

Interviewer: What is missing from the test score that doesn't capture what it means to be a good teacher?

Teacher 104: Personality, [and] all those human elements. Test scores are [just] cut and dried, right or wrong.

The first response suggests a lack of trust in test scores as a measure of teacher efficacy. "…There are plenty of factors that affect test scores…" indicts tests scores as being subject to external bias and influence from non-instructional factors. In addition, learning is not a "…straight line trajectory…", suggesting a single test score presents an insufficient reflection of how students are progressing, that is, regardless of instructional competence, students are never at the same point their learning at the same point in time. Thus, the responses above imply that test scores are (1) incomplete in what they measure, (2) biased/impacted by outside, non-content related, factors, and (3) ignore students' developmental pathways. Here again, embedded in the collective narrative is a distinction between short-and-long term measures. A fourth teacher adds to this view of bias and influencing factors:

> There's been just so much change in how a student can show they learned. Now, it seems, a lot of it is that pencil and paper bubble test. You can have the student in the room do a concept, and they do it wonderfully. Then they take one test, and that ultimately decides if they know it or they don't know it. There's no background. You don't know what happened to the kid on the way to school. You don't know if he had a fight with his mom on the way to school. You just have this one piece, whereas I feel like teaching is kind of, well, it's just that blobby kind of thing. It takes up a lot of things. (Teacher 108)

Here, the view is test scores can be influenced by many factors all of which are outside of teacher's control. This same teacher further comments:

> I feel like a good teacher could do all the good things, especially if you have a student who may be identified Special Ed, or you have Gifted students who are

just, honestly, on their own little planet [laughter], and you know they are going to be successful in life, but they're not going to show it on a test.

Interviewer: So, taken on their own, are test scores an adequate indicator of good teaching?

Teacher 108: No. Because I feel like there are, again, too many elements. You don't know [if] the goldfish died last night. You don't know that if the kid doesn't even have sheets on his bed. You don't know if he had heat last night. There's so many things that we get [when students] leave our schools, and go to their homes that are, I would say for most of us, safe, nurturing, loving environments. We have too many students who don't go home to that.

There seems to be frustration in this teacher's voice: "… I feel like a good teacher could do all the good things … but they're not going to show it on a test…" The frustration is partially rooted in student sub-populations such as special education and gifted (saying these students are "…on their own little planet…"), personal experiences ("…the goldfish died last night…"), and home environments ("…too many students who don't go home to that…"). Still another teacher expresses frustration on other influencing factors (i.e., test anxiety, fatigue, lack of motivation), making it difficult for him/her to interpret the test scores:

> Some students may have test anxiety and not test well, so the test score doesn't reflect, I don't know, reflect how a teacher performs. When kids don't do well on a test, I can't tell you why they didn't do well on a test. I can't tell you if he got tired and he just started bubbling C, C, C. I can't control that. (Teacher 102)

Here, the inferential value of the score is being questioned —"… I can't tell you why they didn't do well on a test…" —with the context of it being outside the teacher's control. The implication is that other measures of instructional effect are preferred to the (standardized) test score.

The comments from the five teachers (above) suggest that test scores are unreliable (reflecting random variation, error, and bias not associated with instruction). Arguably, this detracts from the trust teachers place in the data, and the evaluation results overall. A sixth teacher comments:

> The things that my daughter is coming home learning and talking about, just the mental processes that she's going through, amazes me. But I was blown away that [her] benchmark scores went down instead of up because of what I see. Her teacher is very much a documenter so I see all kinds of data from first benchmark to second, and it was like, "What happened here?" (Teacher 106)

In this comment (above), two measures of the student's learning are being compared: the observations made by the parent regarding "… mental processes …", and the data supplied by the benchmark test score. Here, the two sources of information did not agree, creating confusion and raising questions of reliability. Which information is accurate? Which should be assigned more weight? How can the disconfirming *evidence* be resolved? Finally, if multiple sources of evidence do not agree, then how can they effectively contribute to the evaluation of instructional quality?

Some teachers did provide positive reflections regarding test scores. However, most remain qualified by underlying concerns. For example, the same teacher (above) goes on to reflect on the utility of test scores "… in a perfect world…" Under this condition, test scores could be "… part of a good indicator…" However, the qualification remains that "… I think there are other things to be taken into account…" and they are "…just a snapshot in time…" The teacher states:

> I think they [test scores] are one part of a good indicator. I think there are other things to be taken into account.
>
> Obviously, they [test scores] have a place, and in my mind, in a perfect world, they're for me to see, "What did I do well, what do I need to help them [students]

with, what is this kid getting out of it, and why or why not is he not getting this …" So it [test scores] definitely has its place, but I think those tests are just a snapshot in time. (Teacher 106)

A seventh teacher comments specifically on how focusing exclusively on test scores "…seriously narrows the creativity in the classroom…" He/she argues that other types of achievement-related measures might constitute better indicators of learning than standardized test scores.

Interviewer: I'm hearing that focusing on test scores narrows the activities or the focus of the teacher?

Teacher 103: It seriously narrows the creativity in the classroom because you're constantly working on these little quizzes, tests. I do work on them in intervention, I mean you need to. I have seen teachers where that is their entire curriculum and it's a little scary because there's no room for creativity, which is how children learn, at least in my opinion. They create, they do, they're messy, they mess up, they fix it. It feels so robotic when it's number one, A, B, C, D. Number two, A, B, C, D. How boring is that if you're a kid?

Interviewer: Does project-based learning & instruction get at more of those softer skills that you think are important with kids?

Teacher 103: It does, it gets at those 21st Century collaboration type of skills and they also lend themselves very well to students learning on their own. The problem is, it's hard to measure because they're pushing their own learning. I can't have a checklist and go, "Okay, I've taught that. Okay, I've taught that. Okay, I taught that."

Nearly all teacher interviewees express that test scores serve as a poor indicator of instructional quality. They require careful attention to issues of bias, reliability, omitted attributes, and influencing non-instructional factors. At best, they occupy a small portion of a much larger, more complex, construct of professional practice requiring many types of indicators.

Collectively, teacher perspectives seem to differentiate between tested and non-tested aspects of learning. There is the idea of an indirect curriculum (a phrase barrowed from a comment shared by a principal member). Here, indirect curriculum refers to non-content related goals that are seen as foundational in the learning process. It includes activities such as attendance, discipline, and general participation in schooling. Other perspectives focus on affective aspects of learning: student motivation, engagement, emotional well-being, and desire. Finally, there are currently untested content areas such as art, music, health, etc. To the extent these areas remain unmeasured; teachers perceive test scores as inadequate representations of instructional quality.

Test score adequacy - principals. Principals view test scores as an important and necessary part of the evaluation process. With some qualification, many believe the test score component is a suitable measure of instructional quality. One principal (204) states clearly that "… The goal of a good teacher is to help your students obtain mastery of what you're responsible for teaching. We know what we're being measured by, so it's not a surprise…" Similarly, a second principal responds:

> Interviewer: Okay. Are test scores a good indicator of a good teacher.
>
> Principal 206: Yes. Yes, they are. They are an indicator of a good teacher because a good teacher should get good test scores. But 'good' is relative. If a teacher had 100 percent meets, is that good? Well, it sounds really, really, good, except for 50 percent of those kids should've been in exceeds, because they were [there] the year before.

Both principals position achievement as a primary purpose/goal of instruction. Here, test scores convey important information on instructional efficacy, presuming that if test scores are low instruction is ineffective. In this way, test scores become an

accountability measure. That is, if the goal is academic learning, the accountability question becomes "did it occur in your classroom?" A third principal provides an example of this by saying:

> I think that it [test scores] is, it's a one-shot in time over all that stuff. I think that the teachers understand now that they can't just be good planners and know their stuff. It's holding them accountable. They should be held accountable. It's not just "oh, I did my best and they didn't learn it". It's not "I'm a teacher". No, its "I am a tool for learning". That's the biggest outcome and they [teachers] get that now. I think they're starting to realize it's not just "I teach". It's what did they [students] learn. They have to go together.
>
> … My evaluation of what the teacher has done, and has presented, and how they do their instruction, doesn't tell me if the students have really learned. That's the [test score] data. That's the scores. That's where we have to go. (Principal 201)

In the perspective expressed above, the reference "…it's a one-shot in time over all that stuff…" is not being used in the same manner as teachers. For teachers, the concern is that test scores represent *just* a snapshot in time (emphasis on *just*): limiting, restrictive, incomplete, and an inappropriate reflection on instructional effectiveness. For the principal above, the *one-shot* event is used in terms of summation, aggregation, combination, and encapsulation. It is positive and inferentially informative. Here again, the principal believes that teachers are aware of this and value it, stating "… That's the biggest outcome and they [teachers] get that now…" Finally, this principal states emphatically that the Danielson FFT portion of evaluation "…doesn't tell me if the students have really learned…" devaluing the professional practice ratings in favor of the achievement indicators.

The principal above continues to qualify his/her stance on test scores by placing more trust in multiple year measures:

> I believe that it depends on what test scores you use. I'm not a fan of one point in time, one test score. That's just it. That decides if you're good or not. But we don't do that here, which is good. We use three years. (Principal 201)

Here, the "… one point in time, one test score…" refers to a single-year achievement measure. This principal is placing more value on observing a teacher's annual achievement scores over a multi-year period. Presumably, persistently high achievement scores indicate more effective instruction and vice versa. Interestingly, no reflection is made on interpreting inconsistent, or highly variable, longitudinal measures.

Principals were also more likely to make a distinction in the types of test scores used in measuring educator effectiveness, generally favoring growth measures over raw scores. This is a nuanced perspective where growth metrics are perceived as an important representation of instructional quality, presumably because academic growth accounts for students who start the year at varying degrees of prior knowledge. In this regard, a different principal comments:

> Interviewer: Do you think test scores are an adequate measure of good teaching? We may have answered it before, but just so I'm sure about it.
>
> Principal 203: I think the growth component should be. It should be, and I say that because if it's about growth, we should at least show one year's growth for every child with the time that they're spending in our schools with the teachers. I don't think that's too much to ask for.
>
> … we wanna see that growth for the sake of the students. It should be there if the teacher has done everything that they're supposed to do.

However, even this context of *growth* is qualified by a fundamental perspective that content learning (the state's curriculum standards) remains the primary purpose/goal of instruction. This is, in part, due to state-imposed curriculum standards and

accountability. Still another principal articulates this by saying "…This is what kids are supposed to learn…" and then elaborates to make a clear distinction between *direct* and *indirect* curriculum:

> I would say that growth scores are a good indication of how well someone is doing in class. I would say, by and large, that's a good indicator.
>
> … Test scores are a reflection of the direct curriculum. The direct curriculum is the state standards. This is what kids are supposed to learn. The indirect curriculum are all the things like, hey, we wanna teach our kids how to work hard, be on time, help others, those type of - that's the, to me, the hidden curriculum. Those are the things I wanna instill in kids, but it has nothing to do with what the state says "this is what we want kids to learn". Although, there's no one out there who will tell you, "Oh, that's a bad thing. Don't do that. We don't want kids to learn how to work hard. No, no, no." (Principal 207)

For this principal, the direct curriculum seems to be imposed, what the state says a student must learn. Importantly, it is distinct from other softer goals that might be part of the classroom experience: hard work, timeliness, etc. Referring to this indirect curriculum, the principal closes his/her remarks by saying "… no one out there who will tell you, "Oh, that's a bad thing…" suggesting a level of conflict between valuing test scores over other aspects/impacts of schooling. Regardless, the perspective is that test scores are a necessary core component of teacher evaluation that provides suitable, reliable, and inferentially valid information on instructional quality.

The narratives provided by principals did not uniformly suggest that achievement scores alone provided a complete picture of instructional competence. In this regard, participants identified two limitations. First, standardized test scores are an incomplete representation of teacher's professional practice. Second, the subject areas tested by the current state assessment (reading, mathematics, writing, and science) are not aligned to

the curriculum content of many teachers (i.e., social studies, art music, health, etc.). These limitations are represented in the comments from two principal participants:

> A test score does not indicate the level of commitment of the teacher to the school as a whole, which is that professionalism piece … [and] There's that whole human piece that is maybe not able to be seen in just strictly a test score. I think that's a big piece of what teachers are, as well. (Principal 201)

> One of the things that is a concern with AIMS is that, although every teacher should be teaching reading and can incorporate math, I [may be] the Art teacher. I am a highly effective Art teacher, because I've got my state Art things. I've got my kids are growing and learning. You can see what they did at the beginning and the end, and I feel like this Art teacher is highly effective teacher. They're an awesome Art teacher, but [the Art teacher's evaluation is] reliant on what the other teachers do. (Principal 202)

> … The perfect example is the [Group A] teacher that's the outlier, that we had in one grade level last year, where those scores were 30 percentage points below their teammates. Thirty percentage points. That was huge. If you, when you factor that in, that whole grade level, because they're A teachers, the B teachers then are affected by that, those outliers … If the assessment [scores] are outside of the teacher's control it's a little more difficult to justify that. (Principal 202)

The first comment (Principal 201) reinforces the view that test scores exclude affective dimensions of professional practice. The second comment (Principal 202) acknowledges alignment issues between tested content and the content provided by Group B (non-tested subject area) teachers. [Reader's Note: In this district's evaluation framework, teachers instructing content areas not represented on the state's annual achievement test are designated as Group B]. This latter comment raises an issue of attribution bias: if Group B (non-tested) teachers are evaluated based on Group A (core subject) test results, and the Group A scores are (for example) low, the attribution of those scores will adversely, and inappropriately, affect the evaluations of Group B teachers.

Overall, the majority of the principals were supportive of incorporating standardized test scores into the evaluation process. With some qualification, they did not express concern related to reliability or the adverse influences of non-instructional factors. In general, principals viewed inclusion of test scores as necessary and important indicator while at the same time acknowledging some limitations. When criticisms were offered, principals focused on missing affective dimensions of practice and concerns over attributing core-subject scores to non-core teachers.

Test score adequacy - district. All four district participants viewed test scores as poor indicators of instructional quality. The primary concern was one of omission and/or underrepresentation. Teacher professional competency is seen as multi-dimensional and test scores fail to account for this complexity. As expected, the perspectives are nuanced across members. For example, three district participants expressed their own perspectives:

> I think that when you try to quantify 180 days of instruction to identify what a good teacher is based upon three days of an assessment, that you can argue [raise questions about] the validity of the assessment. I think that's when you start to have a problem. (District 303)

> The fact we spend all of this money to individually test every student on just two things and that makes and breaks our entire educational system, not only as a district, as a state, but now as a nation, and that defines us? The system's broken. (District 304)

> [Teaching is about] getting into the lesson planning, the lesson delivery, the evaluation part of student learning, and reflection on student learning. It [the evaluation system] is still, very much, focused on student learning [test scores]. Still, a large part of a teacher's expected job duties are not defined in learning [test scores]. (District 302)

The first individual expresses skepticism that the complex facets of instructional practice may be reduced to "…three days of an assessment…" Attempts to simplify,

condense, or abridge teaching into a test score are fundamentally flawed and result in inappropriate/incomplete representations of quality. The second individual is concerned with scope/coverage, noting that only two subject areas are tested on the state's assessment. This flaw-of-omission is related to curriculum content where the lack of representation leads the person to state "…The system's broken…" Finally, the third individual also views test scores as omitting important aspects of instruction. However, the omission relates to non-academic attributes of good/effective teaching (planning, delivery, reflection, etc.).

A fourth district participant expressed restrained support for including test scores as a portion of the evaluation process, saying:

> My definition of good teachers doesn't incorporate some of the things that the state is incorporating as far as good teaching. … The state's incorporated student achievement scores as part of what measures a teacher as being good or not good. I think that it's a piece of data that you can look at when you're looking at overall teacher performance ... Whether or not the test scores determine for sure that that teacher's a good teacher or not is— there's no evidence for me to support that it definitely says, "This is a good teacher or not a good teacher." I don't think student achievement necessarily gives us that direct information. (District 301)

The individual acknowledges that test scores are "…a piece of data that you can look at…" when assessing "…teacher performance…" However, the qualification of this perspective is substantive, stating firmly "… there's no evidence for me to support that it [test scores] definitely says, "This is a good teacher or not…" The implication is that any type of competency inference made from a test score is highly questionable. In this regard, another district participant (below) continues his/her earlier remarks, offering:

> I have had opportunities to work with teachers who have been in the bottom ten percent of their students' [in test score] performance, [but] who are the most requested teacher on a campus. … An administrator, who has to be accountable for the school's performance, would say that is not a good teacher. [However] If

> you ask the parent, who is on a waiting list to get into that teacher's classroom, perhaps, they would say that is the best teacher at that school. (District 302)

Here, the individual focuses on the unmeasured affective impacts that good/effective teachers have on students. In this case, achievement measures are low, but indicators of student motivation and connection (the *waiting list*) are high. The implication is that these affective measures are important indicators of quality. Indeed, the phrase "… an administrator, who has to be accountable…" implies adherence to an imposed requirement – the pressure to define effectiveness solely in terms of achievement scores. Another district colleague (below) echoes this connection between the state's emphasis on tests cores as an accountability measure and their use in the teacher evaluation system.

> They [two students] are both straight-A kids, [who] love history and would love to be able to go in depth in history. I think [they are] being cheated out of that, and again, those are direct results of the state accountability system. (District 303)

In context, the district participant knows both students personally. Both are good students who score well on standardized tests. However, because the state accountability system is almost exclusively based on achievement measures in reading and mathematics, classroom teachers face pressure to maximize time and effort on these two subjects to the exclusion of other non-tested subjects. Because of this, the person states the students are "… being cheated…" out of important learning. This represents an added consequence of test scores in evaluation, an engineered narrowing of the learning environment

In summary, district members uniformly agreed that test scores acted as poor indicators of instructional competence. The rationale included exclusion of important instructional attributes, learning content and unmeasured impacts on students. In addition,

the heightened emphasis on standardized test scores as a quality measure results in a narrowing of instructional focus to the detriment of students.

Test score adequacy - state. The reader is reminded that the state group (n = 3) is composed of individuals who had involvement in policy decisions impacting the form and implementation of teacher evaluation throughout the state (legislative and/or interpretative rule making). As a group, state participants expressed generally negative views regarding the ability of test scores to adequately represent instructional competence. However, they were more inclined to acknowledge their inclusion as part of a state level system of accountability. Two of the three individuals questioned the representativeness of tests scores to embody all the attributes important for assessing teacher's effectiveness. Another believed that over emphasis on achievement indicators harmed organizational climate and the collaborative culture in schools. One participant noted that different stakeholder groups (parents, educators, citizens, business, and political) held different perspectives on the meaning of effective teaching; thus, the emphasis/credibility placed on test scores varies depending on the audience.

Importantly, the state participants did not suggest that test scores should be removed from the teacher evaluation system. Rather, their comments expressed caution in their application and interpretation. All seemed to feel that test scores were a piece of a larger picture of instructional practice. Regardless, the general feeling was that test scores were incomplete and provided (at best) limited information on instructional competency.

One state member questioned whether emphasizing test scores in an accountability environment resulted in positive academic improvements. Referring to his/her review of existing research, the individual comments:

> My first conversation with Sanders was somewhere between 1993 and '95, so I ran into him early on, and started reading his research papers, and I thought, "Wow, this is manna from heaven. This is the holy grail," but then I started following Tennessee in the NAEP rankings, and they went nowhere. As a matter of fact, they went downhill.
>
> … They did one [evaluation system] in Nashville that was test score oriented - had zero impact. Rand went down to Florida where they did a big one - had zero impact. Rand went up to New York City under Joe Klein, spent $100,000,000 - zero impact.
>
> … The overwhelming problem with all of the teacher accounting ability systems are negative interdependence … As soon as your test score is your [primary] feedback loop … there's 150 years of history of setting up accountability systems around test scores, and there's no success stories. (State 401)

The perspective is one of over emphasis, of making test scores the dominant indicator. This individual reaches two conclusions: accountability systems based primarily on test scores do not successfully result in higher student performance; and emphasizing test scores in an accountability system results in "… negative interdependence…" meaning an erosion in school climate and cooperative collaboration. The inference is that school climate and collaboration are important ingredients for shaping the learning environment and facilitating student learning. For this individual, excessively emphasizing test scores harms the environmental conditions necessary for maximizing academic achievement.

In the following exchange, the participant suggests that additional indicators of teacher effectiveness should be considered:

> Interviewer: Has [state] policy overemphasized test scores? Is it there simply because the public policy environment expects it to be there?
>
> State 401: Yeah. We haven't understood deeply enough the importance of all these more complex things, that you have to have [regarding] test integrity. We've got better than average test

> integrity on AIMS, but there are issues in there. … The other way of dealing with that … would be to have parent ratings, teacher ratings, and student ratings, and to key off them as much of quality measures.

The type of ratings mentioned suggest that various forms of satisfaction measures should be added to the matrix of indices in order to improve evaluation accuracy. Inclusive are both client (parent, student) and peer-review measures. In other portions of the narrative, this participant discussed examples of districts where these types of multiple feedback sources are utilized.

A second state policy member reflected concern on the increasing emphasis being afforded test scores. Here the perspective is one of exclusion, that is, excluding or de-valuing other important attributes of quality teaching in favor of achievement measures. The individual comments:

> We being so pushed toward test scores that we're not looking as carefully as we should at some of those other components, which might have as much emphasis on what a kid learns than the instruction being given by the teacher. (State 402)

Importantly, this person does not suggest eliminating test scores from the evaluation process. Rather, the context is that effective teaching encompasses more than achievement. Thus, the argument is that test scores are an incomplete measure of instructional effectiveness. The same person continues:

Interviewer: On their own, are test scores adequate for identifying, and distinguishing between, good effective teachers?

State 402: I'm assuming when you say “on their own,” you mean absent any other information?

Interviewer: Yes.

State 402: No.

Interviewer: Because of these other dimensions of what it means to be a good teacher?

State 402: Yes. … Because I think [teachers] professional practice encompasses many more behaviors and interactions with the people that a teacher interacts with, including students and parents and peers and community and supervisors, than the test scores do.

Don't misunderstand me. [While] I'm not a proponent of testing, I believe that tests have a place in our system and would not advocate doing away with any of the testing that we're doing.

The reason I said I would not use testing as the sole basis of determining the teacher is because I think that the testing can vary from group to group, and individual to individual, on any given day, week, content, what have you. To say that the test on these three days during the school year, or these ten days during the school year, aggregated to some score for a teacher, I just don't think is representative of that teacher.

Although, I don't want to imply that I would not have test scores. I think testing is important. If they drastically don't agree year after year after year, then I think it's incumbent upon the evaluator to begin to ask some questions. That's why I'm saying test scores are important not only for the kid, but for the teacher.

The narrative places value on test scores as a portion of a larger evaluation context. This larger context includes "…many more behaviors and interactions…" not captured in the achievement measures. The view is of complexity and the need for multiple measures. Test scores are also seen as highly variable ("…testing can vary from group to group, and individual to individual, on any given day, week, content, what have you…") implying that, on their own test scores may be unreliable.

Finally, for this state participant (above), increased credibility is afforded test scores in the context of time. That is, if "…year after year after year…" achievement

measures are consistently poor, additional value should be placed on this measure to the exclusion of others. In this context, if the alignment between professional practice ratings and achievement measures diverge over time (i.e., high observational ratings coupled with consistently low achievement scores), "…its incumbent upon the evaluator to begin to ask some questions…" [bold emphasis added]. The implication is that observational ratings of instructional quality should inherently correlated with achievement results over the long run and, if they do not, the reliability of the observational ratings should be questioned.

A third state participant also views the contribution of test scores in a larger context. Here, the argument is that the inferential emphasis placed on test scores is dependent on the audience:

Interviewer: Can someone be a good/effective teacher and have low/poor test scores?

State 403: I think it depends on what you're evaluating and how you're evaluating. If you're evaluating a state evaluation of that teacher in the context of the academic standards that we require, and if that student, not just the raw scores, but if that student is making gains, or in this context not making gains, and also the student is at very low academic levels, then the answer would be no.

However, again, that principal, or lead teacher for that matter, that's evaluating at a local level might find that teacher has been able to engage the students in other ways, that they can say, "Okay, we may not be getting those kids to meet academic achievement levels that we want to get to, but quite frankly they're coming to school every day, whereas last year those kids weren't coming to school every day. There's different levels of success and different levels of effectiveness that we need to be looking at.

A distinction is made based on perspective. The state may have a perspective that all students need to learn the Arizona curriculum standards, and it is the responsibility of classroom teachers to ensure that happens. From this perspective, achievement measures become the most important measure of teacher efficacy. In contrast, school officials may be more concerned with student engagement ("…they're coming to school every day…"). Here, the measure of efficacy is based on student behavior: motivation, participation, engagement, etc. It recognizes that the context of efficacy is relative, not well-defined, and lacks broad operational agreement across stakeholder communities.

At the same time, this individual asserts that the dominant perspective shaping Arizona's teacher evaluation framework originates from non-educators. The individual reflects:

> Well, I can tell you that the achievement measure always trumps, at least at the state level. Let's be honest. The achievement measure will trump because the achievement measure is the data that parents, that community members and legislators look at. When they pick up The [Arizona] Republic [newspaper] and it shows the scores of the kids, when The Republic shows the scores of every school district in Maricopa County, that's the first thing they look at. … They [the community] don't care a whit about the teacher effectiveness levels. (State 403)

This is a very emphatic statement both about the perspectives held and the influence wielded by non-educators saying "…They [the community] don't care a whit about the teacher effectiveness levels…" Perhaps, this is not a statement of adequacy but of realism ("…let's be honest…"). This person is not saying this is the way it should be but rather the way it is. While educators may value a multi-dimensional perspective of teacher efficacy, non-educators do not. Embedded is a political perspective - that the state is equivalent to the legislature, and that public perspective is aligned to how the state-policy directed evaluation framework was structured. The phrase "…achievement

measure always trumps…" is also a statement of power, no matter what educators think, it is irrelevant in the political/social context.

Interestingly, the ongoing dialog (below) reveals that this individual's personal perspective is nuanced and reflective, less parochial than the realities he/she described above. The participant reacts:

> Interviewer: On their own, are test scores a sufficient or adequate measure of what you believe a good/effective teacher is?
>
> State 403: No. Well, I think it is a tool. It's a diagnostic that allows us to understand, just like when a doctor takes a blood sample of a patient. That blood sample will give you pretty good information on that individual. A test in a particular subject matter can give us a good indication of whether or not those kids are getting the skills that they need, and making the academic growth that demonstrates that teacher knows how to convey that knowledge to children … That data is again a diagnostic.

There is a qualified sense of importance. Test scores are diagnostic, meant to inform rather than indict. Test scores permit inquiry. Achievement is a starting place from which to investigate and question. In his/her continuing dialog (below), test scores rise in importance if they reveal consistency (trends) over time. However, the qualification continues to be dependent on context:

> The way I would use a teacher evaluation, and maybe give it [test scores] more weight from the state requirements, is if that school were underperforming - a C or D or probably an F school - because something's wrong in that school. We have enough [achievement] data points that show something's wrong in that school. I would say that the way they're evaluating teachers is not very good because the entire school is not performing or making academic gains. Then I would say we need to put more weight on that teacher evaluation for those core [academic] areas because something is not right, if that makes sense. (State 403)

Here, the context becomes the Arizona accountability system which annually assigns schools a label of A, B, C, or D. Label assignments are almost entirely based on

academic achievement measures (passing rates and growth measures). For this participant, if the school displays (academically) poor performance, then the test score becomes an indicator of low teacher effectiveness. In this context, achievement becomes the priority, to the exclusion of other non-academic interests.

There is an important embedded perspective, one that is substantively different from teacher, district, and some state colleagues. The presumption is that, at some point, achievement must become the preeminent concern. That if achievement is unacceptably low, something is wrong with instruction. The orders of importance become first fix achievement, and then attend to the other aspects of schooling. In contrast, others argue that to increase achievement, one must first deal with those other aspects (motivation, engagement, climate/culture, affective components of professional practice, etc.).

Throughout this member's discussion, there seemed a sense of conflict between supporting the state-policy perspective of evaluation (test scores, accountability, honoring public perception) and his/her personal sentiments of what constitutes a good/effective teacher. This is evident in the following exchange:

Interviewer:	Okay. Again, if these two things [measures of professional practice and test scores] don't match, which one would you put more stock in?
State 403:	Okay, I'm going to tell you, I'll answer your question. I'll be totally honest with you. As a parent, I put more stock in my daughter having exposure to those teachers that are the transformative teachers, even if it were at the expense of her academic success [test scores], because to me, I would value that more. I think she would be a much better person because of that. Now from the state policy [perspective], as a member of the [position], I have to say that we have the responsibility to show academic attainment against those core academic areas that it's important for us to achieve.

The narrative implies two competing ideas of teacher efficacy. One is political/policy driven; the other personal. The first is tangible, measurable, and determinant. It is easily tabulated, reported, compared. The second is intangible, affective, subjective, and unmeasurable. He/she speaks of transformative teachers imparting something other than content knowledge saying "…I would value that more…" Here, the importance of schooling is the development of the individual as a person, not solely the acquisition of academic learning ("…she would be a much better person … even if it were at the expense of her academic success…").

While state members value inclusion of test scores as a portion of the efficacy construct, they nevertheless expressed substantive concerns over their suitability to represent instructional quality. First, accountability systems that place emphasis on test scores negatively impact school culture and collaboration and fail to invigorate student learning. Second, test scores as a primary measure exclude important affective attributes of teaching. Third, test scores fail to capture a critical aspect of the purpose of education, to develop the individual as a person.

Metric preference (MP). Arguably, the state-policy imposed evaluation framework operationalizes the presumption that measures of professional practice and test scores are aligned (PP $\Leftrightarrow$ TS): high/low relative values on one are associated with high/low values on the other. That is, each contributes similar information to the TIQ construct. To the degree that stakeholders believe this to be true, the construct proposed by the evaluation system is supported. Disagreement with this premise raises question on the construct's integrity and the inferential value of the evaluation outcomes.

Metric Preference (MP) examines the relative importance placed on test scores and professional practice ratings to inform on good/effective teaching. The motivation for addressing this was twofold. First, published studies suggest that the alignment (correlation) between test scores and other forms of non-academic professional practice ratings are generally weak-to-moderate **(**Milanowski, 2004, 2011; MET Project, 2010; Amrein-Beardsley, 2008; Berliner, 2014**)**. Second, examination of the measures employed in this study indicates similarly weak correlations ($r = .254$, $n = 238$, $p < .001$). This suggests that when principal and teachers review the "live" evaluation results in a high stakes, consequential, setting, they will be faced with contradictory information: one measure suggesting high and the other low instructional competency. The question becomes which measure they place more trust in and value more/less as a reflection of teacher quality?

To examine this, participants were asked to address the following hypothetical condition: if measures of professional practice (PP) and test scores (TS) lead to different interpretations of teacher efficacy, which metric provides a more accurate representation? The generalized interview protocol framing this question was structured using the following hypothetical scenario: *Consider a context in which the test score measure and the (Danielson) professional practice measure did not agree*, *Which would you place more confidence in for identifying a "good/effective teacher": evaluator observations or test scores? Why?* It was hoped that stakeholder responses would provide a deeper, more nuanced, perspective on evaluation and the components used to judge instructional efficacy.

In some of the discussions, the question of preference was preceded by a question of alignment: for example, *Would you assume that if a teacher receives a high score on the Danielson [PP] components, they should also exhibit high growth scores on the achievement test?*, *Do you assume, or expect, that if a teacher were to get very high ratings on the Danielson elements that test score growth should also be high?,* or *Can you be a good teacher and have low test scores?* Often, this provided a basis from which to further examine the question of preference and the relative importance of the two evaluation metrics. Stakeholder views of *Metric Preference* (MP) are examined below as the final component of Research Question RQ1B(d).

Metric preference – teachers**.** Teachers predominantly viewed professional practice measures as being more suitable for representing instructional quality. However, teachers reflected mixed perspectives on whether, ideally, the two measures should be aligned. Throughout the narrative, teachers seemed to struggle with conflicting perspectives. On one hand, it makes sense to believe that good/effective teachers elicit higher levels of achievement from their students while at the same time being resistant to accepting test scores as a suitable primary metric for assessing instructional quality. The narrative provided by one teacher (below) reflects this perceptual struggle:

> Interviewer: Would you think it reasonable that if you have higher achievement measures, you should also be seeing higher professional practice rating?
>
> Teacher 103: Yeah, because at that point then you would have more student-led type activities and be more in those distinguished categories. The kids would be leading it. If the kids are leading it, then they're taking accountability and ownership for their learning, which you would think would account for more [academic] growth or higher [test] scores, just because they are more independent and have a buy-in of their own education.

Interviewer: OK, but if Danielson [PP] and test score measures didn't match up, one high and one low, which one would you say is a better indicator of what it means to be a good teacher?

Teacher 103: I think the Danielson evaluation. It's more in depth. It looks, I know it looks at a snapshot for a day, but your evaluator knows you for the year. I mean they know what you're doing and how you're working with your team, or with your kids, or with the community at large. Those test scores are a snapshot of one, I mean are truly a snapshot of one day. That's scary.

The person is placing higher value on the observation measures based, in part, on the relationship it establishes with the evaluator over a long period of time. The test score is seen as an all-or-nothing (snapshot) event that might present an inaccurate or biased perspective of true instructional competence. It values collection of longitudinal information over short-run data. A different teacher shares this perspective:

Interviewer: OK, so if measures of professional practice and test score growth don't align, one is higher than the other, which would you place more confidence in for conveying what it means to be a good/effective teacher?

Teacher 106: In my opinion, the professional practices. Obviously, they [test scores] have place. In a perfect world, they're for me to see "What did I do well, what do I need to help them [students] with, what is this kid getting out of it, and why or why not is he not getting this, …," so it [test scores] definitely has its place, but I think those tests are just a snapshot in time.

What's on my evaluation is what I've done for the school year, … I think she [the administrator] knows me very well, and she is very good at pinpointing exactly what [my] strengths and weaknesses are, and we talk a lot about that, and I think [the principal's] opinion … to me, is more important than just that one snapshot [test score].

The perspective (above) aligns test scores and professional practice ratings "…in a perfect world…" But the person notes that the world is not perfect - test scores

represent "…just a snapshot in time…" and therefore lack some credibility. In contrast, the relationship established with the evaluator ("…knows me very well…") is more highly valued, in part, due to ongoing communication and dialog ("…we talk a lot …" and "… [the principal's] opinion … to me, is more important than just that one snapshot…"). Regardless, test scores are seen as limiting and uncertain.

The next teacher is a less sure of an alignment between test scores and measures of professional practice. Here, there is recognition that good teachers should elicit some level of academic growth from his/her students, but the connection is qualified:

Interviewer: Would you assume that if a teacher receives a high score on the Danielson [PP] components, they should also exhibit high growth scores on the achievement test?

Teacher 102: I don't know that they would have necessarily high [test] scores. They may show growth, but they may not show as much growth, as say, their neighboring class that's doing the same exact thing. Just the way the teacher taught it, maybe the specific group of kids, like I said, so I don't know that.

The qualification placed on test scores is that they are unreliable, being influenced by external factors such as student demographics ("…maybe the specific group of kids…"), and/or class composition. Upon more direct questioning, the teacher concludes that they don't have to align. The teacher responds:

Interviewer: These things don't have to align?

Teacher 102: I don't think so, no.

Interviewer: OK, let's say that they don't align – for example, a high Danielson score and low growth [test] scores, or vice versa. If they don't agree, which one tells you more about the effectiveness or the quality of the teacher? Which one would you put your most trust in?

Teacher 102: The Danielson, only because that one's based off of evidence. Well, because if I were to go in, and I have high scores on my Danielson - obviously I've had my evaluations my walk-throughs or whatever - then I have evidence with student work and student samples of everything, and I can show on the computer all these kids doing whatever they're doing. I can prove what I'm doing as a teacher. When kids don't do well on a test, I can't tell you why they didn't do well on a test. I can't tell you if he got tired and he just started bubbling C, C, C. I can't control that.

Interestingly, this teacher doesn't view the test score as evidence, selecting the FFT measures "…because that one's based off of evidence…" The Danielson ratings are more valued because they capture activities under the control of the teacher whereas test scores do not ("…I can't control that …"). Here, the evidences of learning ("…student work and student samples…") can be reified, supported, and authorized ("…I can prove what I'm doing as a teacher…"). In contrast, test scores cannot. The concept of control assigns credibility to FFT ratings as a preferred measure of instructional quality. This is further expressed by a third teacher:

Interviewer: We get evaluated on two big components, some kind of test score and the Danielson ratings from your evaluator. Which one of those, those are the only two, are better reflective of a good teacher?

Teacher 105: If I had a choice between the AIMS test or Danielson, I would say Danielson because I have control of that. I can say, "Hey, in my classroom, I need to do that. I need to do that. I'm poor in this. I need to do this. I'm great at that." I have control of that. Everybody else, I don't have any control of. If you're gonna be evaluated on somethin' that you can't control, how good is that?

Not all teachers agreed that professional practice rating constituted a better measure of instructional quality than test scores. Yet even for these teachers, conflict between the two measures remained. For example, another teacher responds to the question:

Interviewer: So, you believe that if you are a good and effective teacher, then you should have high test (growth) scores?

Teacher 104: I do.

Interviewer: OK, let's say the professional practice measures and the growth scores didn't align (one is much higher than the other), which measure would you place more confidence in as an indicator of a good teacher? The test growth score or the evaluation score?

Teacher 104: The growth score.

Interviewer: Why?

Teacher 104: It goes against everything I just said. Because we're teachers and we need these kids to grow in education and curriculum and self-skills and mastery of self-skills and then the next layer of what they need to know academically. I'm not here to raise my own kids. I'm here to raise these kids, and I need to educate them with [the] curriculum as well as everything else.

On the surface, this teacher seems to provide conflicting dialog ("…It goes against everything I just said…"). Previously, the teacher focused on the affective elements of learning (engagement, motivation, etc.) believed to be unmeasured by test scores. During that earlier dialog, he/she responded:

Interviewer: On their own, do you think that test scores are a good measure of what it means to be a good teacher?

Teacher 104: No.

Interviewer: Why?

Teacher 104: Because for our situation, I have a lot of kids that can't even read. They can't understand word problems, but yet they're brilliant in mathematics. …

Interviewer: From your view, what is missing from the test score that doesn't capture what it means to be a good teacher?

Teacher 104: Personality. All those human elements. Test scores are cut and dried, right or wrong.

This individual first expresses concern over important omitted elements of teaching in achievement measures. However, later in the conversation he/she concludes that achievement is the more important indicator of effective teaching. An additional portion of the conversation sheds light on this apparent conflict in reasoning. The same individual responds to a direct question regarding whether one can be a good/effective teacher and still have low test scores:

> Interviewer: OK. Can you be a good teacher and have low test scores?
>
> Teacher 104: No, because if you're a good teacher, according to my definition of a good teacher, the [test] scores are going to grow because the students are going to grow. The [student's] confidence level is going to grow, and the [student's] risk taking level is going to grow.

This statement provides additional perspective. The clarification lies in an implied causal pathway: that for test scores to be high, students must first be engaged in the educational process and exhibit necessary foundational personal characteristics (i.e., confidence, risk taking). The concept of causal pathway was also expressed in other stakeholder narratives, that an important, unmeasured attribute of good/effective teaching is the ability to empower students by instilling aspects of trust, self-value, belief, desire, and perseverance. These attributes are associated with motivation, desire, and engagement. Without these, it is reasoned, students will not take advantage of the learning process resulting in lower achievement measures. In this way, good/effective teachers attend to this casual pathway: first develop the foundational attributes necessary to support learning and then facilitate acquisition of new knowledge.

For the individual above, test scores are seen as a final outcome of the learning process and teachers are responsible for getting students to this end outcome ("…Because

we're teachers and we need these kids to grow in education and curriculum …"). However, this is qualified by a causal pathway: "…mastery of self-skills and then the next layer of what they need to know academically…" In this way, instructional responsibility is multi-dimensional: "… to educate them with [the] curriculum as well as everything else…" And missing from the tests are "…All those human elements…"

It is arguable, and left unexplored in this data collection, whether or not the teacher views test scores as a mandated (imposed) requirement of the profession, therefore consenting to their value as a terminal measure of instructional quality despite their limitations.

In summary, when asked to reflect on the relative value given to test scores and ratings of professional practice, the latter seems to be viewed as more representative of instructional quality. Test scores are perceived as limited by their snapshot assessment of learning, susceptibility of non-instructional factors, and omission of affective impacts on students. In contrast, professional practice measures are considered to be based on relationships, composed of longitudinal information, more evidence-centered, and inclusive of the more affective attributes. Finally, compared to test scores, professional practice ratings are believed to be under the control of the teacher.

Metric preference - principals. Uniformly, principals believed that test scores and professional practice ratings should be highly correlated: high values of one are equated with high values of the other. Unlike teachers, principals predominantly value test scores as the more important and legitimate measure of instructional quality. Importantly, there is a sense by principals that if the two measures do not agree, then the evaluator needs to re-think previously assigned professional practice ratings: that the evaluator is somehow

not applying the evaluation rubric properly. Here, test scores are viewed as more objective and reliable; principals are less willing to question their accuracy. Finally, principals are more likely to view test scores as the primary purpose of instruction, elevating academic achievement above other aspects of the learning/instructional process. The following interaction exemplifies this general perspective:

> Interviewer: If you're a good teacher, should you also have high growth scores?
>
> Principal 201: Yes.
>
> Interviewer: Okay. Which would you place more trust in, the professional practice, Danielson, measure, or the test score, if those two things didn't align?
>
> Principal 201: If they are totally different, I would have to go with, ultimately, you'd have to go with the outcome of what the students have learned, which is the test score. My evaluation of what the teacher has done and has presented, and how they do their instruction, doesn't tell me if the students have really learned. That's the data. That's the scores. That's where we have to go.

The individual places exclusive importance on the test score (i.e., as a direct measure of instructional efficacy). It becomes the singular purpose of instruction to have students learn what is tested. The narrative explicitly de-values professional practice ratings as a primary outcome ("… [it] doesn't tell me if the students have really learned…"). It is possible that the person views the professional practice criteria specified under the Danielson framework to be foundational, a requirement for good teaching which in turn, leads to high achievement. However, if the achievement measures are low, then the teacher cannot, at the same time, be effectively executing the Danielson instructional behaviors.

Engaging in the same type of exchange, a different principal shares:

Interviewer: Do you believe that the Danielson [PP] ratings should be highly correlated with test scores? The better the teacher, the higher the test score?

Principal 205: For the most part, yes. If I have a great teacher, then they are fulfilling their obligations in Domains 2 [Classroom Environment] and Domains 3 [Instruction]. Obviously if they're doing that, they're getting Domain 1 [Planning], which is that preparation. If they can prepare well, they're executing it well. Usually those great teachers that go above and beyond are also very professional [Domain 3] in what they do, so yes, I do.

Interviewer: OK. If the two measures don't match up, which one do you believe tells you more about whether this teacher is a good teacher or not?

Principal 205: [Whispering] Oh, God. If the test scores were done accurately and everything was given the right way, I would say the test scores, because Danielson is still, although it shouldn't be subjective, there's still that subjective piece - because I know you really well, and I know how hard you've worked, and I know how much effort you've put in.

There's still that personal part of that job. There's still that intrinsic part with that you built a relationship with that person. You have to separate that at times and I think that's difficult for some, especially the ones you've put in so much effort with and you know how much they've grown. I'd go with test scores over the other one [i.e. FFT scores].

For this principal, test scores and PP ratings are also tightly linked, answering "…For the most part, yes…" to the question of correlation. Interestingly, there is a causal pathway within Danielson framework starting with Domains 1 and 4 (Planning and Professionalism) leading to Domains 2 and 3 (Classroom Environment and Instruction). Indeed, a great teacher is "… fulfilling their obligations…" and this, in turn, leads to more academic learning.

At the same time, professional practice is (again) de-valued as primary measure of quality. It is subjective, stating "…Danielson is still, although it shouldn't be subjective…" Here, application of the rating rubric becomes biased by the relationship between the evaluator and teacher. This relationship causes inflated ratings "…because I know you really well, and I know how hard you've worked…" Eliminating this bias is difficult and therefore test scores become a superior measure of teacher efficacy. Finally, the principal (above) does qualify his/her comments by "…if the test scores were done accurately…" implying the possibility that achievement tests may have some limitation. However, the individual does not elaborate.

A third principal qualifies his/her trust in tests scores by valuing multi-year trends over single year measures. Here, the principal sees single data points as less reliable and subject to variation. Multi-year trends, however, correct for this and provide more accurate representations of teacher efficacy. Embedded is this perspective that evaluator ratings may be subjected to revision while trends in testing information are not. The individual comments:

> I think if you have trend [test score] data on that teacher, if you've got three years of student growth, I think student growth is very telling. Student growth is going to become more telling … we are holding the teachers accountable in various ways, and now [its] tied to the evaluation. I do think student growth is an important indicator of how effective the teacher is.
>
> Interviewer: I'm hearing you question, and reflect on, your Danielson rating in light of the test score. I'm not hearing your narrative question the test score?
>
> Principal 204: Again, I would want to be looking at three years [of test scores]. If it's one year, it could be a fluke; it could be the kids. But if for three years consistently, they have had extremely high percentage of students [showing] growth, and yet their Danielson [is low], I would be questioning myself [the

PP ratings]. I'd have a conversation with the teacher. Am I missing something? Tell me why are your students consistently achieving higher than the other kids at this grade level? Tell me what you're doing. [In contrast] If I'm giving proficient and distinguished [PP ratings], and I've got a 22 percent growth, what's going on? What am I not doing?

Interviewer: What you've just said tends to say to me that you see this relationship between good teaching and high achievement, and when you don't have that relation, something's amiss?

Principal 204: Well, that's the goal of a good teacher to help your students obtain mastery of what you're responsible for teaching. We know what we're being measured by, so it's not a surprise.

The comments bring out an important aspect of the principal group perspective: That if evaluator ratings and achievement measures do not align, then it is the ratings that need to be re-examined and adjusted. There is implicit trust being afforded the standardized test scores as an instructional quality measure that is not being afforded to the reviews of professional practice. After testifying to the superiority of test scores as the primary metric by which to judge teacher effectiveness, another principal shares the following:

Interviewer: Okay. I would infer from what you said that you have a belief, or you have a trust, that if you're a good teacher you probably should have high test scores. By "test scores" I mean growth.

Principal 208: Right. Yes. And I think that's where it can be difficult if you're doing somebody's evaluation early on in the year. Since we have so many [teachers to evaluate], we do have to start all the way [back] in August, so when it comes down to that, sometimes it's like you need time to gather information and things. I think the earlier you're doing evaluations the more challenging it can be, but again, when there is those red flags of a teacher not doing well, you can go back in and definitely provide additional information in there and do another performing rating.

This dialog implies the following: observation-based ratings of teacher professional practice, done in the absence of external achievement measures, may be imperfect. That is, to properly evaluate a teacher, the evaluator needs to be looking at achievement indicators and that these indicators should be shaping the application of the Danielson rubric criteria.

Importantly, the Danielson system of evaluation contains no such requirement or directive. Indeed, it is intended to measure specific attributes (behaviors) of professional practices, not academic outcomes. As a consequence, the extent to which an evaluator considers factors not explicitly embedded in the rating criteria biases the results. The same principal continues:

Interviewer:	If over the course of time, after you've done your walkthroughs and perhaps some of your formals, if you then start to get this evidence that from the outcomes, the test scores, the formatives, that things aren't as shipshape as you'd like them to be; how does that influence or how do you interpret that in the context of the Danielson evaluation that you've been giving those teachers?
Principal 208:	Then it's our responsibility to go back in and request that [additional PP rating observations]; especially, if they don't [already] need two eval's and they only need the one. For us [evaluators] to be able to go in and actually say, "We're requesting another one." So then we can impact the [overall TEval] performance rating. Then by the end of the year, yeah, then they would have a truly, more of a true rating.

When tests score do not align with ratings, this principal is emphatic that "…it's our responsibility to go back in…" and re-evaluate, re-think, the ratings previously assigned – regardless of the evidences collected during the initial rating process. The person concludes that this results in "…more of a true rating…" Implicitly, the view is that it does so specifically because the ratings should align to the test score. That is, the

test score is correct, the ratings may be wrong, so go back in and change the ratings (after more observation). The individual concludes this part of the conversation as follows:

> Interviewer: Interesting. It sounds like you put your trust in whether the teacher's a good teacher or not into the test score, or the growth score, or the outcome measure. And if the two don't agree you'd be more inclined to go back to your observational environment or activity and re-think that?
>
> Principal 208: Through more observations and walkthroughs, yep.
>
> Interviewer: Right, which implicitly says that you're putting trust that the test score measures are, in fact, telling you something that's sort of fixed.
>
> Principal 208: Yep.
>
> Interviewer: It's objective? It's not subject to very much error?
>
> Principal 208: Right.

A different principal summarizes this view by explicitly stating "…I think, ultimately, that test score is, it validates the evaluation…" It exemplifies the importance most principals placed on measures external to their personal evaluation ratings.

> Interviewer: OK, so to clarify, if these two things [PP & TS] did not align, which one are you going to value more, the test score or the Danielson rating?
>
> Principal 207: If they didn't align, and they didn't consistently align with my evaluation, and then looking at the growth scores, I would have to say, well, I would say I'd have to look at my practice. I'd have to say that I probably value that test score more than I would my own rating of how someone is doing because I think, ultimately, that test score is, it validates the evaluation … If there weren't, again, that would cause me to reflect and think "what am I doing wrong".

This perspective, that test scores provide a superior indicator of efficacy and that its misalignment with professional practice ratings permits re-evaluation of the initial

ratings, raises significant construct validity concerns: It creates a one-way association between test scores and ratings; It degrades the inferential value of the rating process by violating the integrity of the rubric criteria; It artificially increases the association (correlation) between the two independent measures; and it interferes with the interpretation of the Danielson sub-component ratings as a source of critical reflection and behavioral improvement. In essence, it reduces the evaluation process to a single measure while purporting to be composed of multiple independent measures.

Out of the eight principals interviewed for this study, only one differed in his/her perspective on the relative value between the two measures. Like many teachers, this principal valued the long-term relationship built-up between evaluator and teacher as a source of important information. Test scores are seen as a *snapshot*, a single point in time, and unreliable ("…The kids took a one day [test], and you're putting a human element to it…"). The individual comments:

> Interviewer: Which do you value more as a good measure of instructional quality, professional practice ratings or test scores, what if they don't match up?
>
> Principal 206: It's gonna be the professional practice. Because the student growth percentile, that piece, is based off of one day. The kids took a one day [test], and you're putting a human element to it. Whether or not that kid was off … whatever happened in the 35 lives of those kids that day, so you're putting a human element piece to it [test scores] that the teacher is not in control of, and you're putting it [on] one day. That's the reason why … with the amount of times we're in the classroom, I'm able to see some other pieces, and plus because [test scores] only takes one measurement.

In summary, principals position academic achievement as an important measure of instructional efficacy. Other attributes of professional practice are viewed as servicing

this end and are therefore afforded less weight. There is a causal link between instructional practice and learning such that both should be highly correlated and aligned. If the two measures are found to be out of alignment, seven out of eight principals place higher confidence in test scores as an indicator of teacher competency.

Metric preference - district. District participants were uniform in view that test scores provide a poor reflection of instructional quality. Lack of confidence in test scores suggested that they did not necessarily have to align with professional practice ratings. Not surprisingly, district members were uniform in placing more confidence in professional practice measures when faced with situations where professional practice and test scores did not align.

Reflecting general skepticism regarding alignment, one district member states:

> … Whether or not test scores determine for sure that that teacher's a good teacher or not is, there's no evidence for me to support that it definitely says, "This is a good teacher or not a good teacher." … because I think sometimes people can have really good test scores and may or may not have been a good teacher. (District 301)

Reacting to which measure is more indicative of quality teaching, the individual goes to say:

> Personally, I would put more confidence in observations that are done by trained people that really know what they're looking for and have an opportunity to ensure that they [teachers] do the dialogue and reflection pieces with it. I think that [with] those you're seeing practice in action. You can experience it in the classroom and experience, see what, hear what the teacher's saying. Hear the tone that's being used with the students; hear the responses that the students are giving during instructional time. (District 301)

Unlike, principals, this district participant places more value on observational ratings made by "…trained people that really know what they're looking for…" Professional Practice ratings trump test scores because they are constructed from

experience, "… seeing practice in action….", and from the evaluator being witness to instructional action. It is argued that these experiences provide superior representations of quality than do test scores.

A second district participant (below) shares this preference saying:

> Interviewer: Which would you place more confidence in, in identifying a good teacher, the evaluator observations or these objective test scores?
>
> District 302: It'd be the evaluator observations, absolutely. As we try and meet that expectation, and eliminate the ineffective teachers out of the classrooms, there is that responsibility for the observer, the evaluator, to balance that and say, "I know that this is an excellent teacher, has good relationships with students, has fair to good performance for the students' academic performance," but the [test] instrument only measures one part of that, where the evaluator can capture, more globally, the teacher's performance of all aspects of their daily job. It [the test] captures the score, the number on the students' performance. The evaluator who does the observation is able to account for much more than that.

Here, test scores provide a limited representation of competence ("…the [test] instrument only measures one part of that…" and "…The evaluator who does the observation is able to account for much more than that…"). Indeed, it becomes the "…responsibility for the observer …" to verify test-based interpretation with observational evidence in order to provide "balance." This responsibility is anchored in the participant's lack of trust in test score's ability to present a complete representation of instructional quality (i.e., "…the [test] instrument only measures one part of that…"). Again, it is this act of bearing witness, of experiencing, of understanding the complex environment in which instruction takes place that provides authority to the professional practice ratings over the achievement measures.

Reacting to the question of preference, a third district member makes an emphatic statement, choosing observational evidence over test scores:

Interviewer: Which would you put more confidence in, in identifying a good teacher, a test score or an observation?

District 303: I would say an observation, yeah, not even a question.

The person justifies this position based on the perceived inferential inadequacy of test scores ("…3 days of an assessment…") and the potential for score bias ("…doesn't take into account the composition of your class. I don't think that that's fair…"). (Note: some of this excerpt was referenced in the previous section concerning test scores):

> … I think that when you try to quantify 180 days of instruction to identify what a good teacher is based upon 3 days of an assessment, that you can argue the validity of the assessment as well. I think that's when you start to have a problem … To just look at purely a cut score piece and include that as your achievement perspective, doesn't take into account the composition of your class. I don't think that that's fair to the teacher that gets the gifted cluster in their classroom and the same grade-level teacher that gets the Special Ed cluster in their classroom. (District 303)

The fourth district interviewee (below) continues this trend, responding:

Interviewer: So, if you had to choose between the Danielson ratings and test scores [to inform on good/effective teaching] you would place more confidence in the Danielson score?

District 304: Absolutely, there's more flexibility in that. We need to have flexible tools that will be able to give us feedback as things change.

Here, the rationale is that test scores lack flexibility, adaptability, and inclusion of new types of (21st Century) outcomes. During the discussion the individual offered the following perspectives that help clarify the position (Note: portions these exemplars were reference in the previous section concerning adequacy of test scores):

> When we're looking at that profile of a [21st Century] graduate, it is more than what's measured on an AIMS test or any national test. It [good/effective teaching] is not just about being literate in terms of reading and writing, and being capable in math, and really that's the only thing we test … Using that reading score and that math score to really then determine whether that science teacher or that CTE automotive teacher did a good job in preparing the student doesn't make sense to me. There's this disconnect … the system's broken. To keep trying to define it in this very limited box doesn't make sense. It just doesn't make sense. (District 304)

> We have to establish other ways to get that information about how successful teachers are. I don't think any of us as educators want to walk away from any type of accountability. Yes, we think along the way we need to know where students are academically, but right now it's as if these national assessments are the only way we know if schools and teachers are doing a good job. If we want to compete globally, what's more important, a student with high AIMS scores, or PARCC scores, or whatever we want to call those, or a student who can speak three languages? (District 304)

Test scores are perceived as incomplete, omitting skills and knowledge recognized as important for 21st century graduates. This elevates the importance placed on professional practice ratings. Referencing "…that profile of a graduate…" the individual believes that achievement measures need to expand beyond reading and math saying "… what's more important, a student with high AIMS scores … or a student who can speak three languages…" In this dialog, it remains unclear whether this perspective would change if the test score domain were to expand to include a more contemporary and expansive set of outcomes.

As stated, district participants were very clear in their preference for professional practice ratings over standardized achievement measures when faced with conflicting interpretations. They viewed test scores as providing a limited representation of instructional competency while at the same time valuing the use of trained evaluators who bear witness to classroom/instructional activities. One district participant expressed

his/her preference in terms of responsibility, to validate or disconfirm inferences made from test scores with evidence of professional practice collected through observation. This perspective aligns with those expressed by teachers but is antithetical to those expressed by principals, who believed that observational ratings should be made to align with achievement scores.

Metric preference - state. The metric preferences expressed by state-level participants more closely aligns to those of teachers and district representatives. However, the state perspective seems more nuanced, contextual, and qualified. For example, one state participant provided the following initial dialog:

Interviewer:	Do you have an assumption that good teachers produce high test scores?
State 402:	Hmmmm. No. The answer to that question is no
Interviewer:	The quandary in the policy framework is that we have two major components: Professional practice measure and test scores. If those two things don't match (one is high, one is low), which one do you put more trust in to tell you about whether or not this is a good or effective teacher?
State 402:	Good professional practice.
Interviewer:	Why would that be?
State 402:	Because I think the professional practice encompasses many more behaviors and interactions with the people that a teacher interacts with, including students and parents and peers and community and supervisors, than the test scores do.

In this context, test scores provide limited information and are viewed as omitting important attributes of quality instruction. The choice is to value professional practice information over achievement indicators. However, the same individual qualifies this

perspective in the context of time: if test scores continue to indicate low levels of achievement, then their importance in the evaluation process rises. The person explains:

> [But] Don't misunderstand me. [While] I'm not a proponent of testing, I believe that tests have a place in our system … To say that the test on these three days during the school year, or these ten days during the school year, aggregated to some score for a teacher, I just don't think is representative of that teacher. … [However] There's also the element of time too. If they [Test scores & PP] drastically don't agree year after year after year, then I think it's incumbent upon the evaluator to begin to ask some questions. (State 402)

Interviewer: So if the test scores over time are out of line with the professional practice score, you would then begin to question what?

State 402: The evaluator.

The context is nuanced in that at any given point in time, test scores are a less valued indicator. But if learning does not increase over time, and observational ratings continue to suggest instructional excellence, then something is wrong with the evaluation measures. The implication is that achievement will increase over time in the presence of quality instruction. In this way, the measures should be longitudinally, positively, correlated. But at any specific time, the two measures may not completely agree.

A second state participant also offered a slightly different reaction to this context of metric preference. The dialog starts by making a distinction between stakeholder groups – different stakeholders will have different perspectives on what is important (note: portions of this individuals dialog were referenced in the previous section on test score adequacy):

Interviewer: Can someone be a good/effective teacher and have low/poor test scores?

State 403: Again, I think it depends on what you're evaluating and how you're evaluating. If you're evaluating a state evaluation of

> that teacher in the context of the academic standards that we require, and if that student … is making gains, or in this context not making gains, and also the student is at very low academic levels, then the answer would be no.
>
> However, again, that principal, or lead teacher for that matter, that's evaluating at a local level might find that teacher has been able to engage the students in other ways, that they can say, "Okay, we may not be getting those kids to meet academic achievement levels that we want to get to, but quite frankly they're coming to school every day, whereas last year those kids weren't coming to school every day. There's different levels of success and different levels of effectiveness that we need to be looking at.

Here, the state-policy perspective emphasizes test scores as a primary outcome while district-located leaders may value more affective instructional impacts on students. However, there remains recognition that test scores do not necessarily embody all aspects of a multi-dimension construct. This individual further describes how the general public and state policy-makers would "always" value test scores above measures of professional practice, saying

> Well, I can tell you that the achievement measure always trumps, at least at the state level. Let's be honest. The achievement measure will trump because the achievement measure is what parents, community members, and legislators look at. … They don't care a whit about the teacher effectiveness levels. (State 403)

However, when reflecting on his/her own perspective the person went on to respond:

> Interviewer: Okay. Again, if these two things [measures of professional practice and test scores] don't match, which one would you put more stock in?
>
> State 403: Okay, I'm going to tell you, I'll answer your question. I'll be totally honest with you. As a parent, I put more stock in my daughter having exposure to those teachers that are the transformative teachers, even if it were at the expense of her academic success, because to me, I would value that more. I think she would be a much better person because of that. Now

> from the state policy, as a member of the [position] I have to say that we have the responsibility to show academic attainment against those core academic areas that it's important for us to achieve.

For this person, the comment reveals a subtle distinction in perspective: parent versus state policy leader, where positionality impacts the emphasis placed on different evaluation measures.

In summary, state policy participants valued professional practice ratings and recognized limitations of test scores as a primary indicator of instructional quality. However, the perspective was qualified by factors of time and positionality. Over the long run, academic achievement represents an important educational outcome, and instructional competence cannot continually be depicted as high while learning fails to improve. In the short run, test scores fail to represent important aspects of good/effective teaching. Finally, the value placed on tests scores differs by stakeholder group. It is posited that the general public and state-policy environment focus almost exclusively on test scores as the primary indicator, while educators and individual parents may place more value on the affective aspects of schooling as represented by professional practice ratings.

Overall construct summary. Overall, teacher and district stakeholders trusted professional practice ratings over test scores as a primary indicator of instructional quality. In contrast, principals generally afforded test scores higher status. State members discussed substantive limitations of test scores but qualified their critique in the context of time (scores should increase in the presence of quality instruction) and stakeholder perspective (i.e., general public and legislators versus educators and parents).

Petite assertions - RQ1b(d).

General efficacy.

Petite assertion #1. For teachers, district, and state members, the evaluation system fails to provide an adequate representation of Instructional Competency.

Teacher, district and state participants provided generally negative reflections on the evaluation system's ability to accurately assess TIQ. Teachers were more likely to include concerns over rater reliability and attempts to reduce teaching to a fixed set of measurable activities. State participants cited the large proportion of (Group B) teachers lacking direct measures of the instructed content and the lack of early-grade (K-2) indicators. District and state participants shared their reticence regarding the lack of a common definition of instructional quality.

Both teacher and district participants commented on missing/unmeasured attributes of quality teaching as well as agreement that test scores were not good measures instructional impact. District narratives conveyed a sense of ongoing improvement (i.e., getting better, doing the best we can) in spite of the state-policy imposed constraints. Finally, state participants reflected that the quality of evaluation systems varied across Arizona districts and that the single mentor-evaluator model could be detrimental to school climate and culture.

Petite assertion #2. In contrast to teacher, district, and state members, principals express positive sentiments on the ability of the evaluation system to adequately represent instructional quality:

The principal group held more positive sentiments regarding the evaluation system's ability to produced adequate representations of instructional quality. They based

this perspective on the quality of training they (and teachers) received, clarity and specificity of the Danielson framework, and the use of objective evidence (both test scores and FFT ratings) to assess instructional quality. Principals also believed that these factors lead to high levels of trust and acceptance (in evaluation results) among teachers. For some principals, growth measures were the preferred utilization of test scores as an indicator of instructional efficacy (Reader's note: Growth scores refer to year-to-year changes in relative achievement while a status score refers to the amount of content mastery attained at a particular point in time).

*Time, frequency, and observation***.**

Petite assertion #3. Lack of adequate time/frequency allotted to the evaluation of individual teachers was a dominant concern expressed, to differing degrees, across all stakeholder groups. Lack of time/frequency has a negative impact on the ability of the evaluation system to adequately represent teacher instructional quality.

Lack of time/frequency was raised as a substantive concern for teachers, district and state members. It is portrayed in the following contexts.

- First, teacher and state level narrative suggested that the process of conducting evaluation for all teachers on a campus prevents allocating adequate time/frequency to any single teacher. This inhibits the evaluator's ability to obtain a deep understanding of teacher practice.
- Second, state level narrative suggested that the substantive time required to conduct campus-wide evaluations limits an administrator's ability to attend to other equally important instructional leadership responsibilities.

- Third, teacher narrative suggested that insufficient time/frequency in classrooms increases the possibility for gaming the system, activities presented during formal observation events may not adequately reflect day-to-day instructional activities of classroom teachers.
- Fourth, state-level participants suggested the principal-as-single-evaluator model may harm school climate, culture, and collaboration.

In contrast to the above, most principals did not believe that time/frequency of observations was an issue. Indeed, most believed that they allocated sufficient time to the evaluation process and that this increased the teacher's trust in the evaluation results.

Test scores/measurement (adequacy and metric preference).

Petite assertion #4. Teacher, district, and state member groups generally viewed test scores as a poor indicator of instructional quality.

Across stakeholder groups, numerous limitations associated with test scores were expressed. These include:

- Limited academic content representations
- Unmeasured attributes of good/effective teachers including affective impacts on students and affective attributes of teaching
- Unmeasured, indirect, curriculum (non-academic goals and objectives)
- Influence of non-instructional factors (bias, reliability)
- Limitations associated with a single event, snapshot in time, indicator
- Unintended consequences including the narrowing of curriculum and restriction of instruction

- Some state members suggest that the published research reveals that use of test scores as a primary accountability/quality measure has not lead to higher achievement
- Over emphasis on test scores harms organizational climate, culture, and collegial collaboration

While recognizing the limitations, state members were more likely to value inclusion of test scores as a legitimate component of a larger evaluation mosaic than were teachers or district representatives. This was partly due to the perceived need for some basic level of accountability and their belief that the general public values test scores over other, less objective, forms for measuring instructional efficacy.

Petite assertion #5. Unlike other stakeholder groups, principals were singularly positive about the importance and adequacy of test scores as an evaluation metric.

Principals were more likely to view test scores as a main objective of instruction. When principals did raise concerns regarding the use of test scores, the comments were related to missing affective attributes of professional practice (teacher commitment, engagement, passion), and the attribution of score to non-tested content area teachers.

Petite assertion #6. Teacher and district members generally viewed professional practice measures as providing a more accurate representation of instructional quality.

Principals voiced the contrary perspective, emphasizing achievement as a more trusted indicator. State members generally devalued test scores as a suitable indicator except in the context of needing a base-level metric of achievement outcomes for ensuring district/school/teacher accountability and reporting purposes.

Research Question 1C: Consequential Evidence (RQ1C)

RQ1C (a) and RQ1C (b). In what way has implementation of the teacher evaluation system affected the PP of classroom teachers (instruction, student learning, professional capacity building, job satisfaction, etc…)? Do the perspectives of efficacy and system affect differ across stakeholder groups (Teachers, Principals, and Policy Makers)? The approach is semi-structured interviews. The measures are coded interview responses.

This research question examines stakeholder perspectives of the evaluation system's impact on teacher (and administrator) professional practice. Many measurement theorists argue that examining consequential evidences is a critical requirement of construct validation (Kane, 2001; Kimball & Milanowski, 2009; AERA et al., 1999, 2014; Amrein-Beardsley, 2008, 2009). Specifically, validity is viewed not as an attribute of data, but rather of the "…uses, interpretations, and claims…" being made from data (Linn, 2008). Messick (1989a) states that validation is concerned with "…the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). In this way, consequential evidences become an active component of validation study.

Arguably, any process meant to evaluate professional competency necessarily leads to efficacy judgments on the individuals being evaluated. In turn, these judgments may have both intended and unintended consequences (AERA et al., 1999 2014; Amrein-Beardsley & Collins, 2012; Collins, 2012). The purpose, intent, and implementation of Arizona's state-directed teacher evaluation framework is explicitly connected to action

and consequence. That is, one of the main intents is to improve the professional practices of teachers and, by association, student achievement. Therefore, the validity inquiry questions whether or not this goal is accomplished inclusive of any unintended, unforeseen, consequences?

Process. To explore the consequential dimensions of the evaluation system, participants were asked to respond to the following general interview prompt: *How has implementation of the teacher evaluation process impacted your instructional practice?* Variations of this theme were initiated depending on stakeholder membership:

Teachers:	How has participation in the teacher evaluation process impacted your professional/instructional practice?
Principals:	How has participation in the district's teacher evaluation process impacted your professional practice? How has it impacted the instructional practice of classroom teachers?
District:	How has implementation of the district's teacher evaluation process impacted the instructional practice of classroom teachers?
State:	How has implementation of the state's teacher evaluation framework impacted the instructional practice of classroom teachers?

The prompt remained intentionally vague, allowing participants to reveal the context of their thinking regarding change and impact. Within the narratives, stakeholders shared on a mix of positive and negative aspects of system impact. To reflect these perspectives, the feedback from each stakeholder group is organized according to these two dominant viewpoints.

Within the positive perspective, two sub-components became apparent: *Clarity, Focus, Structure (includes precision, reliability)* and *Reflection, Communication, Dialog,*

and Goal Setting. Similarly, the negative concept was further distinguished by two sub-components: *Conformity/Reductionism* and *Fear/Stress*. Reflections from each stakeholder group (teachers, principals, district, and state) are discussed below within this structural context.

Positive impacts.

Construct overview – Positive impacts. As will be examined in this section, stakeholders identified a number of positive impacts associated with the evaluation process. It is argued that two dominant themes are identified in the narratives: *Clarity, Focus, Structure* and R*eflection, Communication, Dialog*. Importantly, R*eflection, Communication, Dialog* is posited to be a causal function of *Clarity, Focus, Structure*—that is, the structure and operational detail specified by the evaluation process permits targeted reflection and dialog on instructional practice. Without this well-defined structure, this type of focused communication would not be possible.

Clarity, Focus, Structure is represented by the following perspective:

> The evaluation structure exhibits a clearly defined set of elements and expectations. The Danielson professional practice (PP) components are operationally identified, as are their rating criteria. This clarity helps focus instructional practice. Classroom teachers have a clear understanding of the behaviors they are expected to address and be evaluated on (to attain a high rating). In addition, the operationally defined structure provides a sense of reliability, consistency, in evaluation ratings (applied to everyone).It is recognized that test scores are part of the overall evaluation rating process. However, reflections of impact are mostly attributed to aspects of the professional practice portion of the system. For teachers, it is unclear from the data whether they adjust practice to fit criteria out of conformity, the desire to be rated highly within the evaluation context, or because they see the components as legitimate elements of instructional improvement.

The concept may be described by the following codes and identities. See Figure 43.

standardized	objective
clarity	evidence-based
consistency	organized
common standards	focused

Figure 43. Codes and identities related to *Clarity, Focus, Structure*.

Reflection, Communication, Dialog is represented by the following perspective:

> The evaluation system promotes personal reflection on teaching. In addition, it promotes dialog and communication with evaluators (administrators). These attributes arise directly from the evaluation structure - a result of its structural/process requirements, its conduct, and its implementation by evaluators (multiple meetings, walkthroughs, observations, pre/post conferences, etc…). Communication and dialog facilitate relationships, which strengthens the accuracy of the final evaluation. Goal setting is an outcome of the reflection process. Reflection, dialog, and communication are facilitated by attributes of clarity and organization.

This concept may be described by the following codes and identities:

discussion	empowerment
communication	feedback
(critical) reflection	conversation
dialog	

Figure 44. Codes and identities related to R*eflection, Communication, Dialog.*

Stakeholder narrative – Positive impacts.

Teacher (positive). Many teachers reflected on the benefits of having a clearly delineated evaluation structure. Clarity and structure seem to ease teacher anxiety in

terms of explicitly knowing what and how they will be evaluated. In addition, clarity and structure help teachers align their instructional practice to the measured evaluation components: the 22 behavioral elements outlined under the Danielson framework. Interestingly, positive attributions of evaluation clarity and structure focused heavily on the Danielson framework. Comparatively, positive attributes regarding test scores were infrequent.

One teacher comments on how clarity and structure has made a positive impact on his/her instructional practices:

> I think the actual [Danielson] rubric is definitely a positive for me because it gives me something on paper in black and white to do. It probably could actually be even more detailed or maybe some more suggestions listed on there for some alternate, maybe some special needs, students. (Teacher 101)

There is a notion of conformity in the reflection, but here it is seen as a good thing, an improvement. It results in greater focus; it is "black and white", tangible, something that can be acted upon. Clarity permits purposeful and positive modifications to instruction. The same individual goes on to say:

> I think now in the last year and a half, I think everybody has the same rubric. Everybody has a distinguished, a proficient, so whatever school you're at, even though the principal is different and, the principal, you might not have trust with that principal or whatever, but that rubric is the same for everybody. For me, it's helpful to know that if I want to do my best, these are the things that I have to do. That's really helped me as a teacher is having that rubric in place saying if you want to get highly qualified or distinguished, these are things that we're looking for, these are things that have to be done in your classroom, with your students, with your parents. (Teacher 101)

Embedded is a concept of consistency (i.e., "…everybody has the same rubric…" and "…that rubric is the same for everybody…"). There is comfort in knowing this because it implies a common definition of what is expected. In addition, the adequacy of

this definition is not being questioned. Standardization also seems to permit reliability in ratings across different evaluators. Interestingly, one might argue that this person's perspective is not so much about improving instructional practice as it is about *chasing the rating*. That is, what one does to get a good evaluation may be substantively different than what might be done in the absence of the evaluation system. This distinction is not fully distinguished in the narrative. Regardless, the impact of the evaluation process is positive, "… it's helpful to know that if I want to do my best, these are the things that I have to do…"

Another teacher (below) shares a positive reaction to the evaluation process, stating:

> It's made me think a lot more about what is it that I'm supposed to be doing. It's not just "plan the lesson" and "do the lesson". It's made me think more about, "How have I been involved in school or do I need to step out of the comfort zone?" (Teacher 106)

Here, the evaluation system has increased the individual's critical reflection of practice. The clarity of structure and components permits a deeper focus on instructional actions. It transforms instructional action from the simplistic "… just "plan the lesson" and "do the lesson…" to the reflective and personal "…do I need to step out of the comfort zone…" This person continues:

> I think when I first started working in the district it was kind of like, "Okay. Here's your evaluation. Good job." Then we didn't really think about it much more, but, now, there are so many more components, and it's so specific. We [administrators and teachers] really talk about them [components] ahead of time - "What do you want me to look for? Are there any specific areas that you want advice on or feedback on?" I think it's helped in the planning of executing what you're gonna do in the classroom, but, also, the reflection piece, I think, has gotten much more important because it's not just teach it and move on - it's teach it, look at the data, and then where are we gonna go from here?" (Teacher 106)

For this teacher, the impact of the system has been increased dialog, communication and critical self-reflection (i.e., "…we really talk about them [components] ahead of time... the reflection piece…"). This has "…helped in the planning…" of the instructional activities "…in the classroom…" and this new environment is substantively different than prior evaluation programs (i.e., "…when I first started working in the district it was kind of like…") The source of this perspective is the articulation of components (i.e., clarity and structure), saying "…[there are] so many more components… it's so specific…").

Throughout the collective teacher dialog, positive impacts from the evaluation process are tied to this idea of clarity, structural definition, and specificity. These attributes facilitate instructional improvement by permitting focus. A third teacher (below) expresses this in his/her comments:

Interviewer:	How has our evaluation system impacted you as a teacher? Your professional practice - if at all?
Teacher 102:	It has. I mean when I go to have my annual evaluation, I look at Danielson and I'm looking at like, "Okay, well how can I make sure that I'm exceeds," or what is it now proficient? I mean, am I doing these things to make sure that I can be in these two top categories? Am I making sure that I'm not in these two bottom categories and if I am, how can I improve myself? I'm always looking for ways to improve myself and I definitely never want to be in those last two categories. Every year that goes by I'm looking to see that I'm having more and more distinguished instead of just the proficient. What can I do to make sure that I'm doing that? It makes me think about my craft more.
Interviewer:	Is there a link between the things that you're working on to get the higher rating, and your belief system that "If I work on these I actually am getting better"?

Teacher 102: Yes, because I don't want to be that average teacher. I want to be that above average teacher and, as teachers, most of us want to be that more-than-average.

For this person, the main impact of the evaluation system is "…It makes me think about my craft more…" There is a causal pathway between the evaluation structure/process and changes to instructional practice. The person expresses a desire to become a better teacher (i.e., "…I'm always looking for ways to improve myself …"). The pathway/criteria for improvement are provided by the components/rubrics within the Danielson framework (i.e., "…am I doing these things to make sure that I can be in these two top categories…"). Finally, there is no indication that the Danielson Framework is insufficient, inadequate, or unrepresentative of best practices.

As before, it is an arguable interpretation of whether the individual (above) is simply defining improvement as attainment of a higher rating or whether he/she believes higher ratings are synonymous with improved practice. Regardless, throughout the teacher narrative there is a reliance on the framework to identify changes in instructional behaviors. To this end, a fourth teacher (below) explicitly refers to the evaluation framework as a *guide* for adjusting practice saying "…What am I supposed to be doing in each of these categories to grow…" As before, growth is tied to clarity, clarity is a form of conformity to best practice, and best practice is defined by the behavioral criteria identified under the framework rubrics.

> We wanna be that person [a teacher that gets a high evaluation rating]. I know that many colleagues of mine strive to get into those columns [Proficient & Distinguished on the FFT]. Has it helped with growth? Well, I mean it's helped guide where we need to grow. At least it's some sort of a guide as to, "What am I supposed to be doing in each of these categories to grow?" (Teacher 103)

Interviewer: I'm hearing you say it brings clarity to the things that you do?

Teacher 103: Yes.

Interviewer: Has it been a positive change? Are you a better teacher because of the new evaluation process?

Teacher 103: That's interesting. Before this evaluation was brought through, I went through the National Board process. I think that was something that really impacted my teaching more than the evaluation piece.

I got poor scores on a couple of [categories] with the new evaluation. I got some poor scores. Not terrible, but ones that I went, "Excuse me? I've always been in this category for this. Now I'm approaches?" I went back and fixed it and went back to my evaluator and said, "No, I fixed this, I fix my scores." [Laughter] "Come and look, I fixed it."

Interviewer: You changed your activities?

Teacher 103: Yeah, especially that communication with families. That's the one I come back to because that was the one that was low for me. I am, this year for sure, e-mailing much more, calling much more, making sure things are getting out.

Interviewer: That's a positive.

Teacher 103: Yeah.

An interesting aspect of this reflection (above) is the comparison to the "National Board process." Here, the teacher seems to value the national criteria above the Danielson-based criteria saying "… I think that was something that really impacted my teaching more than the evaluation piece…" At the same time, the person acknowledges the impact of the local evaluation system, citing positive changes/improvements in parent communication. However, there is a sense that if a different evaluation structure (besides Danielson) had been adopted, the teacher would be adjusting his/her practices accordingly while still placing higher value on the National Board criteria.

Teacher summary (positive). Overall, teachers indicated a number of positive attributions regarding the evaluation systems impact on professional practice. However, it remains arguable whether the perspectives are driven predominantly by the desire to attain higher evaluation ratings. The distinction is one of conformity versus improvement. Conformity predicts that the same narrative would be exhibited if any one of a number of alternative standards-based evaluation systems had been adopted. Improvement would suggest that instructional practices improved as a result of the evaluation structure, teachers become better at their craft leading to the betterment of student learning. This is an important distinction because conformity to the system is not necessarily an ascription of personal value. Regardless, the narrative presented suggests teachers believe the district's evaluation system has positively impacted professional practice.

Principal (Positive). For principals, the *clarity, focus, structure* theme is more closely aligned with the concepts of *evidence, objectivity, and reliability*. That is, the foundation of the system is data, objectively assembled from a prescribed process of gathering evidence and assigning ratings based on established common criteria. Again, attributes of *evidence, objectivity, and reliability* are artifacts derived from the evaluation system's *clarity, focus,* and *structure*.

The narrative suggests that structure provides comfort to principals in the form of confidence and assurance, that outcomes are accurate representations of instructional competence because they are derived from a well-defined process. For principals, *Clarity, focus, structure* impacts their professional practice (PP) by enforcing systematic review, conducting evaluation in the same manner for all teachers.

Like teachers, principal narratives also discussed aspects of R*eflection, Communication, Dialog* as an important positive aspect of the evaluation system. The structure of the system incorporates opportunity for reflection and communication. Interestingly, for principals, self-reflection is a key attribute, that they [principals] become better evaluators as a result of having to think deeply about specific areas in the Danielson Framework for Teaching (FFT).

The importance of evidence as a characteristic of the *clarity, focus, structure* concept is expressed by a principal:

> I think there are two big areas for me. The tool itself, the Danielson's rubric, and the model that we're using online with the CES system, has made me much more aware of evidence-based. I like that a lot, because then it's cut and dried. Here's the evidence. It's all right here in the script. Here's the things that you brought me - etcetera, etcetera. I can go in and look, "Oh, you haven't updated your grade book in four weeks. I don't know what else you want me to say, but you didn't score very high in this area. That's a professional responsibility." (Principal 202)

Interestingly, by relying on evidence, this principal feels less compelled to enter into discussions about the accuracy of the evaluation rating: "… I don't know what else you want me to say, but you didn't score very high in this area…" Here, evidence is equated to objectivity and accuracy, which in turn, permits confidence and assurance. There is a sense that a *benefit* of the system is the reduction in judgment required to evaluate: "…it's cut and dried. Here's the evidence. It's all right here…" Further, objectivity is a function of "…the tool itself…", it is provided/delivered, rather than worked at, strived for, or engineered by the evaluator. Importantly, evidence/objectivity is a product of the evaluation system's structure and requirements, to evaluate you must gather evidence.

Like some teachers, another principal depicts the clarity/structure concept in terms of guidance*:*

> I think as a new administrator it has definitely given me some guidance as to what I'm looking for. It's kind of given me a path to follow as to what I should be looking for, using vocabulary like evidence-based so I'm not just, "Oh, you seem like a good teacher." Now I'm able to go back to a rubric and state, "This is where you're at, and this is why," so I feel like the ratings are much more justified. To have been trained on it, I think that's a huge piece. (Principal 203)

Here, guidance is aligned to a well-organized evaluation process and the use of evidence, the elimination of subjectivity. By following the "path," the principal's ability to accurately evaluate is enhanced because "…the ratings are much more justified…" There is a confidence in the narrative that the new system is better than previous methods because of this reliance on evidence, structure, and procedure. Evaluators are *empowered* due to this confidence. This same principal goes on to say:

> I just feel more empowered in supporting teachers to become that "good teacher." I feel like we have more information. I've been informed. I feel like I have more information to support our teachers in moving forward. Again, all of it is data. It's evidence-based and "show me the numbers" and "show me how you do this". Just telling me isn't going to be enough and then for us, at the end, to wonder why our scores aren't where they need to be. (Principal 203)

For this principal (above), empowerment leads to support/mentoring of teachers. It permits evaluators to improve teacher's professional practice because feedback is evidence-based. The statement "… Just telling me isn't going to be enough*…"* is a requirement of proof, forcing teachers to look objectively at their practice and justify their actions. For this individual, the more that feedback is based on *data,* the more it will impact instruction. A different principal talks both about the empowering aspects of evidence as well as the benefits of collegial dialog between evaluators:

I'm much more focused when I go in the classrooms and I think I know what to look for more. I think I've learned more through this by talking with colleagues. My assistant principal and I have spoken so many times and go into rooms together, so that we can see that we're on the same page. I think every day I learn something … [Also] It's allowed me to take out more of my being subjective. I think it's allowed me to say, "I don't have the evidence that supports distinguished." I think it's kinda' given me a spine in a way because it's not so much about my perception. It's "This is the rubric and if this is what you want, then this is what you hafta' demonstrate", so it took me out of it I feel.

Interviewer: I'm hearing you say it's given you a little bit more comfort in that you're using a more objective measure to rate your teachers?

Principal 205: Yep.

Interviewer: And you feel that that's helped you focus? Did I paraphrase that right?

Principal 205: Yes, absolutely. I just feel like I'm not nervous about going into some of these meetings because I have a rubric to look at. I have that as my backbone, that if a teacher's gonna argue something, well, here's all the evidence. Here's the rubric.

The principal (above) is infused with confidence and empowerment because he/she is "…more focused … I know what to look for … it's allowed me to take out more of my being subjective… it's kinda' given me a spine…" The source of the principal's feeling is the evaluation process itself, its structure and clarity of criteria. For this individual, objectivity becomes authority. The results are beyond reproach: "… if a teacher's gonna argue something, well, here's all the evidence. Here's the rubric…"

The principal group narratives surrounding *clarity, focus, structure*, which are anchored in concepts of objectivity and evidence, are predominantly shared in the context of their own personal improvement: Principals believe they are better evaluators due to the process, structure, clarity, objectivity and reliance on evidence inherent in the system.

They are empowered by confidence and authority through the use of common, objective evidence, and they view this as a key improvement in the evaluation process.

Like teachers, some principals also cited increased reflection, communication, and dialog as a positive impact of the evaluation process. For some, the context was expressed in terms of peer collaboration, while others focused on evaluator-teacher dialog. In the exemplar immediately above, the principal expresses the benefits of peer-to-peer communication saying "…I've learned more through this by talking with colleagues. My assistant principal and I have spoken so many times and go into rooms together, so that we can see that we're on the same page. I think every day I learn something…" For this person, collaboration becomes possible because of the common structure of the system. The two administrators are able to view the same instructional event, utilize the same (scoring) criteria, and then discuss how they applied the evidence to arrive at their independent ratings. The implication is that this type of peer dialog leads to increased accuracy of evaluation.

Similarly, a different principal reflects on the benefits of evaluator-teacher communication:

> I had a two hour post [conference] on Friday with my eighth grade science teacher, and I think I grew as an administrator throughout that two hour conversation. In order for me to help guide and facilitate learning and conversations, I'm continually rethinking. I have to be reflective as far as what I'm gonna say, how is it going to be perceived, and how is it gonna be received. I have to be reflective as an evaluator to make sure that I'm continuing to grow and refine my conversations as far as giving that feedback. (Principal 204)

For this individual, the system permits personal reflection ("…I'm continually rethinking…") and the result is personal growth ("…I grew as an administrator…"). This opportunity to reflect was created from the requirement to conduct a *post-conference*.

Here, conferencing forces him/her "…to be reflective…" Importantly, this reflection activity is connected to a sense of personal responsibility, a duty, as an evaluator "… to make sure that I'm continuing to grow and refine my conversations as far as giving that feedback…" Regardless, the *reflection* concept is aligned to personal growth as an evaluator.

As under the *clarity, focus, structure* concept, reflection on communication/dialog is framed in the context of principal's own personal growth as evaluators: one from communication with peers, the other with teachers. The personal reflection concept is expressed by still another principal, saying:

> When I've been going into the classrooms, I'm much more focused, I'm scripting. But I'm also thinking "where it's gonna go". I'm starting to recognize that I'm not just scripting for the sake of scripting. I'm not copying everything the teacher is saying for the sake of "I have to hurry up and do everything". I'm thinking about where the teacher is going. I'm thinking about "oh, this is gonna be a great piece for the question discussions because this is definitely a higher-level question". As I'm typing it, I'm starting to think where things are going.
>
> I think it's [the evaluation process] helped me communicate [with] teachers because and it's helping them to be better. When I'm having conversations with them [teachers], it's not just about the lesson. It's portions of the lesson that were really good and what about it was really good. We're not just talking about the activity that the kids did. It's what makes it [the lesson] good. It makes the conversations with the teachers a lot better. (Principal 206)

The comments from this principal (above) suggest that *clarity/structure* provides the foundation for reflection, *communication/dialog*. The outcome of observing instructional behavior for specific (measured) attributes is more targeted focus on what is happening in the classroom ("…When I've been going into the classrooms, I'm much more focused, I'm scripting…"). In turn, the process of scripting provides the foundation for questioning, dialog, and exchange ("…I'm thinking about 'oh, this is gonna be a great

piece for the question discussions…'"). The end result is better communication with teachers ("…it's [the evaluation process] helped me communicate [with] teachers…").

Principals – Positive impact on teachers. As mentioned above, principal reflections focused mostly on the impact the system had on their own personal activities and development. To redirect their thinking, principals were asked to reflect specifically on teacher's professional practice. The dialog presented below examines this perspective. Reacting to the question on teacher impact, one principal shares:

> Interviewer: How has it [the evaluation system] impacted teachers' professional practice in classrooms?
>
> Principal 203: I find that they [teachers] continue to go back to the rubric to determine how to better their practice …-they're [teachers] reading the rubric, because they understand why they're receiving the scores that they're receiving - it helps them take that next step to better their practice.
>
> Interviewer: So a positive impact?
>
> Principal 203: I firmly believe so, firmly believe so.
>
> The other piece we're starting to hear more and more is we have teachers telling teachers, "Well, haven't you looked on line? Haven't you looked at the rubric? This is how we can get to this piece." If something comes up during PLCs, they do make reference to the Charlotte Danielson rubric. We're getting to the point where it is going to become our common vocabulary and something that is just a part of our practice, which I like.

The perspective is emphatic, "…I firmly believe so, firmly believe so…"; the evaluation system is having a positive impact on teacher's professional practice. The rational for the belief is adherence. That is, teachers are being evaluated by a clearly delineated set of criteria and they are able to reference this criteria "…because they understand why they're receiving the scores that they're receiving…" By adhering to the

rubric, and the framework's behavioral components, "…it helps them take that next step to better their practice…" There is a direct connection between the FFT and good teaching. If you follow the FFT, if you score highly on the component rubrics, you are a good teacher. Through adherence, you improve practice.

The individual also comments on the gradual incorporation of this *adherence* into the teacher's belief system: "… we're starting to hear more and more … teachers telling teachers … 'Haven't you looked at the rubric?'" This is held as evidence of positive impact. The logic is (1) the FFT defines what it means to be a good teacher, (2) if you follow the FFT and score highly on its components. then (3) you are a good teacher. Because teachers are increasing telling their peers to refer to the rubric is evidence of the evaluation system's positive impact on classroom instruction. This line of thinking is also expressed by another principal:

Interviewer:	How do you think it [the evaluation system] has impacted the professional practices of your classroom teachers?
Principal 204:	They're [teachers] more aware of what we're expecting every time we come in. They are more aware; they know that the objective needs to be aligned. I think just by having them know, now that we've informed them of the four domains, the 22 components, the rating, the everything. I look back at when I first came to this district in 2004, and it was almost like a mystery for them [teachers]. Ooh, it's evaluation time. They didn't know the rubric. … At least now, teachers know exactly what is expected, so I think that has changed the majority of teachers.
Interviewer:	That's been a positive change?
Principal 204:	Absolutely, because they are aware of how they're being evaluated, which we should be aware of, how we're gonna be measured.

Interviewer: Would you say that participation in the evaluation system has improved the instructional practices of your classroom teachers?

Principal 204: I would. I would say that it has improved the majority of instructional practice, because now I think teachers are better able to label what they're doing. They know what the planning and preparation piece, what the criteria, what those components are, what it's consisted of.

As before, the basis of believing that the evaluation system is improving instructional practice is adherence to the components within the FFT. This is evident throughout the narrative: "They're [teachers] more aware of what we're expecting … they are aware of how they're being evaluated … just by having them know … we've [the district] informed them …" In this way, awareness and adherence is the basis for the improvement: "… at least now, teachers know exactly what is expected, so I think that has changed the majority of teachers…" This connection is very apparent in the exchange when asking if the changes have been positive, responding "… Absolutely, because they are aware of how they're being evaluated…."

Another principal shares how his/her teachers go out of their way to showcase alignment to the FFT:

Interviewer: How, if at all, has the evaluation system impacted the professional practices of your teachers? Have you seen it change their instruction?

Principal 206: The one thing I've seen different is that they are asking me to come in for specific things, like, "We're doing a really neat thing today, and we'd love for you to come in and watch this. What you're going to see is," and they start referring to the rubric. "What you're gonna see is some great discussions. The students will be leading the discussions+ and that kinda stuff.

Interviewer: Because that's part of the distinguish rubric?

Principal 206: Right.

Interviewer: Yeah, so a positive change?

Principal 206: Right. Yeah.

Here, the desire to adhere to the FFT, and to receive high performance ratings, is so strong that these teachers engineer review outside of the formal evaluation structure. The purpose is to ensure their administrator (evaluators) witnesses their highly aligned instructional activities. By implication, the fear is that administrators will not see these activities during the normally allotted evaluation time. The motivation is to attain the highest possible score (Note: FFT component rubric scores range from zero to three; to receive a three requires evidence of student-directed learning). Again, for this individual, the evidence of positive impact is this adherence to the FFT based on a belief that it equates to good/effective teaching.

This equating of instructional quality, adherence to the FFT, and evidence of positive impact is continued by still another principal (below):

Interviewer: If I were to ask your teachers "Has the evaluation process changed your professional practice?" What would you think they might say?

Principal 208: I would think that the majority of them would definitely say that it has impacted them positively … they [teachers] don't really question the [Danielson] components at all. I believe, that they feel, that that's what good teachers should be doing, as well. Yep.

Like his/her colleagues, there is a perception by the person (above) that teachers share in the belief that the evaluation systems has resulted in positive changes in professional practice. The principal asserts that teachers see the framework as representative of good teaching ("…they feel that that's what good teachers should be

doing …"). The presumption is that by following the framework teachers will improve their practice.

From previous analysis of teacher narratives regarding the pros/cons of the evaluation system, it becomes a debatable question whether teachers reflect the same level of conviction/certainty as principals. Of debate might be the relative influence/importance that adherence/conformity plays in shaping teachers conviction—is conformity a stronger incentive than a true belief in the representativeness of the FFT to accurately represent what it means to be a good/effective teacher? Interestingly, this same principal goes on to comment:

> Interviewer: OK. If we were to take the Danielson framework and just get rid of it, would teacher's classroom practices in your school change very much?
>
> Principal 208: No, I don't think so. No. Especially, because to me, I would still be getting in [classrooms]. I would be getting in and doing those informals and providing that feedback, and coaching and things. Yep. Usually, those are more of what impacts them, I think, just the conversations that we might have.

This is telling statement because it colors the perspective offered by the individual. Here, the implication is that the most important aspect of the evaluation process is not derived from the FFT components themselves. Rather, the benefit arises from the supporting/mentoring relationship formed between evaluator and teacher. By stating "…those are more of what impacts them…" this principal is valuing the dialog/communication required under the evaluation process. An interesting question is whether this *perspective* is widely shared by his/her colleagues? Unfortunately, the data provided herein does not address this. However, it does contribute to the question of adherence/conformity—would similar narratives be expressed regardless of the formal

evaluation system adopted, as long as that system had clearly delineated structure and measurement criteria?

Finally, another principal reflects on the system's impact on teachers not in terms of the structure, but from the general perspective of process – evaluation leads to improved instructional practice:

> Interviewer: At this point in time, how has it [the evaluation system] changed teachers' practice, if at all?
>
> Principal 205: I think they [teachers] think more about their practice because they know we're looking at it all the time. I think, especially here, I'm doing two to three evals a week; so is my assistant principal. That's six teachers, on top of getting into two to four teachers a day, at least, and the same with her [the AP]. We're in and out of the rooms all the time. We're sending that feedback immediately. They're seeing it categorized by the certain domain so that those domains keep going in their face.

Here, it is the act of evaluating that leads to instructional improvement. It is the process of consistently evaluating, collection information, that facilitates meaningful reflection in their practice: "… because they [teachers] know we're looking at it all the time…" and "… those domains keep going in their face…" The implication is that without evaluation, teachers would not engage in this reflection. Frequent communication is a key attribute: "…We're sending that feedback immediately…" Finally, the evaluation system's measured components provide structure for this reflection process. The structure enables principals to provide feedback "…categorized by the certain domain…" (Reader's Note: Only two of the eight principals referenced test scores in their discussion of impact. In both cases the perspective was that test scores aided in focusing on an important goal of education: student learning.)

Principal summary (positive). Overall, principals reflected that aspects of *clarity, focus, structure* positively improved their ability to accurately evaluate instructional competence and provide useful feedback to teachers. This conviction is primarily founded on concepts of evidence and objectivity. Because FFT ratings require assembly of evidence, principals feel empowered and confident in their ability to accurately assess instructional practice. In addition, clarity, focus, and structure provide the foundation for personal reflection, development, and growth as an evaluator. There is a sense that by becoming better evaluators, teacher's instructional practices will improve. Principals also seem to indicate that clarity, focus, structure help teachers to focus on the measured FFT components which, in turn, improve their professional practice.

Adherence is an important concept in the principal narrative. Adherence refers the benefits realized by focusing on the evaluation system (FFT) components. There is an underlying belief that the FFT components represent good/effective teaching. By extension, principals believe that adherence to the FFT necessarily results in improved professional practice. In this regard, the structure of the system permits clear communication with teachers on what/how they are being evaluated. This, in turn, leads to teacher's increased focus and instructional improvement. Arguably, embedded in the narrative is an implied perspective that teacher competency improves as a result of the evaluation process in general.

District (positive). Interviews were conducted with four district-level policy makers involved with the evaluation system. Unlike the majority of principals, only one out of the four shared predominantly positive sentiments regarding the impact from the evaluation process. A second participant provided more tentative/qualified support. In

both cases the support was specific to the Danielson framework and did not address test score issues.

The individual expressing generally positive sentiments served in numerous roles within the district: classroom teacher, instructional coach, school administrator, and currently a district policy maker. The individual has had experience with previous incarnations of the Danielson framework. Reflecting on the evolution of the evaluation district's system, he/she initial shares:

> I worked with my own teachers at my own campus where we had an understanding. Then at the district level when I'm working with teachers from across the district and administrators from across the district, it's allowed me to see how inconsistent we've been, and really impacted my desire of really to make it something that we have that we're doing effectively, consistently, and fairly across the district. (District 301)

The perspective is that previous implementations were inconsistent. In contrast, the current formulation is seen as an improvement—effective, consistent, fair. The individual goes on to explain how previous versions were not helpful to classroom teachers primarily because they did not provide suitable feedback and reflection from which to make targeted improvements:

Interviewer: How has your work in the teacher evaluation activity process impacted you as a professional, as an educator?

District 301: It has. Yes. As a teacher I felt it was not valuable at all because I didn't ever get any feedback. Barely, I would get something that was signed, like, "Here it is. Sign it." There was really no reflective conversations.

But I became an instructional coach when I left the classroom. A large part of my role was to be trained on Danielson's model and work with teachers on it as a coach. Everything that I did, I based it on the rubric and the way professional development was targeted around it. I would identify those certain areas. From that point on, making teacher evaluation relevant,

> important, and meaningful to teachers became a goal of mine because I didn't get it as a teacher. It's something that I have been passionate about for a long time, actually.

Here, the main issue with prior implementations was lack of feedback provided to teachers. By implication, this prevented/limited improvement – teachers were not receiving specific critical reflection: "…Here it is. Sign it." There was really no reflective conversations…" Upon becoming an instructional coach "…Everything that I did, I based it on the [Danielson] rubric…" This was an attempt to bring specificity, clarity, and purpose to the evaluation process. Indeed, allocation of professional development and training were being directly tied to the FFT components. By implication, teachers are better able to improve practice if they have this type of specific feedback. The individual goes on to say:

> When I moved into this [district] role … because of the state legislature and because of the changes, it became a priority … It's also allowed me to work within this [evaluation] committee.
>
> I worked with my own teachers, at my own campus, where we had an understanding. Then at the district level, when I'm working with teachers from across the district, and administrators from across the district, it's allowed me to see how inconsistent we've been, and really impacted my desire of really to make it [evaluation] something that we have that we're doing effectively, consistently and fairly across the district. (District 301)

While not specifically stating that the evaluation system has had a positive effect, the perspective tracks improvement in the system from a context of providing biased, inconsistent, and insufficient information to that of effective, consistent, and fair reflections on teacher competency. By implementing the latest system, it is presumed that these changes have had a positive impact on instructional quality. However, a second district participant (below) offers a more qualified reflection, saying:

> …I would say Danielson, while it's one of the better models, it still narrows the scope of what people are looking for. My opinion of that is - good or bad - it has narrowed what teachers do in a classroom… (District 303)

The impression is that the FFT is as good as some competing evaluation frameworks. However, it fails to fully incorporate all aspects of quality teaching. Strict adherence to the framework therefore "…narrows the scope…" and "…narrowed what teachers do…" In this regard, the person is conflicted. On one hand, failure to fully capture the attributes of effective teaching arguably results in an incomplete measure of quality ("…narrows the scope of what people are looking for…"). On the other hand, for those attributes actually measured within the FFT, conformance/adherence may improve instruction. For remainder of the discussion, this individual focused on the negative impacts of utilizing a test score-centric evaluation system.

District summary (positive). Only one district representative provided a generally positive view of evaluation system's impact. The perspective is based on a comparison of current evaluation components, process, and procedures to previous incarnations of evaluation system. For the individual, the current version brings added structure and consistency, which in turn mitigates measurement bias, affords fairness, and improves accuracy. Most important for this individual is the addition of critical feedback and reflection, which provides teachers the necessary foundation from which to improve professional practice. A second district participant felt the new system brought added structure and focus but that the effect of this structure may be mixed. That is, the FFT may be missing important attributes. Exclusive adherence to the measured attributes necessarily narrows instruction. This narrowing can be positive to the extent that the measured attributes are characteristics of quality teaching: if more teachers focus on these

attributes, their competency will improve. In contrast, not focusing on important, but unmeasured, attributes may harm instructional quality.

State (positive). Three state-level participants each involved in policy aspects of the evaluation framework were interviewed on the impact of the state's system. Two of the three expressed generally negative perspectives while one provide cautiously optimistic views. This latter person reacts to the inquiry as follows:

Interviewer: Over the last year and a half, how do you think the state's new framework is impacting the quality of teaching and education in Arizona?

State 403: Well, I'd like to hope that it is, and the thing that I am hopeful, the reason why I believe it is, is that, at a minimum, and I've seen this across the state, there is at least a discussion now. With a lot of districts that didn't have good systems, they're at least talking about how they develop good systems. Or they're looking to districts that are noted to be strong, and looking at teacher evaluation, and then looking at how they determine whether or not that's an effective teacher. At a minimum, whether or not we're there yet, and fully functioning, and we're hittin' our stride, I don't think we are. But I think we're starting to get there.

I'm not so sure they [districts] understood what the data was telling them, even at the superintendent levels. I don't believe they could see the data and say, "Okay, here's where we're falling down and why." They just didn't. Now you're starting to get more and more individuals to have those conversations, to seek out those individuals that can explain it to 'em. They're starting to get it, and that to me is very heartening.

Interviewer: So, it's advancing the public discourse on what quality teaching, quality education is? You think it's a benefit to us?

State 403: How they're using the data when it comes to 'em.

The individual seem cautiously optimistic that the state's evaluation framework is having a positive impact: "…I'd like to hope that it is…"). The view is one of ongoing

improvement and development: "…I think we're starting to get there…" The primary catalyst is reflective discourse on quality teaching – an increase in communication and collegial dialog: "…there is at least a discussion now … at least [districts are] talking about how they develop good systems … starting to get more and more individuals to have those conversations …" Regarding the current status the individual remarks "…whether or not we're there yet … I don't think we are…", but "…they're starting to get it, and that to me is very heartening…" The perspective (above) is that districts did not previously have adequate teacher evaluation systems ("…a lot of districts that didn't have good systems…"). The positive role of state policy has been to raise this issue, provide a structured framework, and initiate necessary reforms. By implication, as districts improve in their ability to effectively evaluate classroom teachers, student learning will increase.

Summary – positive impacts. Across stakeholder groups, positive reflections on the impact of the evaluation process concerned aspects of *Clarity, Focus, and Structure* and *Reflection, Communication,* and *Dialog*. Here, the system provides an organized structure for implementing, communicating, and evaluating changes to instructional practices.

For teachers, *Clarity, Focus, and Structure* facilitate adherence to the required measured behaviors and activities. It provides teachers with a blueprint of what/how. A key nuance to the teacher perspective is adherence: by focusing on the specified evaluation components and associated performance criteria results in higher evaluation ratings. For teachers, a by-product of this is more focused communication and dialog with administrators on the evaluated criteria. A key question arising from the narrative is

whether adherence equates with conformity. The former implies agreement and internalization while the latter requires only compliance.

Principals provided a nuanced perspective on the *Clarity, Focus, and Structure* theme. For this group, *Clarity, Focus, and Structure* are equated with evidence, objectivity, and reliability. In this regard, the system's structure requires use of objective evidence assembled from a systematic process using defined criteria. It is believed that this leads to unbiased, accurate, and consistent measures of instructional quality. Because the information is objective, it permits formative communication and dialog with teachers regarding instructional quality. The structure also injects confidence, empowering principals to see themselves as effective evaluators: they follow the process, utilize evidence, and objectively evaluate resulting in accurate/reliable assessments of competency. It authorizes them to have focused conversations and dialog regarding strengths and weaknesses. In this way, the professional practices of teachers are improved. For principals, there is no distinction between adherence and conformity: since the evaluation structure represents the actions/behaviors of good/effective teaching, instruction is improved regardless of the personal beliefs of the teacher.

Only one district and one state level participant provided generally positive reflections on the impact of the evaluation system. For the district participant, *Clarity, Focus, and Structure* represent a substantive improvement over previous methods of evaluating teachers. The new system formalized provision of critical feedback and reflection and increased the amount/quality of administrator-teacher communication. Here, consistency of structure allows for focus on important attributes of teaching not emphasized in earlier incarnations. For the state level participant, the framework requires

districts to focus, reflect, and improve methods for evaluating teachers. Prior to the legislation, schools and districts were evaluating teachers ineffectively and inconsistently. The framework brings a structure to promote critically reflection and improvement.

Negative impacts.

Construct overview – negative impacts. Stakeholders identified a number of negative impacts resulting from the evaluation process. The two interrelated themes may be classified as *Fear/Stress* and *Conformity/Reductionism*. The *Fear/Stress* component is represented by the following perspective:

> Fear/Stress: The act of evaluation creates stress, harms morale, and damages the professional identity of teachers. There is trepidation associated with use of test scores as part of the evaluation process. Inherent is a perception that the evaluation components are incomplete (missing elements). Fear/stress forces conformity and parody to attain favorable ratings and job security. It presumes an element of unfairness and a lack of accuracy.

This concept is characterized by the codes and identities in Figure 45.

fear	punitive
stress	personal
negative impact on morale	emotional
concern	top down
judged	loss of job
labeled	high stakes
evaluated	consequential
lack of accuracy	

Figure 45. Codes and identities related to *Fear/Stress*.

Conformity/Reductionism is represented by the following perspective:

> Conformity/Reductionism: As a result of implementing a well-defined evaluation process, teachers are forced to conform to a finite set of requirements and behaviors. Conformity narrows both content delivery and instructional variety. It is restrictive and reduces instructional flexibility, creativity, and risk taking. The system fails to operationally capture the complex, multi-faceted nature of

teaching. Reductionism leads to parody and parody distorts the representation of instructional quality.

This concept is characterized by the codes and identities in Figure 46.

narrowing	excluding
reduction	parody
conformity	dog'n pony
limiting	performing
restrictive	parody

Figure 46. Codes and identities related to *Conformity/Reductionism.*

Teachers, district, and state members consistently expressed dimensions of these two themes: perceiving the embedded attributes as negatively impacting instructional quality. In contrast, principals de-emphasized the impacts, viewing fear/stress as a natural byproduct of evaluation and conformity/reductionism as a necessary adaptation for improving instructional practice. In this way all groups recognized these *negative* themes but differed in their perception of impact.

Stakeholder narrative – negative impacts.

Teachers (negative). Four of the seven teachers interviewed expressed a variety of adverse impacts attributable to the evaluation system: concerns over the use of test scores as a primary indicator, a reduction in teacher morale, and narrowing of instruction and learning. One teacher shared concern over the ability to conform to all the required components measured by the FFT, saying:

> I don't have a way [in my day-today teaching] to say, "Okay, standard number two, A, B, C, has been fulfilled." I think people want that because they're scared to lose their job, is basically the end result. (Teacher 103)

The view suggests difficulty in breaking down, and measuring, the complexity of teaching into individual, itemized, behaviors. By implication, there is concern that they cannot be effectively measured in a piecemeal manner. Regardless, teachers attempt to conform to this approach out of fear: "… they're scared to lose their job…" The comment seems to merge this idea of conformity with reluctance to view teaching as measureable, compartmentalized activities. This suggests a lack of trust in the evaluation process. The same teacher goes on to comment:

> It [the evaluation process] seriously narrows the creativity in the classroom because you're constantly working on these little quizzes, tests. I have seen teachers where that is their entire curriculum, and it's a little scary because there's no room for creativity, which is how children learn, at least in my opinion. I think we've lost project-based learning because of it. People are scared to take a risk because if I take a risk, I'll miss a week of test preparation. I'm noticing in the years that … after AIMS everybody does projects. Before that, it's real scary … because I wanna make sure my tests are good. (Teacher 103)

For this person, the evaluation process "…seriously narrows the creativity in the classroom…" because each individual component of the FFT must be measured, assessed, and/or accounted for. To do this, teachers are forced to consistently give "… these little quizzes, tests…" In addition, the focus of these assessments is restricted only to content that is explicitly tested by the state assessment, further narrowing instruction and learning. As a result the teacher suggests a loss of instructional creativity and de-emphasis on project-based learning "…which is how children learn…" There is a strong motivation to conform in order to protect job security and to attain high test scores.

A second teacher sees reduction of morale as a negative outcome of the evaluation process.

> I see teachers demoralized and ruined, and [their] personalities and hopes and dreams crushed because they didn't quite fit into what an evaluator was looking

> at. Maybe they didn't quite cut the data or certain rubric items that is on the evaluator's rubric. Yet I feel like they [the teachers] brought a lot to the table and a lot of potential. I think that if more teachers were supported instead of feeling like the evaluation was a reason to get rid of them, they would feel a little more strength and courage and realizing that they're not alone. I see so many educators leaving because they feel like they're alone. They're in a no win situation, and they are good teachers. They're not supported enough. (Teacher 104)

The teacher feels the evaluation process has substantively harmed the professional identities of his/her colleagues using terms like demoralized, ruined, crushed. The source of this is lack of support ("…if more teachers were supported…") and a view that the main purpose of evaluation is punitive: "… a reason to get rid of them…" As a result, teachers are leaving the profession "…because they feel like they're alone…" and "…not supported enough …" The statement "…They're in a no win situation…" exposes a forced conformity to an evaluation structure that is believed to imperfectly capture the essence of good/effective teaching ("…and they are good teachers…"), harming morale and professional identity. The same teacher goes on to respond:

> Interviewer: Has the evaluation system impacted you as a classroom teacher?
>
> Teacher 104: It causes a lot of drama. It causes a lot of emotion. It causes a lot of - seriously, she [administrator] just walked in and we're in transition. I just didn't get the objective on the board yet … It's like man, it's crazy, because they're just snapshots. Exhausting.

This comment reveals a causal link between time/frequency of observation and adequate measures of instructional quality. Here, infrequent observation raises the influence of non-instructional factors ("… just walked in and we're in transition…") potentially biasing the resulting performance ratings. This reduces trust in the system, raises *drama*, and "…causes a lot of emotion …" The individual continues:

Interviewer: Has it [the evaluation system] changed the way you teach?

Teacher 104: No, never. I'm too old for that. There's no way I'm going to … It is what it is. I will never change, even though we kind of have an ongoing joke within our grade level that when they [administrators] walk in and we jump to the board, it's like the Vana White show. Then we start going through the motions and the hoops then because we know they're looking for that.

This teacher indicates the evaluation process has had no impact on his/her instructional practice. From the narrative, the teacher is older and seems unwilling to change regardless of any type of administrative review. However, the comment raises this recurring issue of conformity to what is being measured ("… the Vana White show…" and "…going through the motions…"), suggesting evaluation results fail to capture the day-to-day instructional practices of some classroom teachers.

A third teacher is a bit more positive regarding the impact on practice while at the same time admitting to conforming to the evaluation requirements:

Interviewer: Has it changed the way you do your teaching?

Teacher 105: Well, somewhat. I look at those things [Danielson components] and I'm thinking', "Oh, wow, I don't do that. I need to put that into my lessons." … I have done things where my teaching is more towards my observations. I look at whenever I know I'm gonna get an observation. I do my lesson plan and I'm thinking, "Can I bring these things in?"

The reflection is not being expressed as a negative. However, the individual is acknowledging conformance as a consequence of its use – if the system was not in place, he/she would not be incorporating "… these things…" into the lesson. In addition, the perspective might be different if he/she did not have prior knowledge of when the observation was taking place. Lack of prior knowledge would limit this ability to conform for the evaluation period.

A fourth teacher sees the use of standardized test scores as a source of concern and stress:

> … you've got the whole standardized testing thing, and I think that makes us all very nervous. That's all definitely in our minds, I think, all the time. "What is this gonna mean for us once those numbers and labels get released, and then once next year comes, and we've have two years of that, what does that mean for us?" I think there's lot of nerves for teachers in terms of the data part.
>
> Interviewer: By the data part, you really mean the test score part?
>
> Teacher 106: The test scores, yeah. It's very scary for us [teachers] to think that their performance for three or four days [of testing] could determine our fate as teachers. That's how we look at it even though three-quarters of it is based on what our admin thinks of our total teaching. We can't get that part out of our minds. The Danielson rubric is very tangible; it has behavioral traits that you can follow, but growth on a test, not so much?

Here, the use of test scores, coupled with a perceived lack of control, causes higher stress and concern: "… makes us all very nervous … there's lot of nerves … It's very scary …" Intellectually, the teacher knows the Danielson observations ratings are weighted more highly than test scores but "… we can't get that part out of our minds…" There is a fixation on test scores, giving them more weight and importance than actually reified in the system ("…determine our fate as teachers…") There is teacher uncertainty about how the test score component will impact their evaluations over the long run and the comment suggests a mistrust in test scores as a suitable measure of competence: it is a snapshot in time, restrictive, and narrow.

Teacher summary (negative). The narrative provided by some teachers related to fear/stress suggest the following negative outcomes from the evaluation process: (1) the evaluation system attempts to inappropriately itemize/compartmentalize instruction for the purpose of measuring; (2) the evaluation structure forces conformity to the measured

attributes for the sake of job security; (3) the process leads to a narrowing of instructional creativity to what is measured and tested; (4) the evaluation harms professional identity and morale, making retention "good" teachers more difficult.

Principals (negative). Narratives provided by principals hold a unique view of the system's *negative* impacts. Principals expressed *Fear/Stress* as the main issue. However, from their perspective, *Fear/Stress* arise as a natural consequence evaluation. As such, this consequence is not seen as detrimental to the validity of the evaluation results or harmful to teacher's professional identity. Here, the new evaluation process provides an accurate representation of instructional competence. Those teachers displaying weakness in their practice will be properly identified and asked to address the shortcomings. In this way, teachers with previously unidentified weak practices should be concerned. Importantly, this perspective is different from the three other stakeholder groups which view negative aspects of the system as being inherently detrimental to the proper identification of good/effective teachers.

One principal speaks about the general level of stress caused by the evaluation process:

> I think they're [teachers] getting used to it [the evaluation process]. I don't know that anyone's ever comfortable. Even though I feel like my teachers are comfortable with me and I'm in all the time, they still freak out when they know it's their formal observation. They still don't like being observed like that. Even though I'm in their rooms five other times, the pressure they feel is on this one [observation], and even though I keep saying it's a culmination of everything, it's that one moment in time. I will say they're still not comfortable. (Principal 205)

The principal's comments suggest that teachers are not fully aware of the multiple sources of evaluation information utilized to arrive at final evaluation results; "…I'm in their rooms five other times… I keep saying it's a culmination of everything…" Even

though the principal conducts multiple walk-throughs, it is the weight teachers are assigning to the single full-lesson observation event that is creating the stress: "…they still freak out…; …the pressure they feel is on this one [observation] …; … that one moment in time…" However, this stress is not detrimental to the evaluation process. Rather, it is a natural reaction to being evaluated. The principal believes that " … they're [teachers] getting used to it… my teachers are comfortable with me…" Teachers, or anyone, simply "…don't like being observed like that… they're still not comfortable…" These remarks devalue the possibility of unintended consequences such as declining morale, commitment to the profession, or acceptance of evaluation results.

A second principal reflects that teachers have not expressed much concern about the new evaluation process. However, a natural *nervousness* is something that should be expected:

> Interviewer: Have your teachers given you much feedback about the evaluation system?
>
> Principal 208: No … they [teachers] haven't given me feedback about the process. I think it's [evaluation] just something that they're used to. [However] I think they have the natural nervousness like, "Oh my principal's coming in" or something like that, and "This is the formal. I want to do really well." They always, obviously, want to do their very best.

A third principal also sees the *fear/stress* caused by the evaluation process as a legitimate by-product for those teachers displaying less than acceptable instructional practices:

> I think that, by and large, when I get done with an evaluation, and a teacher sees "oh, that's where I'm at" I think they're pretty satisfied. There's a couple that probably feel like I ought to worry a little bit, and they should because it's not there.

> I don't think it's gonna catch anyone by surprise, but I think it will raise the level of concern. I think that's a good thing cuz, I think, then they'll start reading their rubric a little more. (Principal 207)

Here, the majority of teachers are both comfortable and satisfied with the evaluation process. However, a select few "…ought to worry a little bit…" because the quality of their instructional practices may be in question. Stating "… they should because it's not there…" indicates a confidence that the evaluation process will accurately expose weak areas. This also implies teachers acknowledge the accuracy of the process and thus will become concerned.

The reader is reminded that, for the first time, the district's new evaluation process produces a single summated assessment of overall teacher performance. This is expressed in the form of a classification label whose terminology is specified in legislation. The performance descriptors districts are required to assign are *Unsatisfactory*, *Developing*, *Effective*, and *Highly Effective* (Ariz. Rev. Stat. §15-203A.38, 2010). Interestingly, legislation does not provide operational definitions of these terms or specific analytic methods by which to identify them. In this regard, the same principal (below) believes that the transition to the new evaluation process and the assignment of these descriptive terms are a source of concern for teachers. The individual comments:

Interviewer:	Do you think that the teacher evaluation process has positively impacted your teachers' ability to be better teachers?
Principal 207:	Some of them, yes. Others, maybe not … I haven't come across anyone who just thinks "well, evaluation is dumb, and whatever happens, happens". I don't have anyone like that … [However], I have a feeling that once scores [labels] come out, then there will be a different story. It's gonna raise the level of

concern. Some are gonna go "oh, my gosh." Some are gonna go "wow."

Again, the principal downplays the significance of any fear/stress created by the evaluation process. Indeed, the person suggests that teachers are quite accepting of evaluation in general. The concern originates in the assignment of descriptive performance labels and whether those labels are interpreted to align with teacher's self-impression of competency. For example, a teacher that has previously been strong in a number of areas and relative weaker in others may not view him/her-self as being *Unsatisfactory* or *Developing* or even *Effective*. These terms have personal interpretations outside of the analytic method being used to derive them. The principal suggests that some teachers will disagree with the label stating "… some are gonna go 'oh, my gosh.' Some are gonna go 'wow…'"

While discussing aspects of new approach for quantifying and labeling teachers, a fourth principal comments:

Interviewer: Is that a concern for teachers?

Principal 202: I think so. I mean, there are just a lot of questions out there.

Interviewer: Is it this idea of the evaluation system coming out with some end label that's different than evaluations in the past?

Principal 202: Oh, yeah. What's my label? Am I gonna be highly effective? Previously, really, you'd go down [to the teacher] and say, "Hey, you're doin' a great job. You're doin' a great job," and it wasn't tied to anything. Now [they are] actually getting a label.

This principal makes an explicit comparison with former approaches for communicating teacher performance. Previous communications were based on relationship and dialog with the evaluator ("…Hey, you're doin' a great job…" There

was no summative label, ranking, or overall performance metric. The stress inherent in the new system is that "…now [they are] actually getting a label…" Arguably, this label is seen as a form of branding—assigning a highly reductionist term which is intended to summarize the complexity of instructional competency. The fear/stress stems from this loss of depth and descriptive understanding.

While discussing the labeling aspect of the new evaluation system, a fifth principal (203) agreed that teachers are concerned about this reduction to a single term stating "... I'm anticipating it [concern over the label]. I'll be honest with you. I'm anticipating it more…" This principal goes on to say:

> I think some of them are going to be a little concerned, a little concerned because, and again, this is just that different perspective on things. Having that we're inheriting these students every grade level, and we are being held accountable for the last however many years based on [test scores] … I am now inheriting a student and whatever it is they brought over the last eight to nine years. But again, I go back to if we are responsible for showing growth, which should not be of concern. If we can show that growth, we should be okay. There should be nothing to worry about. (Principal 203)

The narrative introduces another perceived source of stress for teachers – the use of test scores in the evaluation rating. This concern is coupled with a perceived inability to control for student's lack of prior learning attributed to teachers in earlier grade levels: "…I am now inheriting a student and whatever it is they brought over the last eight to nine years…" The principal argues that teachers believe their evaluation will be adversely impacted by the poor (i.e. ineffective) instructional practices of teachers in the lower grade levels (i.e., "…I am now inheriting a student…") However, at the same time the principal dismisses this as unfounded stating "…we are responsible for showing growth, which should not be of concern… there should be nothing to worry about…"

The addition of test scores as a source of teacher stress is echoed by another principal (below). As context, this principal's school was assigned a state accountability label of "D" based primarily on low achievement indicators on the state's most recent reading and mathematics test. The distinctions in the narrative are between test scores measured in terms of absolute performance (i.e., annual passing rates) versus growth measures which tend to nullify disparities based on student demographics and prior performance. In this context, the principal comments:

> For the teachers, I think, right now, it's still a little kind of stressing for them … I think that they, at the beginning, when they heard that [test scores] scores were coming to be part of [their] evaluation, it was a little bit of a shock to them … At first, I had teachers crying when they first heard that you're going to be ineffective if your [test] scores are bad. What they're hearing is - they immediately hear my effectiveness is gonna be based on [test] scores. Then they kinda shut off. They are nervous, I think.
>
> I really had to work with a couple of them to kind of, okay, get rid of the stress, that little switch that went off in your head that went "oh, my gosh". … When I was able to one-on-one talk to them and explain to them that they're [the District Office] adjusting for that [low prior achievement] and go through what you had shared with us, then it was like, ah. Okay.
>
> [However] If you're [teachers] doing what you say you're doing or what you need to be doing, there shouldn't be a problem … [For] some of them it [the evaluation system] will heighten it [stress], and if it does, it's because it's needed. (Principal 201)

Here again, the principal's perspective of teacher concern/stress is different from the teachers themselves. The perspective is that his/her teachers had misinformation and that once explained ("…they're adjusting for that…") the stress/concern becomes unwarranted. As with his/her colleagues, the stress caused by the evaluation process is legitimate for selected teachers because it will reveal the true level of their instructional competency: "… if it does, it's because it's needed…"

Principal summary (negative). Principals expressed limited concern for negative consequences of the evaluation system. For most, fear/stress was the primary concern of teachers. However, this concern was a natural attribute of being evaluated. Most people become anxious when face with evaluation.

Principals cited the reduction of performance down to a single composite rating as one source of teacher concern, including the assignment of state-mandated descriptive performance labels. A few principals cited the use of test scores in the evaluation process as an added source of teacher concern. However, for virtually all principals, the fear/stress/concerns raised by teachers were discounted. Indeed, principals viewed the evaluation process as providing accurate assessments of instructional quality. Based on this premise, they felt that selected teachers should become concerned because the new system would be accurately exposing areas of documented weakness.

District (negative). As discussed below, three out of the four district members shared generally negative reflections on the capability of the evaluation process to positively impact instructional practice. Two commented directly on the impact to teacher practices and one on its ability to change academic outcomes underlying the purpose of evaluation. Finally, one district member reflected a generally positive perspective of the evaluation system's ability to improve instructional quality.

Reflecting on the question of system impact, one district member comments:

Interviewer:	How do you think implementing our evaluation system is impacting teachers and their professional practice?
District 304:	It has a negative impact. Teachers are so concerned about that number part [quantitative rating] and the impact of student assessment results on their career that, again, it makes us focus

> on what we all know is not going to give us the best results for those students. My objection all along has been that we have taken this AIMS [test], sliced and diced it and thrown it into everything: now to give letter grades to schools, to define students, now we're defining teachers … We're stuck with the one thing because it's being mandated that we are stuck with that.

For this individual, the overall impact is emphatically negative. The perspective is based on the adverse influence of over emphasizing test scores saying "…it makes us focus on what we all know is not going to give us the best results…" In addition, there is concern about reducing instructional quality to a quantitative measure ("…that number part…") and how it may result in adverse, high consequence, impacts "…on their career…" The comment reveals a lack of trust in the evaluation results because of this over reliance on [AIMS] test scores: "… [we] sliced and diced it and thrown it into everything…" Finally, there is an objection to having a test score requirement externally imposed: "…it's being mandated that we are stuck with that…" It forces conformity to a less than adequate system of evaluation. The implication is that a test-centric evaluation misses important non-academic outcomes, fails to adequately represent instructional quality, and re-directs instructional activities to behaviors not reflective of good/effective instruction. The same individual goes on to say:

> I really have fear that it's having negative rather than positive. I think as a district we've done something right, and that is we've met with people. We've explained it. We've worked with them. We've tried to show them that at least we're using growth … we've done everything that we could in what I consider a negative kind of situation, a negative mandate that was given to us. We've done everything that we could to help people understand that we're gonna be supporting their success. But our reality is they're gonna have a number attached to them. I think that's a shame because I think that there are parts of education that are shunned, being left out, at the expense of [preparing] for the things that are on the test. (District 304)

In discussing his/her perspective, the person reveals frustration with trying to design and implement the best evaluation possible in the context of restrictive policy mandates: "… we've done everything that we could in what I consider a negative kind of situation, a negative mandate that was given to us…" However, the reality is that assessment of teacher competence is being reduced to a number, "…they're [teachers] gonna have a number attached to them…" As a result, important aspects of teaching are crowed out, "shunned," because teachers are forced to prepare for a test. Here, conformity to the system is harmful because the test is an inadequate representation of competence.

A second district participant sees the system as having a negative impact on teacher morale:

> People [teachers] are freaking out about it [release of the evaluation results]. I've had teachers say to me, "Well, I'm gonna leave because I don't want any part of this evaluation system. I'm goin' to another district where they don't do this." Oh, yes they do. This is state-wide, by the way, kind of a thing. (District 303)

The stress is caused by imposing a system that arguably doesn't effectively measure teacher competence. Imposing this measure causes teachers to consider leaving the district. Correctly, the person notes that all districts are required to implement systems inclusive of test scores. The harmful effect is that teachers nevertheless feel concern which negatively impacts relationships (i.e., culture, climate, morale). The same person goes on to share (Note: a portion of this exemplar was previously quoted):

> I would say Danielson, while it's one of the better models, it still narrows the scope of what people are looking for. My opinion of that is - good or bad - it has narrowed what teachers do in a classroom.
>
> I would argue that teachers are flat out teaching to tests and why wouldn't you? To be honest with you, if I was in the classroom today, given everything that our teachers are dealing with, I would absolutely be teaching to the test. Not a question in my mind, and I think that that's a shame, in that students suffer from

[not receiving] a broader perspective … Because these kids are gonna be tested in this, and that's gonna be part of whether or not I'm effective, whether I keep my job, whether I get paid. [District 303]

Interviewer: [Teachers are] actually giving it [test scores] more weight than our system gives it?

District 303: Absolutely. Yeah, I would absolutely agree with that statement because of the emphasis that's put on it, not only at a state level, but at a national level. The [school] accountability's 100 percent of the test score therefore it's a huge component in their [teachers] minds. I think that's the only thing they focus on.

Again, over emphasis on test scores leads to a narrowing of instructional focus with harmful consequences: "…teachers are flat out teaching to tests …" However, for this person, teacher's reaction is a rational response. The evaluation system is imposed on them, and it forms the basis of a high consequence outcome, "…why wouldn't you..." conform to its requirements? Teachers are recipients of the system, not designers. The argument is that they must conform. This leads to harmful impacts: narrowing instruction where "…students suffer from [not receiving] a broader perspective…" Interestingly, it is recognized that perception and reality are different. Regardless of the actual weight given to test scores in computing the final results, teachers believe that test scores represent the majority of the competency measure.

A third district participant (below) reflects on system impact from a broader perspective: the purpose of education and the goal of having all students reach content mastery. The logic is that if the purpose of education is student learning, and evaluation systems are implemented to ensure this outcome, then the impacts of such systems have been, at best, neutral. Embedded is the reoccurring view that emphasizing test scores as a primary outcome measure is insufficient/inappropriate.

> There's ample evidence to show that teachers do those things [adhere to the FFT components] and [yet] the spread [in achievement] continues. There's still your Falls Far Below." There is still your Approaching … We have never eliminated the bottom 50 percent of student achievement with the best evaluation system and expectations that has ever been created in the history of America. (District 302)

This comment questions the assumptions underlying teacher evaluation systems: Teacher evaluations reify all of the causal factors impacting student learning, adherence to system attributes leads to greater learning, and the act of evaluating inherently leads to improved professional practice. The reflection suggests that quality teaching is not simply represented by academic outcomes alone; there are other important non-academic goals characteristic of good/effective teaching. Thus, the impact of evaluation has been, at best, neutral—teachers may be classified as proficient, but this alone has not lead to high achievement for all students. The same individual goes on to comment:

Interviewer:	That's an interesting statement … Do I interpret your latest statement as saying, "Wow, we've been doing teacher evaluation for a long time, and, by virtue of the evaluation process, we haven't eliminated the achievement gap"?
District 302:	We're trying to use the evaluation process to solve the problem that 100 percent of the students are not achieving the highest score. We're putting a lot of responsibility on that assessment system, and it is much more than that.

This individual's narrative questions the premise that evaluation in and of itself leads to better student outcomes. For the person, the evidence suggests that this is not true. Something is misaligned, omitted, or remains unclear (i.e., "…it is much more than that…"). The argument being presented is that evaluation (in general) has not changed the distribution of student outcomes and is therefore insufficient. In this way, there is a suggestion is that evaluation systems are too narrowly defined; there is more to teaching

than test scores, evaluation systems that are test-centric will not realize the desired impact of improving outcomes. The person continues:

> There's billions of dollars of intervention money spent on education, to reach out to the underperforming or the underprivileged, or whatever may be causing that bottom 25 to 50 percent to not perform well, but since we're talking about the evaluation system, no, it's never gonna fulfill its intent.
> That being said, there is an expectation, outside of the 33 percent, that the performance part of a teacher's evaluation still accounts for a lot of academic type performance of a teacher; again, getting into the lesson planning; the lesson delivery; the evaluation part of student learning; reflection on student learning. It is still, very much, focused on student learning. Still, a large part of a teacher's expected job duties are not defined in learning. (District 302)

Here, the impact is no impact*:* evaluation has not led to desired results. The misalignment infers that test scores are the sole measure of instructional quality ("…Still, a large part of a teacher's expected job duties are not defined in learning…") and that evaluation is "…never gonna fulfill its intent…"

District summary (negative). District participants held generally negative views regarding the impact of the evaluation process. Harmful impacts were primarily associated with an over emphasis on test scores. This emphasis leads to increased stress, lower morale, and misdirected instructional focus. Important learnings and experiences are shunned and evaluation outcomes misrepresent teacher competency.

State (negative). Two of the three state participants reflected generally negative views on whether or not the evaluation system is improving the instructional capacity of teachers. The first individual cites the many simultaneous changes/initiative taking place in education. The multitude of mandates is overwhelming teachers with a harmful collective impact. The narrative from this individual is provided below:

> I think teachers are overwhelmed. If it was only that they [districts] were implementing a new teacher evaluation system, maybe it wouldn't be quite so bad.

> But the timing of the teacher evaluation system, with the school letter grades, with the third grade retention, with the common core content, [with] instructional shifts, [changing] strategies, and the potential change of AIMS [test] to something else, I think you put all of that together. Then the tying of all of this to money, to financial incentives, and the classification system, I think it's [all] having a negative impact on teachers.
>
> Interviewer: That's because the teacher evaluation system is embedded in all these other things that are happening at the same time and collectively that's just having a negative impact?
>
> State 402: Yeah. If we hadn't put in all of these other things, if all this was, we're going to revise the teacher evaluation system and give test scores some weight, I think people would have been okay with that.

For this person, the impact of the evaluation process is negative because it is being implemented along with so many other substantive education initiatives that directly impact teachers: a revised school/district accountability labeling system, legislation concerning third grade retention, a change in the state's curriculum standards, and transition to a new state assessment. Politically, these changes are increasingly being tied to issues of school funding, teacher compensation, and public perception of school quality. Teachers are recipients of these changes; they are burdened by the implied added responsibility of which they have no control. The person concludes that "…it's [all] having a negative impact on teachers…" The same individual connects these stressors to staff morale and the ability of schools to attract and retain high quality teachers:

> I think teacher morale … based on a small sampling of superintendents over the last couple of months, they all have vacancies. They have teachers who have left the profession within the first 14 weeks of school. Just up and gone … The vacancies were as a result of (1) not being able to fill it at the beginning of the year, (2) somebody leaving after they signed a contract, and (3) people just generally feeling put upon. I think it [the evaluation system] is having a negative effect for the people who are staying here, but it's also having a negative effect in terms of not getting people to come into the profession and claim that this is their career and stay here. (State 402)

Arguably from this narrative, an unintended consequence of the evaluation process is an inability to attract and retain teachers (i.e., "…they [districts] all have vacancies… teachers who have left the profession…"). For those that stay in the profession, this state official believes that the evaluation system acts to lower the morale of continuing teachers (i.e.,"…people just generally feeling put upon…"). Overall, the system is "…having a negative effect…" both on teachers "…who are staying…" in the profession and for "…getting people to come into the profession…"

The same person goes on to comment:

> The legislation says if you're ineffective two years in a row, or you lose your continuing status, or whatever, [then] you're classified as a probationary teacher. I think most people are gonna be careful as to how far they push those statutes so they don't get to that line. I think we're just creating a new widget. (State 402)

Here, a secondary unintended consequence is an incentive for districts to be very calculating in how/who they label as *Ineffective*: if fewer teachers are entering and/or remaining in the profession, districts will take care to retain their current pool teachers regardless of their instructional competency. Because the evaluation system is seen as externally imposed and having a negative impact overall, this person believes that districts will attempt to minimize the identification of poorly performing teachers.

The interesting phrase in this comment (above) is "…we're just creating a new widget…" The individual views the evaluation policy as a troublesome imposition on an already beleaguered education system that doesn't accurately assess instructional quality. In its current form, the state-imposed evaluation system is burdensome, inconvenient, and lacks validity. Consequentially, it makes conduct of education more difficult. The system

forces compliance but will not result in meaningful outcomes. This state member continues:

> … In many ways, it's moot. All this discussion is moot. This is what we are required to do, and, as compliant people that we are, and I'm talking about the educators, we are breaking our backs to make this work. We are doing everything we can to come up with an evaluation form that has all these little categories and all these little rubrics … We're going through all of this, and we've put all of this effort, and I just question the yield from that effort. (State 402)

There is a sense of frustration, helplessness, and unwarranted imposition in the narrative. The good-faith effort being put forth by districts greatly exceeds the benefits (validity, impact, outcome) realized from the policy.

A second state member echoes this general lack of efficacy. Here, the evaluation framework is being implemented incorrectly and is emphasizing the wrong components. As a result, school culture and climate are harmed, negatively impacting morale, and lessoning chances of meaningful improvements. The individual begins the discussion by reacting to an initial prompt:

> Interviewer: It's our third year of the evaluation policy, but the first year of its implementation. Do you think it's making a difference, or will make a difference, consequentially? Will it improve teaching?
>
> State 401: I think it's gonna have some negative effects, and I don't know how much, so my sense is that very few are getting it right, and it's taking teachers' energy away, it's discouraging them a little bit, it's leaving them a little less energized and a little bit less positive for their students. To me, it's a tragedy.

The individual is direct in his/her response by stating "…To me, it's a tragedy…" However, the premise is that districts are implementing the evaluation process improperly. It is not that the system itself is improperly defined. An unintended consequence is that it is "…taking teachers' energy away…", lowering morale, and

damaging professional motivation. As a result, students are negatively impacted because the learning environment created by teachers is "…a little bit less positive…" The comment suggests that there is a better approach to teacher evaluation than currently practiced by Arizona districts: "…very few are getting it right…" The interviewee goes on to say:

> I think everybody [districts] has relatively simple [evaluation] models in mind in education, and in my view, those models have traditionally been, for the most part, fairly destructive. A few principals can carry it out, because they are highly skilled personalities and they can sense how to keep the whole thing balanced, but it's rare.
>
> The worst managed ones [evaluation approaches], they're having the manager do all the evaluations, and if just my manager is evaluating me, now all of a sudden, it changes my relationship with my peers. To a certain extent, it creates negative interdependence, so it's unidirectional. I'm very concerned that almost all of our schools in Arizona have gone to unidirectional evaluation of teachers. (State 401)

For this individual, the primary concern is implementing "…simple [evaluation] models…" Simplicity fails to capture complexity; simplicity results in an inappropriate weighting of key attributes: "…to keep the whole thing balanced…" Simplicity leads to systems that are "…fairly destructive…"

Only a select few principals are able to implement the system properly. For this person (above), the culprit is the principal-as-evaluator model of evaluation: "…the worst managed ones…" Principals that serve as both mentor and evaluator acts to harm relationships, school climate and culture, creating "…negative interdependence…" by altering the "…relationship with my peers…" A principal-as-mentor relationship permits two-way communication, dialog, reflection while a principal-as-evaluator imposed a unidirectional relationship: "… almost all of our schools in Arizona have gone to unidirectional evaluation of teachers…" The principal-as-evaluator model is also

criticized for the substantive amount of effort/time required to conduct the process, stating:

> It's overloading principals. In my view, the place I want my principal is not in his office doing paperwork, doing evaluations—I want him out in the parking lot in the morning, in the afternoon greeting parents, talking to them, assessing the situation. Letting them know that they're creating that positive interdependence and welcoming of the parent, knowing that this school belongs to them because the principal's out front greeting people. (State 401)

There is an implied trade-off in time and priority. Inefficiently implemented evaluation models take away from the principal's role as school leader, responsible for building community, relationship, and trust. The negative impact of the current system is a reduction of efficacy in these areas.

State summary (negative). State participants provide generally negative reflections on the impact of the evaluation system. Concerns included reduction in teacher morale, increased inability to attract and retain quality teachers, reduction in school climate/culture, excessive time requirements placed on administrators that detract from their ability to address other important leadership duties.

Summary – negative impacts. Teachers shared a number of substantive concerns regarding the overall impact of the evaluation process. The difficulty/inability to reduce instructional activities down to a small set of measured criteria raises concern and trust issues. Reductionism leads to a narrowing of instructed content, restricts instructional creativity, and forces conformity to attain favorable evaluations and protect job security. The interaction of reductionism, conformity, lack of trust, and fear/stress acts to harm the professional identity of teachers, lowers morale, and impacts the organization's ability to retain/attract of quality educators.

Principals expressed limited concern over issues of *fear/stress* and *conformity/reductionism*. They are viewed as natural by-products of the evaluation process. Indeed, *fear/stress* leads to increased focus on important instructional behaviors resulting in improved professional practice. In contrast to teacher, district, and state stakeholder views, principals believe conformity improves practice because the components of the evaluation system represent the essence of quality teaching.

Most district members shared negative views on the impact of the evaluation system citing test scores as a primary concern. Test scores raise stress/fear, harm teacher morale, and act to narrow instruction due to conformity. Test scores are generally seen as an inappropriate proxy for instructional quality resulting in a lack of trust in their use as a primary evaluation metric. For district members, their inclusion forces conformity and therefore crowds out other important aspects of teaching. There is a view by this group that evaluation, on its own, will not lead to improve practice or an increase in student learning.

State members also expressed generally negative views on the impact of the evaluation system. Issues included a reduction in teacher morale, increased stress, and harm to school climate and culture. In addition, the time spent on evaluation is excessive and limits the ability of administrators to engage in other important leadership activities. The impacts imposed by evaluation make it more difficult to retain/attract quality educators.

Petite assertions: RQ1C (a). *In what way has implementation of the teacher evaluation system affected the PP of classroom teachers (Instruction, student learning, professional capacity building, job satisfaction, etc…)?*

Petite assertion #1. Clarity/Focus/Structure permits modification and alignment of instructional practices. The new evaluation system provides *Clarity, Focus, and Structure* regarding the expectations and behaviors required to attain a favorable rating. The organized structure of the system and process permit teachers to modify their activities to align with these expectations and behaviors.

Petite assertion #2. Adherence/Conformity to the system's Clarity/Focus/Structure is a theme in the narratives. The evaluation system is imposed. That is, teachers are recipients of the system, not designers, collaborators, or developers. Instructional modification may be seen as adherence, conformity, and/or compliance. Teachers adhere/conform to the system components in order to ensure favorable evaluation outcomes and to ensure job security. Some see adherence as neutral/positive because the system's components are generally believed to reflect aspects of good instructional practice. In contrast, others see adherence as compliance. Here, teachers represent their evaluated behavior as different from their day-to-day practices. In this way, adherence, conformity, and/or compliance represent a negative impact.

Petite assertion #3. *Clarity/Focus/Structure* supports Communication, Dialog, and Reflection. Due to the well-defined and organized structure of the evaluation process, teachers are afforded increased opportunity to have focused communication and dialog with administrators regarding the measured instructional components.

Petite assertion #4. There are a number of negative effects associated with *Clarity/Focus/Structure*: narrowing of content, loss of instructional creativity, lower morale, harm to professional identity, and forced conformance/compliance. The evaluation components are viewed as incomplete. This reductionism raises trust concerns

leading to an increase in stress/fear. The impact is a narrowing of content and a loss of instructional creativity. Forced conformity/compliance harms morale and the instructional identity of teachers.

Petite assertions RQ1C(b). *Do the perspectives of efficacy and system affect differ across stakeholder groups (teachers, principals, and policy makers)?*

Petite assertion #1. Principals express substantively different, and more positive, perspectives on impact than teachers, district, or state members. For principals, *Clarity/Focus/Structure* is equated to concepts of evidence, objectivity, and reliability. This permits feelings of comfort, confidence, and assurance in the evaluation process and results. Structure leads to consistency, and objectivity leads to accuracy/fairness. Principal competence as evaluators is improved due to this clarity and objectivity. Principals believe the evaluation components adequately reify high quality instruction. Thus, distinctions between adherence, conformity, and compliance are devalued because all improve the professional practices of teachers. Principals view stress/fear as natural by-products of any evaluation process. In this regard, stress/fear leads to increased focus on the measured evaluation criteria which, in turn, improve instructional efficacy. Principals reflected that the evaluation structure improved their own practice as evaluators because it provided clarity and structure on important. By being better evaluators, teacher practices are similarly improved.

Petite assertion #2. District and state participants expressed generally negative perspectives regarding evaluation impact. Concerns from district participants focused primarily on the negative impact of using test scores as a primary indicator of instructional quality. Their use leads to increased stress/fear, lower morale, and a

narrowing of curriculum and instruction. As a result, important learning and experiences are crowded out. There is a general skepticism that test-centric evaluation systems are able to positively improve teacher practice. State members were equally negative citing lower teacher morale, inability to attract/retain quality teachers, a reduction in school climate and culture and the excessive time required by administrators to complete the evaluation process on their campuses.

Positive impacts cited by the district and state members included increased focus on issues involving teacher evaluation – improving organizational dialog, communication, and reflection. For these groups, a formalized evaluation system brings clarity and structure to the issue.

Research Question 1D: Reliability Evidence (RQ1D)

RQ1D. What are the reliability indices for the PP and VAM scales used to form measures of TIQ? The approach is scale and sub scale reliability indices, correlation, measurement error, tests of normality. The measures are PP scale and sub-scale item correlations, tests of VAM regression assumptions, prediction error (SEM).

This primary research question investigates attributes of the data/scales upon which inferences of instructional competence are made. Under the teacher evaluation framework, two sources of data (VAM and ratings of PP) are used to construct a composite score which is then interpreted as a measure of instructional quality. To examine the inferential validity of this (composite) measure, it is necessary to examine the reliability of both sources of data.

Specifically, reliability evidence informs on the stability (consistency) of measures obtained from repeated testing **(**AERA et. al, 1999, 2014; Thompson, 2003;

Wainer & Brown, 1988). In this regard, data reliability is concerned with attributes of measurement precision, score variation, and error, all of which are dependent upon the nature of inter-item, item-score, and test-retest correlations **(**AERA et. al, 1999, 2014; Thompson, 2003; Wainer & Brown, 1988). Data that are highly reliable are invariant across repeated sampling and are accurate/precise in their measure of the true population scores. That is, high reliability implies that the observed data embody low degrees of measurement error and thus accurately represent attributes present in the population. This concept is expressed in classical true score theory by the relationship of *Observed Score = True Score plus Error* (Carmines & Zeller, 1979; Thompson, 2003; Crocker & Algina, 1986). As measurement error decreases the observed score more closely approximates the true score. Reliability estimates are therefore concerned with informing on the size and direction of measurement error and the data's ability to accurately represent the construct of interest—in this case Teacher Instructional Quality.

Thompson (2003) states that "…it is important to evaluate score reliability in *all* studies, because it is the reliability of the data in hand in a given study that will drive study results…" (p. 5). Indeed, poor levels of reliability directly reflect on the data's ability to adequately inform on the construct of interest as well as the study's ability to yield "noteworthy" effects in the form of both statistical and practical significance (Thompson, 2003). Finally, as stated in the 1999 issue of the *Standards for Educational and Psychological Testing*, "… it should be the goal of test developers to investigate test reliability as fully as practical considerations permit. No test developer is exempt from this responsibility…" (AERA et al., 1999, p. 27).

Reliability may (in part) be affected by two sources of error: random and/or systematic (Crocker & Algina, 1986). Systematic errors "… consistently affect an individual's score because of some particular characteristic … that has nothing to do with the construct being measured…" (Crocker & Algina, 1986, p. 105). It is very difficult to isolate and measure systematic error since its origin is external to the measurement activity. In contrast, random errors "… affect an individual's score [either positively or negatively because of purely chance happenings…" (p. 106). The concept of chance (probability) becomes discernable under conditions articulated by measurement theory including sampling design, scoring, and statistical testing procedures (Allen & Yen, 1979). For the latter, reliability indices may be assessed through a number of empirical approaches depending on the type of data (scale) and the context under which the data is constructed (Crocker & Algina, 1986).

Reliability is directly aligned with the concept of validity. Validity (i.e., construct validity) is concerned with the suitability of inferences being made from observed data (Thompson, 2003**;** AERA et. al, 1999, 2014; Messick, 1989a). Thus, validity is not an attribute of the data themselves. Rather, validity is a property of the judgments, conclusions, and claims made on/from data. It follows that in order to be able to make appropriate inferences from data, the data must first be reliable (consistent in measure with relatively low amounts of error). Thus, reliability of measure becomes a necessary, but insufficient, condition for claims of inferential validity (AERA et. al, 1999, 2014; Wainer & Brown, 1988; Messick, 1989a).

The following sections discuss reliability for each component used in construction of the teacher evaluation composite score: PP scores and VAM scores.

Reliability measures of PP ratings. All PP ratings used to construct teachers' final evaluation scores are resultant from a single observational event, a single evaluator assessing the competence of a single teacher on each behavioral component. While it is true that the ongoing evaluation process incorporates a series of informal (5-10 minute) classroom walkthroughs, pre-post conferences, and reflections on a variety of instructional artifacts/documents, the actual assignment of component PP ratings is a singular event much like that of a student taking an achievement test on a given day and time. In this context, reliability indices based on measures of internal (item) consistency are appropriate (Crocker & Algina, 1986; Carmines & Zeller, 1979; Allen & Yen, 1979)[21]. Specifically, adaptations of Cronbach's coefficient alpha, standard errors of measure, and principal components analysis are employed to assess the reliability of the rating scales.

PP reliability under conditions of data ordinality. It has been previously discussed that the PP data in this study are measured using a four-point ordinal (Likert) scale. In addition, the data show a pronounced negative skew and cannot be considered distributed as multi-variate normal. Reliability indices that rely on computation of linear correlations under assumptions of continuous, multi-variate normal, scales may be biased. Gadermann, Guhn, and Zumbo (2012) note that considerable technical debates have been ongoing in the published literature regarding use and interpretation of reliability indices under such conditions. The authors state that "… [these] technical debates are … critically relevant to practitioners and researchers in the social sciences in

[21] For a complete treatment of reliability theory and related measures of internal consistency, the reader is referred to chapters 6 and 7 of Crocker and Algina's (1986) *Introduction to Classical and Modern Test Theory*. In addition, Carmines and Zeller (1979) present a cogent discussion of the application of Principle Components Analysis to issues of reliability and validity.

general. In fact, using Cronbach's alpha—or any other reliability coefficient—under circumstances that violate its assumptions and/or prerequisites might lead to substantively deflated reliability estimates. A substantively deflated estimation of a test's reliability, in turn, might potentially entail some misinformed inferences, such as discarding a test due to its seemingly low reliability…" (p. 1-2). This reflection is made with an understanding that Cronbach's coefficient alpha already represents a lower bound estimate of true test reliability using even the best-behaved data (Crocker & Algina, 1986, p. 121; Carmines & Zeller, 1979, p. 45).[22]

For this reason, this study adopts the three methods for examining reliability of the PP measures. First, Cronbach's alpha of internal consistency is replaced with estimates of ordinal alpha (Zumbo, Gadermann, & Zeisser, 2007; Gadermann et al., 2012; Bonanomi, Ruscone, & Osmetti, 2013). Ordinal Alpha is computed from a polychoric correlation matrix rather than the Pearson covariance matrix (Bonanomi et al., 2013).[23] Gadermann et al. (2012) note that when the data are ordinal and skewed, alpha indices based on the Pearson correlation are understated and should be replaced by estimates based on polychoric correlations:

> It has been shown that the Pearson correlation coefficient [r_{ij}] severely underestimates the true relationship between two continuous variables [i, j] when the two variables manifest themselves in a skewed distribution of observed responses. A tetrachoric/polychoric correlation, on the other hand, more accurately estimates the relationship of the underlying variables … The non-ordinal coefficients focus on the reliability of the observed scores by treating the observed item responses as if they were continuous, whereas the ordinal

[22] Cronbach Alpha assumes that all possible combinations (split-half's) of items represent perfectly parallel subtests with identical error. This assumption is usually without merit and leads to alpha being interpreted as a *lower bound* estimate under ideal testing conditions (Crocker & Algina, 1986, p. 139).

[23] For Cronbach's alpha, the Pearson covariance matrix is routinely used; for example, as a default in statistical software programs, such as SPSS and SAS. An important assumption for the use of Pearson covariances is that data are continuous, and if this assumption is violated, the Pearson covariance matrix can be substantively distorted (Gadermann et al. 2012, p. 2)

> coefficients focus on the reliability of the unobserved continuous variables underlying the observed item responses. In this way, the ordinal coefficients are nonparametric reliability coefficients in a nonlinear classical test theory sense. (Gadermann et al., 2012, p. 2-3)

To compute the polychoric ordinal alpha, the Mplus version 7.11 was first used to generate the polychoric correlations between FFT rating items applying a weighted least squares mean/variance- adjusted (WLSMV) estimator to account for the ordinal scaling and non-multivariate normal characteristics of the PP data (Muthen & Muthen, 2012). Following the formula provided by Gadermann et al. (2012), Ordinal alpha is computed as $\acute{\alpha} = \frac{k * \bar{r}}{1+(k-1)* \bar{r}}$ where $\acute{\alpha}$ distinguishes ordinal alpha from Cronbach's alpha, k is the number of components in the PP domain and $\bar{r}$ is the average polychoric correlation between the component scores.

Second, ordinal theta reliability coefficients are constructed for each of the FFT sub-scales using an exploratory factor analytic approach (Carmines & Zeller, 1979; Muthen & Muthen, 2012; Green, 1977; Zumbo et al., 2007; Li & Wainer, 1997).[24] Mplus Version 7.11 was employed to conduct an ordinal scale-adjusted EFA analysis using a weighted least square means and variance adjusted estimator (WLSMV).[25] This procedure generates the proper variance components (i.e., eigenvalues corrected for categorical scaling) needed to compute ordinal theta reliability coefficients (Zumbo et al., 2007).

---

[24] Carmines and Zeller (1979) and Zumbo et al. (2007) utilized Principal Components Analysis to compute Ordinal Theta. For this study, algorithms available in SPSS do not adequately accommodate issues of ordinal scaling and non-normality when conducted factor analysis. In addition, Mplus properly construct EFA from polychoric correlation tables but does not include Principal Components Analysis routines. For this study, it was decided to use EFA common-factor analysis on the polychoric correlation within Mplus. Since common factor approaches replace the diagonal terms in the correlation matric with 8estimates of common variance (r-square), the resulting eigenvalue estimates will be smaller than those ideally reported under PCA. In this way, the theta estimates presented herein represent a lower bound to the theta values than those reported in the referenced literature.

In this approach, EFA estimates variance attributable to each extracted component based on the item correlation matrix of the measured data (Pett et al., 2003). Under assumptions of unidimensionality, it is expected that the pattern of extracted factors identifies a single dominant factor. This primary factor should account for the majority of the observed data's variability while remaining components should account for minimal variance. In addition, the item loadings on the primary component should be substantively larger compared to their loadings on the remaining components. Interpretatively, components with relatively small eigenvalues represent error (residual variance) that is unrelated to the primary latent construct. In this way, EFA provides a means to explore the reliability of the measures by assessing the amount of explained variance associated with the primary component in much the same way that coefficient alpha depends on the strength of correlations among all scale items (Armor, 1974; Carmines & Zeller, 1979; Zumbo et al., 2007; Li & Wainer, 1997).

Coefficient theta utilizes the eigenvalues (variance account) of the first primary factor to compute a reliability coefficient for the scale (Armor, 1974; Zumbo et al., 2007). This takes the form of Theta = $\frac{N}{N-1}$ * (1 - $\frac{1}{\lambda_1}$) where N equals the number of items in the scale and λ_1 is the largest (first) eigenvalue of the extracted factor (Armor, 1974; Carmines & Zeller, 1979; Zumbo et al., 2007; Li & Wainer, 1997). As a measure of internal consistency, Theta may be shown to be a special case of coefficient alpha (Armor, 1974). The important distinction is that coefficient alpha constitutes an estimated lower bound of scale reliability while Theta represents a maximized upper bound (Zumbo et al., 2007; Li & Wainer, 1997).

The third method for assessing scale reliabilities is to estimate the standard errors of measurement on each of the FFT domain subscales and the overall PP scores (Crocker & Algina, 1986; Allan & Yen, 1979). These standard errors are then used to construct confidence intervals to inform on the range of error present in the PP scores used in the evaluation process. Expressed as confidence interval, the precision of the evaluation scores are interpreted based on the range of possible true scores. Allan and Yen (1979) note that

> … the main advantage of the use of confidence interval [in reliability analysis] is that they make it clear that an observed score contains a certain amount of measurement error. If the test score is unreliable, confidence intervals for true scores will be very wide, indicating that observed scores are not good indicators of true scores … (p. 90)

For this study, the range of confidence interval is proportionally compared to the range of the total scale to provide a common comparative metric of the measurement error.

PP reliability indices – ordinal alpha. Ordinal alpha reliability indices computed for each of the FFT domains and the composite PP score are provide in Table 56. Included is a comparison to the unadjusted Cronbach alpha reliability coefficient computed from the raw rating data using the Scale Reliability module in SPSS Version 21.0. The supporting Mplus Version 7.11 PP polychoric correlation table is provided in Appendix I.

Table 56

Ordinal Alpha Reliability Estimates for FFT Evaluation Domains

PP Domain	No. of Components	Average Polychoric Correlation	Estimated Ordinal Alpha	Cronbach's Alpha
Domain 1: Planning & Preparation	6	0.721	0.939	.856
Domain 2: Classroom Environment	5	0.654	0.904	.810
Domain 3: Instruction	5	0.655	0.905	.816
Domain 4: Professional Responsibilities	6	0.683	0.928	.854
Composite PP Score	22	0.643	0.975	.949

As shown, the ordinal alpha reliabilities associated with each PP domain and the composite score exceed .90 indicating high degrees of internal consistency. Gadermann et al. (2012) interprets the size of ordinal alpha by stating:

> … typically, the psychometric literature recommends that alpha for a scale should not be smaller than .70 when used for research purposes, at least .80 for applied settings, and greater than .90 or even .95 for high stake, individual-based educational, diagnostic, or clinical purposes…(p. 5)

Using these criteria, the reliability data reported for the PP scores seem to satisfy the requirement for use in teacher evaluation.

Including the comparative (unadjusted) Cronbach alpha reveals the difference between the two approaches. Per Gadermann et al. (2012), the Cronbach alpha for each of the subscales is substantively less than the polychoric ordinal alpha estimate reflecting both the impact of a relatively small number of components and the assumption of continuous multivariate normality. The discrepancy lessens on the composite PP score

reliabilities because of the large increase in the number of components (increasing from 5-6 to 22) revealing more sensitivity in Cronbach alpha indices to changes in the number of test items.

Finally, interpreting the unadjusted Cronbach alpha might lead researchers to suggest that the PP domain sub-scales may not be appropriate for use under high-consequence conditions (Gadermann et al., 2012). However, values of ordinal alpha might provide more confidence in their use in teacher evaluation decisions.

PP reliability indices – coefficient theta. As discussed above, indices of ordinal theta reliability are computed based the eigenvalue revealed by the most dominant factor extracted from the PP data's polychoric correlation matrix (Armor, 1974; Carmines & Zeller, 1979; Zumbo et al., 2007). Table 57 reports coefficient theta reliabilities based on eigenvalues estimated by Mplus.[26]

[26] When indicator variables are specified as ordinal (categorical), the WLSMV estimator first computes a sample correlation matrix (tetrachoric, polychoric) and then fits the model to that, thereby estimating [all] the model parameters (http://www.statmodel.com/discussion/messages/23/646.html?1341197843)

Table 57

Ordinal Theta Reliability Estimates for FFT Evaluation Domains

PP Domain	No. of Components	Dominant Component Eigenvalue	Estimated Ordinal Theta
Domain 1: Planning & Preparation	6	4.607	0.940
Domain 2: Classroom Environment	5	3.621	0.905
Domain 3: Instruction	5	3.624	0.905
Domain 4: Professional Responsibilities	6	4.423	0.929
Composite PP Score	22	14.548	0.976

As with ordinal alpha, the data indicate strong ordinal theta reliabilities for each of the PP sub-scales and the overall composite scores. Indeed, ordinal alpha and ordinal theta are very close in magnitude.

PP reliability indices – standard error of measure. In this approach, estimates of ordinal alpha reliability and scale standard deviations are combined to construct true-score confidence intervals in order to inform on the size/magnitude of the measurement error inherent in the PP rating data (Crocker & Algina, 1986). The general form for constructing a 95% confidence interval around an examinee's score is X ± 1.96 $\widehat{\sigma_M}$ where $\widehat{\sigma_M}$ is the standard error of measure. In turn, the standard error of measure is estimated by the equation $\widehat{\sigma_X}\sqrt{1 - \widehat{\rho_{xx'}}}$ where $\widehat{\sigma_X}$ is the standard deviation of observed scores and $\widehat{\rho_{xx'}}$ is the estimated score reliability. In this form, the interval is interpreted as being 95% confident that the examinee's true score lie in the interval X ± 1.96 $\widehat{\sigma_M}$. Table 58 and 59 reports the confidence interval information for each of the PP domain

scales. All CI estimates are constructed using the mean PP score of the observed measures.

Table 58

FFT Scale Descriptive Statistics

PP Domain	No. of Components (Total Points)	Scale Mean	Scale Standard Deviation	N
Domain 1: Planning & Preparation	6 (18)	12.782	2.4552	238
Domain 2: Classroom Environment	5 (15)	11.286	2.0027	238
Domain 3: Instruction	5 (15)	10.634	2.0676	238
Domain 4: Professional Responsibilities	6 (18)	13.164	2.4689	238
Composite PP Score	22 (66)	47.866	8.2241	238

Table 59

Ordinal Alpha Reliability Estimates for FFT Evaluation Domains

PP Domain	Ordinal Alpha	SEM	CI Lower Bound	CI Upper Bound	CI Range	CI Scale Proportion
Domain 1: Planning & Preparation	0.939	0.606	11.593	13.971	2.377	13%
Domain 2: Classroom Environment	0.904	0.621	10.070	12.502	2.432	16%
Domain 3: Instruction	0.905	0.637	9.385	11.883	2.498	17%
Domain 4: Professional Responsibilities	0.928	0.662	11.866	14.462	2.597	14%
Composite PP Score	0.975	1.300	45.317	50.415	5.097	8%

For each of the sub-domain PP scores, the size of 95% true score confidence intervals (CI) range from 2.4 to 2.6 points, or between 13 and 17% of their respective point scales (*M* = 15%). Comparatively, the size of the 95% CI for the composite PP score is 5.1 points (twice the amount compared to the sub-scales) but proportionally only 8% of the total composite scale—a reduction of approximately 47%. That is, while the absolute range of the true score CI is larger for the composite scale, the relative magnitude of the error is substantively reduced. Thus, the composite score may be seen to be more precise than any of the sub-scale scores by a factor of approximately two. In this way, inferences made from an examinee's sub-scale score are more tenuous than those made from his/her composite PP score.

VAM reliability indices. The second primary component of the teacher evaluation composite score concerns measures of student academic growth based on multi-level value-added models (VAM). Here, estimates of measurement error reflect the degree that VAM models are able to predicted student outcomes with precision. In this regard, VAM model reliability indices are constructed to inform on the degree of measurement error present in the predicted estimates. The reliability measures explored in this study include the following: (1) proportion of variance explained (PVE), (2) reduction in the intraclass correlation coefficient (ICC), and (3) model standard errors of estimate (SEE) including construction of true-score confidence intervals (CI) and error range statistics re-expressed in terms of absolute number of test items.

VAM reliability indices – proportion of variance reduction. Estimation of multilevel models does not permit computation of explained variance (R^2) in the same manner as single level regression methods such as ordinal least squares (OLS). The

reason is that variance decomposition under multi-level, multi-equation, model specifications are more complex and multi-faceted than computed from single-equation (OLS) models (Raudenbush & Byrk, 2002; Snijders & Bosker, 1999; McCoach & O'Connell, 2012). Essentially, multilevel data requires specification of predictive equations for each level of nesting, each containing its own error term. Model variance estimates are composed of linear combinations of these error terms, preventing computation/interpretation of a single OLS-like R^2 reliability measure (Luke, 2004).

To address this problem, a number of pseudo r-square values have been proposed. For this study, the proportion of variance explained (PVE) proposed by Raudenbush and Byrk (2002) and McCoach and O'Connell (2012) is utilized as one measure of model precision. This approach compares estimated variance components produced by fully unconditional and conditional model specifications. A fully unconditional model represents the simplest form of a multilevel specification and is equivalent to a one-way analysis of variance (ANOVA) model with random effects devoid of any explanatory (predictor) variables (Luke, 2004). A conditional model incorporates the desired set of predictors at each level of nesting. The reduction in variance estimates attributed to the addition of predictor variables serves to inform on the explanatory power of the conditional model.

For this study, of particular interest are the level-1 reliability estimates of student academic achievement after correcting for the multi-level interdependencies caused by students nested within schools. The Level-1 PVE index is computed as $\frac{\sigma^2_{Uncoditional} - \sigma^2_{Conditional}}{\sigma^2_{Unconditional}}$ where $\sigma^2_{Unconditional}$ is the level-1 residual variance of the

unconditional model and $\sigma^2_{Conditional}$ is the level-1 variance in the conditional model. Table 60 reports the PVE indices for each grade/subject VAM model used to compute 2013 student residual growth measures.[27]

Table 60

PVE VAM-Model Indices by Grade by Subject

Reading - Residual Variance*				Mathematics - Residual Variance*		
Unconditional Model	Conditional Model	PVE Ratio		Unconditional Model	Conditional Model	PVE Ratio
1957.120	522.316	0.733	Grade 3	2425.814	712.428	0.706
1717.472	441.746	0.743	Grade 4	2353.830	644.858	0.726
1667.844	442.476	0.735	Grade 5	1934.310	493.857	0.745
1686.353	412.197	0.756	Grade 6	1934.546	452.611	0.766

* All variance estimates significant at $p < .001$.

The data indicate that the conditional model specifications for both reading and mathematics accounting for between 71 and 77% of the level-1 (student) variance.

A critical question is whether this amount of explained variance is sufficient to warrant making high-stakes, consequential decisions based on the model's outcomes. Arguably, having between 25 to 30% of (error) variance left unexplained might raise some concern. One on hand, Goldberger and Hansen (2010) cite (VAM) R^2 values ranging between .72 and .73 in their reflection on the suitability of using VAM models to make high stakes teacher tenure decisions (they concluded such models are suitable to the

[27] The same model specifications were estimated on 2012 achievement, resulting very similar variance estimates. 180 (75%) of the 238 Group A teachers had two years of VAM information. Differences were due to changes in employment status and/or position/school.

task). In contrast, Gadermann et al. (2012) argue that in high stakes decision making reliability (error) indices need to reflect substantively lower amounts of measurement error with r > .90 (implying explained variance ratios of r^2 > .81). In addition, there is considerable literature implicating the stability (reliability) of VAM models suggesting substantively high reliabilities are required for consequential decision making (Wayne, 2010; Papay, 2011; Newton et al., 2010; Darling-Hammond, Amrein-Beardsley, Haertel, Rothstein, 2012; Au, 2010; Amrein-Beardsley & Barnett, 2012).

VAM reliability indices – intraclass correlation coefficient (ICC). The intraclass correlation coefficient (ICC) provides a measure of the amount of variance in the dependent variable (student achievement) attributable to the grouping (schools) variable (Luke, 2004; Raudenbush & Byrk, 2002; McCoach & O'Connell, 2012). When membership in a subgroup influences values in the dependent variable the assumption of independence (that all observations are independent and identically distributed – iid) is violated and the standard error estimates become understated (McCoach & O'Connell, 2012). If the model's standard errors are biased downwards, associated reliability indices are also biased possibly leading to incorrect inferences and conclusions. For this reason, it is important to measure the amount of variance attributable to (2-level) group membership and the degree to which the final (consequential) model specification has minimized this effect.

For a two-level nested model, the ICC is computed as the 2-level error variance divided by total model error variance. For clarity, first consider the specification for a fully unconditional 2-level model (equating to a one-way ANOVA with random effects). The 1-level equation becomes $Y_{ij} = \beta_{0j} + r_{ij}$ where Y_{ij} is the test score for the ith

student located within the j[th] school, β_{0j} is the intercept term for the equation representing the mean test score for school j, and r_{ij} is the random 1-level error term for student i within school j. For the 2-level specification, the intercept term β_{0j} is presumed to vary across schools (that is, each school has a different mean test score due to differences in school characteristics). Thus, the 2-level equation depicts this condition as $\beta_{0j} = \gamma_{00} + \mu_{0j}$ where γ_{00} is the grand mean of test scores across all schools and μ_{0j} is the random 2-level error term for school j. For estimation purposes, the two equations may be combined as $Y_{ij} = \gamma_{00} + \mu_{0j} + r_{ij}$. This is the equation estimated via multilevel modeling techniques in SPSS or HLM. More complex representations of this basic form are constructed when additional covariates (predictor variables) are added to each of the level equations.

From the fully unconditional (no predictor) model, it is easily seen that estimated values of the dependent variable (student test scores) Y_{ij} is now a function of two error terms, μ_{0j} (the 2-level school effect) and r_{ij} (the 1-level student effect). Further, in that γ_{00} is a fixed amount (the grand mean of all test scores), the variance associated with the estimated student scores Y_{ij} is Var (Y_{ij}) = Var ($\mu_{0j} + r_{ij}$) = Var (μ_{0j}) + Var (r_{ij}) = τ + σ^2 where τ and σ^2 are the 2-level and 1-level error variances. The 2-level ICC may therefore be computed as ICC = $\frac{\tau}{\tau+\sigma^2}$, its value indicating the amount of variation in the student test scores accounted for from student membership within schools. If this value is *low*, little influenced is excreted from within-school correlations and the model's standard errors are not substantively biased. If the ICC is high, within-school influences are substantive and the model standard errors become increasingly understated.

Because the fully unconditional model does not attempt to account for any student- or school-level characteristics, the ICC from this specification provides information on the degree to which the student scores are uniquely influenced by the within school correlations. In turn, comparison of the ICC values after inclusion of predictor variables permits comparative analysis of error variance reduction – i.e., a measure of model fit and reliability.

ICC for both the unconditional and conditional VAM models in Reading and Mathematics for grades 3 to 6 were computed. Table 61 reports the ICC values obtained along with the variance reduction attributable to the inclusion of the predictor variables.

Table 61

ICC Variance Components for Unconditional and Conditional VAM models

	Unconditional Model			Conditional Model			
Grade Level	1-Level Variance	2-Level Variance	ICC	1-Level Variance	2-Level Variance	ICC	Reduction
Reading							
3	1957.120	169.147	0.080	522.316	19.205	0.035	0.554
4	1717.472	183.126	0.096	441.746	20.496	0.044	0.540
5	1667.844	142.533	0.079	442.476	0.953	0.002*	0.973
6	1686.353	178.567	0.096	412.197	15.270	0.036	0.627
		Reading Mean:	0.088			0.038	0.674
Mathematics							
3	2425.814	269.853	0.100	712.428	25.411	0.034	0.656
4	2353.830	208.008	0.081	644.858	65.909	0.093	-0.142
5	1934.310	202.642	0.095	493.857	19.883	0.039	0.592
6	1934.546	251.146	0.115	452.611	43.911	0.088	0.230
		Math Mean:	0.098			0.064	0.334
		Grand Mean:	0.093			0.053	0.505

* Value was double checked against original SPSS syntax, model specification, outcomes, and subsequent tabulation to see if data processing errors had occurred. The conditional 2-level variance for Grade 5 Reading is accurate as reported by the statistical model

ICC indices for the unconditional models range between (approximately) 8% to 10% in Reading and 8% to 12% in Mathematics. Similarly, ICCs for the conditional model range between 0% to 4% in Reading and 3% to 9% in Mathematics. Finally, ICC reduction estimates range between 54% to 97% in Reading and -14% to 66% in Mathematics. The values reveal substantive improvements in most of the models due to inclusion of the predictor variables. However, two errant results seem to stand out. First, the ICC (.002) for the Grade 5 conditional Reading model implies that the predictor variables virtually eliminate the 2-level nesting effects. Second, the ICC for the Grade 4 conditional model in mathematics increased, resulting in a negative reduction (-14%). This latter result suggests that inclusion of predictor variables did not mediate the influence of the school-level grouping effects.

Finally, although reporting substantive improvement for most models, the subject conditional ICCs report mean cross-grade values of 3.8% (Reading) and 6.4% (Mathematics). School level error variances were reduced, on average, 67% in Reading and 33% in Math. However, collectively the conditional multilevel models still retain 5.3% level-2 error variance. This suggests additional adjustments to the models might be investigated to further account for the between-school influences on the student achievement scores.

VAM reliability indices – standard errors of estimate (SEE). A third approach to examining the reliability of the VAM models focuses on level-1 regression standard error of the estimates (SEE) and the construction of 95% true-score confidence intervals (CI). In the context of VAM regression functions, the general form for constructing a 95% CIs around an estimated mean score is $\hat{Y} \pm 1.96\ \widehat{\sigma_E}$ where $\hat{Y}$ represents the estimated score

and $\widehat{\sigma_E}$ represents the regression standard error of estimate (SEE). The SEE is computed as $\widehat{\sigma_E} = \sqrt{\frac{\sum((Y_i - \hat{Y})^2}{N-P}}$ where Y_i is the actual score, $\hat{Y}$ is the regression equation estimated score, $\sum((Y_i - \hat{Y})^2$ is the sum of the squared deviation (residual) scores, N is the number of observations, and P is the number of estimated regression parameters (N-P equates to the degrees of freedom in the model). Necessarily, larger SEE values reflect greater measurement error (less precision, lower reliability) resulting in larger true score CIs.

CI information may also be evaluated in terms of test items. The base metric for all estimated VAM models are academic achievement measures from the state's standardized achievement tests (AIMS) expressed in the form of scale scores.[28] This metric may be re-expressed in terms of number of test items by regressing the scale score vector as a function of raw score (number of items answered correctly): that is, Scale Score = $\alpha + \beta$(Raw Score) + μ where α is the regression intercept, β is the slope coefficient, and μ is the error term. The estimated slope parameter $\hat{\beta}$ represents the incremental number of scale score points assigned for each additional correctly answered test question. Arguably, this provides for a more concrete understanding of model error as it relates to the accuracy of the student performance estimates.[29] Table 62 presents a summary of the VAM model SEE values, constructed CIs based on mean estimated scale scores, and associated test-item equivalents.

---

[28] Scale scores are based on summed raw scores transformed to a new scale for the purpose of standardization/equating across test forms, correction for item difficulty, and related psychometric adaptations (Tan & Michel, 2011, Embretson, 2000, Bond & Fox, 2007).

[29] It is noted that scale scores have no theoretical upper or lower bound since that are transformations of raw scores based on assumptions related to externally specified mean and variance values. For this reason, the range of CIs expressed in terms of scale scores cannot be evaluated to the low/high bounds of an underlying scale. To make such a comparison, the CI scale score range must be re-expressed in terms of equivalent test questions. The equivalent items may then be divided into the total number of test items to arrive at relative scale proportions.

Table 62

Conditional Model Level-1 95% Confidence Interval

Grade	Mean Predicted Scale Score (SS)	Level-1 SEE (SS)	Lower Bound (SS)	Upper Bound (SS)	CI Range (SS)	Items per Scale Score (#)	CI Range in Item Equivalents (#)	Total Items on Test	CI Item Range (%)
Reading									
3	468.165	18.195	432.502	503.827	71.325	4.503	15.840	54	29.333
4	487.202	15.548	456.728	517.675	60.946	4.135	14.740	54	25.926
5	505.004	15.442	474.738	535.269	60.531	4.222	14.336	54	25.925
6	517.061	14.368	488.900	545.222	56.322	4.350	12.947	54	22.222
Mathematics									
3	379.096	24.814	330.460	427.731	97.270	3.945	24.654	66	37.355
4	391.893	22.693	347.414	436.373	88.958	3.775	23.568	68	34.659
5	400.768	17.220	367.018	434.519	67.501	3.694	18.271	67	27.270
6	420.232	15.746	389.371	451.094	61.723	3.539	17.442	68	25.650

(SS) Scale Score; (SEE) Standard Error of the Estimate

For Reading, the lower/upper CI bounds range between 56 to 71 scale score points which is equivalent of between 13 to 16 items on the test. This translates to 22% to 29% of the possible test items. Similarly, for the Math VAM models, the CI lower/upper boundaries range between 62 to 97 scale score points, equivalent of between 17 to 25 test items and 26% to 37% of the possible test items.

Research Question 1E: Theoretical Construct Articulation (RQ1E)

Methodological Approach	Methodological Type	Participants	Data Sources & Collections	Effect Measures
Stakeholder Interview	Qualitative	Stakeholders (Teachers, Principals, Policy)	Interview	Coded Interview Responses;

Supporting Research Questions:

Item	Description	Approach	Measure
RQ1E (a):	What is the theoretical construct definition held by stakeholders regarding high quality teaching?	Semi-Structured Interview	Coded Interview Responses
RQ1E (b)	Do the theoretical construct definitions differ by stakeholder group? By VAM Group?	Semi-Structured Interview	Coded Interview Responses
RQ1E (c)	What are stakeholder perspectives regarding the purpose and intended outcome of teacher evaluation? Does this perspective differ across stakeholder groups?	Semi-Structured Interview	Coded Interview Responses

Figure 47. RQ1E theoretical construct articulation.

RQ1E (a) and RQ1E (b). What is the theoretical construct definition held by stakeholders regarding high quality teaching? Do the theoretical construct definitions differ by stakeholder group? The approach is semi-structured interviews. The measures are coded interview responses.

Introduction. This research question explores stakeholder perceptions of good/effective teachers. Two facets of this query are explored. First, how do stakeholders define what it means to be a good/effective teacher and do these perspectives differ by stakeholder group.

Because the evaluation process presumes to assess instructional competence, construct validation requires a comparative theoretical context. In this regard, the salient question becomes whether the sentiments held by stakeholders closely align with the metrics and evidences used in the evaluation process? Specifically, are the components

reified by the evaluation model aligned to the attributes exemplifying good/effective teaching held by teachers, evaluators, and policy makers? Close alignment serves to add credibility, while divergence raises questions, regarding inferential efficacy.

Construct overview. This topic was presented during the interviewee activity (generally) as "*What does it mean to be a "good teacher?"* The narrative analysis presented below argues that stakeholder views regarding what it means to be a good/effective teacher are complex, nuanced, but dominated by affective, non-academic, attributes. With some individual exceptions, this is consistent across all stakeholder groups. Specifically, the data presented below argue that the construct of good/effective teaching may be depicted as follows:

1. The construct is complex, encompassing diverse attributes reflective of quality instruction,
2. Affective interventions, impacts, outcomes are emphasized over academic dimensions,
3. Good/effective teachers purposefully engineer a supportive, personalized, learning environment, and
4. Test scores reflect only a minimum expectation for good/effective teaching. As such, achievement measures are an insufficient proxy for assessing instructional quality.

Throughout the narrative, stakeholders referred to attributes in terms of both activities (i.e., building relationships) and outcomes (i.e., highly motivated students). For the purpose of this analysis, both perspectives were conceptualized as *What Good/Effective Teachers Do*. That is, if a participant states that good/effective teachers

build relationships (an action) this is interpreted as both an action and an outcome. Similarly, if a participant distinguishes effective teaching in terms of highly motivated students (an outcome), it is presumed that teachers actively engineer this result through practice. In this way, the construct *What Good/Effective Teachers Do* incorporates both attributes and actions.

Analysis of the narratives suggests that the construct *What Good/Effective Teachers Do* has two primary sub-categories: *Providing a Safe/Nurturing Environment*, and *Mentor, Support, and Motivate*. The first sub-component, *Providing a Safe/Nurturing Environment*, is conceptualized from the data as follows:

> A good/effective teacher provides/constructs a learning environment that permits students to engage in, and struggle with, the act of learning new things. This environment permits/encourages students to take risks and fail without serious consequence. In doing so, it empowers students "not to know" and to grow through the act of learning. Engaging in the learning process in a safe, non-consequential environment, builds confidence and the desire to learn more. This concept is not group-centered. It focuses on the individual. It requires building trust, self-confidence, persistence, independence such that a student may take advantage of, and engage in, the opportunity to learn. In this way, creating a safe/nurturing learning environment is connected to the second sub-category *Mentor, Support, and Motivate.*

Codes and identities exemplifying this sub-category are shown in Figure 48.

- safe learning space
- safe to struggle
- take risk
- fail
- try/attempt
- fail
- without consequence
- persistence in a safe, encouraging, and supportive environment
- lack of judgment/consequence
- students expand, grow and independently explore
- confidence to struggle
- confidence to ask questions
- persevere/persistence
- trust
- relationship

Figure 48. Codes and identities related to *Providing a Safe/Nurturing Environment.*

The second posited sub-category, *Mentor, Support, and Motivate*, focuses on the individual student. The category is conceptualized from the data as:

> Teachers form relationships with, and understanding of, student's personal context. Based on this relationship, teachers are able to mentor students (as individuals) through a process of progressive learning and personal growth. Learning progression is personalized, built on each student's individual context. The focus is on making a difference in a student's life such that learning may occur over the longer run. Good/effective teachers help students overcome limiting factors by understanding needs in a personalized way, adapting instruction, building self-confidence, and engineering pathways to success. Importantly, "success" is much more than content learning, test scores, or academic outcomes – which are generally seen as a minimum and insufficient criteria for defining good/effective teaching. In addition, there is an implied ordered, causal, pathway connecting the nurturing of relationships, development of student self-awareness, and subsequent (long term) academic learning.

Codes and identities exemplifying this sub-category are included in Figure 49.

Affective/More than Academic:

- More than delivery of instruction; more than content; non-academic; help overcome outside factors/limits;

Teacher as Motivator:

- Engaging, build self-confidence, self-worth; affective; motivate/inspire; hope, sense of future; advisor, mentor, comforter; personalized; tap into student interests, strengths; build passion, desire to learn; positive outlook, growth personally & academically; 'belief they can succeed' leads to necessary for learning;

Personal/Relationships:

- Interpersonal; relationship with students and families; relationships #1; mentor; caring, safe, trust (student trust in teacher as friend/supporter); "know baggage students come to school with"; know strengths/weaknesses (academic & emotional/personal, social); individualized; recognition of the personal and how it impacts ability/desire to learn

Causal Pathway:

- Implied causal pathway: relationship/emotional -> increased engagement -> increased content learning -> desired outcome; relationship is correlated with learning; "Good/effective" addresses both emotional and mechanical components of learning; first do emotional side;

Test Scores as a Tool:

- Use frequent measures to show progress to student (build self-confidence & self-worth which leads to better engagement and performance over time); show students they are capable, growing, successful; continuous success in learning;

Figure 49. Codes and identities related to *Mentor, Support, and Motivate.*

Stakeholder narrative and analysis.

Teachers (What Good/Effective Teachers Do). Teachers expressed a myriad of attributes that defined good/effective teaching. As mentioned, the narratives are primarily defined within the sub-categories of *Providing a Safe/Nurturing Environment and Mentor, Support, and Motivate*. In this regard, a teacher participant (below) emphasized the fundamental need to build relationship with students:

> I think a good teacher is one that has a good rapport with the students. If they don't like you and you've done something that has upset them, they might not take your teaching to heart. You have to have that rapport with the students. Then you have to build upon that so that they have that trust with you - to want to learn. They need to be able to trust you that they can raise their hand and say, "Hey, Miss [name]" or whoever, "Will you help me with this?" and to know that if they're not getting it, they can always come to you and get that help. (Teacher 105)

For this person, good/effective teaching is defined in terms of relationship which, in turn, serves as a foundational pathway to learning. First, a good/effective teacher builds rapport with students. This forms the basis for establishing trust. Trust leads to a desire to "…want to learn…" and engagement in the learning process. In addition, trust permits students to feel like they "…they can always come to you and get that help…" In so doing, the teacher ultimately builds an environment where the student feels supported, safe, and able to have difficulty. Here, relationship facilitates the teacher's ability to support, mentor, and assist students: they feel confident taking risks, asking questions, and confronting difficulty. Embedded in the comment is a learning environment which permits students to explore, attempt, take risks, and question their learning process.

A second teacher states that good/effective teachers have to fundamentally "…know how to teach…" However, the context of this phrase is not simply skills-based pedagogy. Rather, it is relationship based, saying:

> A good teacher? You have to know how to teach [laugh]. It's not just coming in, looking at what is needed to be taught and throw the information up there and hope they understand it. You have to know the students. You have to know how they learn. You have to know things that are going on in their lives. It's not just, "Oh, here's 30 kids and let's teach them." You have to know about them [students]. You have to know about their age, you have to know about home, family, school, problems that have happened in the past, problems that may happen in the future … I tell them all about me all the time so that they feel that they can open up to me as well. Then we're all family, so they know my whole family … You're kind of like a mind reader all at the same time, and you're kind of the caregiver that they don't have while they're not at home. (Teacher 102)

For this person, "… know[ing] how to teach…" is not defined by content delivery. Here, good/effective teachers "... know the students… know how they learn... know about them…" This depth of understanding is at the personal/emotional level. It extends into the non-school contexts of student's lives: home, family, and past personal history. In this way, good/effective teachers do not simply deliver instruction. They are not providers. They are partly facilitators, caregivers, who adapt instruction practice by the affective, the personal, and the emotional. A third teacher echoes this connection between nurturing the affective and goal of learning, saying:

> The relationship enhances a test score. It can. To me that can be a correlation to the test score, the confidence … You can believe in yourself. You can do this. You've got the confidence. (Teacher 104)

For the three teachers above, good/effective teachers impact the individual by building a sense of self: self-confidence, self-value, self-belief, which, in turn, leads to greater learning. The foundation is relationship. Arguably, this pathway/connection to achievement indicts test scores as a proxy of effective teaching in the short run: test

scores may be high for selected students already self-empowered, self-motivated, or self-connected to academic (tested) learning, and require little from a teacher except delivery of content. In this way, minimally effective teachers may still display high achievement. However, good/effective teachers are characterized by their ability to alter outcomes that would not have occurred without their intervention because they impact the affective through relationship, support, mentoring, and caregiving. Two additional teachers (below) both capture aspects of this this perspective stating:

> I think it's just really looking at each kiddo, and seeing what their needs are emotionally, mentally, socially, and academically, and trying to give them the best that I can, so those first important needs are met, so that then they can move on, and focus on the academic skills that they need. (Teacher 106)

> That engagement, that level of trust, with the kids, I guess, where they know that they are not going to be put into a specific situation where they're embarrassed, or anything like that. Then you need to have a good rapport with the kids and the families and your school. All of that kind of encompasses everything, so yeah, it's so hard to describe. (Teacher 103)

Teacher 106 outlines a causal pathway grounded in the collective understanding of a student's emotional, personal, social, and academic needs. The main goal is "…so those first important needs are met…" The perspective is time-relevant, "…so that then they can move on…" as well as causal and progressive. In addition, Teacher 103 brings together aspects of relationship, trust, the affective, and the learning environment. Saying "… [students are] not going to be put into a specific situation where they're embarrassed…" is a statement of trust fostered by the classroom experience. This environment is constructed and purposeful. It does not occur by chance. It is a direct artifact of good/effective teaching.

Only one of the seven teachers interviewed cited test scores as a primary indicator of good/effective teaching. However, the context of this reflection is one of "testing as a tool," a means to provide student's feedback on their personal capability, growth, and success. In this way test scores are motivational and facilitate confidence building. Importantly, the discussion is not about application of state-level standardized tests as an indicator of quality instruction. Rather, it is about the use of pre/post testing in the context of day-to-day instruction. The teacher explains:

Interviewer:	What if somebody from outside the school that you met casually, or friends, asked "what is the definition of a good, great, or effective teacher. What would that be?
Teacher 101:	I have to figure out what's best suited for each individual [student]. I kind of look at it, for me, for the evaluation process is "are my kids learning?"
Interviewer:	So a good teacher is a teacher where the kids learn what we teach them?
Teacher 101:	I would agree with that.

The teacher goes on to say:

> I do a lot of pretesting and I do a lot of just "where they started, how they are doing, how do they know this multiplication or this division?" Then as I teach it, and as time goes on, are they showing improvement? Their confidence grows. They're just happier, they're more motivated, and their goals are set. Just from me working with them and other school staff working with them, they just feel more confident with themselves like they've actually achieved what they were supposed to achieve.
>
> … If you can inspire them - there's hundreds of movies on teachers inspiring kids, but if you really are a teacher, a lot of it is true. If you can get kids to come to school, that could be half the battle. Research will say if you can get kids to come to school a certain amount of time, then their success rate, their graduate rate will go up. (Teacher 101)

This teacher initially responds to the question of what makes a good/effective teacher in terms of content learning. But the context becomes qualified by the personal context of a student. Just as with his/her colleagues, the deeper narrative is one of

relationship impacting the affective: "…If you can inspire them…" Motivation, confidence, desire, a realization of personal growth and capability, are all concepts embedded in this view. This teacher views pre-/post-testing as a motivational tool to effect change in students' personal relationship with learning. Impacting the affective becomes a fundamental attribute of what good/effective teachers do ("…if you really are a teacher……")

In addition, this teacher emphasizes the importance of non-academic, non-tested, goals/objectives by noting "…If you can get kids to come to school, that could be half the battle…" There is acknowledgement that good/effective teachers recognize and nurture these conditions in the short run so that academic outcomes may improve at some point in the future. Again, there is a concept of time and progression, beginning with the affective, that eventually leads to improved learning outcomes.

Importantly, it is not being argued that this particular teacher (101), unlike his/her colleagues, sees learning outcomes as an important outcome of good/effective teaching. The narrative suggests he/she does. However, to argue that it represents the essence of what a teacher does would be inaccurate. For this individual the act of being a good/effective teacher is complex, incorporating affective actions and impacts that are foundational.

Teacher summary (*What Good/Effective Teachers Do*). The teacher narrative under the component referred to as *What Good Teachers Do* emphasizes affective impacts on students. It may be categorized in terms of *Providing a Safe/Nurturing Environment* and being a *Mentor, Supporter, and Motivator*. In this context, good/effective teachers purposefully strengthen the connection students have with the

learning experience. This is not an attribute of the student, but explicitly an artifact resultant from teacher action. It is accomplished by developing relationships, understanding the personal, and providing an environment that rewards attempt, effort, risk, and growth. In this way, a good/effective teacher empowers students beyond levels otherwise present prior to entering their instructional setting.

Principals (What Good/Effective Teachers Do). The perspectives of principals regarding what good/effective teachers do under the sub-categories *Provide a Safe/Nurturing Environment* and *Mentor, Support, and Motivate* aligned closely with teachers. Readers are reminded that principals tended to place higher value on test scores as a primary indicator of quality instruction. However, their narrative with regard to *What Good Teachers Do* reveals a nuanced perspective that includes many non-tested dimensions of professional practice. One principal places an emphasis on relationship building and affective impacts as an important trait of good/effective teachers:

> You want someone who's going to connect with students, and build those relationships to reach those hard kids that have this outer shell that cannot crack, and they haven't been motivated. There are probably four or five different teachers on my campus that I would call very good teachers. I've got one who can reach any kid. It doesn't matter how tough they are. Are they the most engaging teacher? No. Do they have the best test results? No. But if I have a tough kid that I need someone to connect with, I know she is the best teacher for that child. (Principal 204)

Here, the competence of these teachers is depicted not in terms of test scores ("…Do they have the best test results…"), but in terms of relationship for "…those hard kids…" It is the ability to identify students who need the personal, the emotional, in order to affect change toward learning. Here, good/effective teachers impact the affective in order to change the outcome. There is recognition that not all students are the same or

require the same relationship. In this way, good/effective teachers understand the individual context of students' lives, concluding "…I would call [them] very good teachers…" The same principal goes on to share:

> A good teacher is someone who makes learning relevant, gives them a reason to learn it, to remember it, and to use it … [The ability] to understand your kids and what they need. Whether it's a whole group, small group, individually, you're able to stretch and support their learning. (Principal 204)

For this principal, instilling personal motivation and a desire to learn through relationship is augmented by tailoring instruction in terms of relevance. The instructional delivery component is personalized, not group-centered. Good/effective teachers merge the personal ("…understand your kids…") with the instructional in order to "…support their learning…" Here, relationship, personalization, and understanding, lead to stronger connection, motivation, and desire for learning. These facets permit good/effective teachers to tailor instruction, make learning interesting and relevant to the student which arguably, leads to improve outcomes.

A third principal describes attribute of good/effective teachers as:

> When I think of my great teachers here at my school, I think of teachers who have built those relationships with their kids, number one. They know their students' strengths and weaknesses academically, emotionally and socially. [Indeed] their job is to build those relationships and get to know our kids because unfortunately, home lives are so different than they used to be. They come with such baggage now and so we have to be that stable role model for them, with good character, responsibility, and just all the things we want our kids to have. (Principal 205)

Again, the narrative describes "great teachers" in terms of relationships where relationship is "number one" This individual defines relationship as "… know[ing] their students… academically, emotionally and socially…" Here, content learning (academics) is positioned as only one of multiple components. The person sees relationship as

fundamental criteria of "their job." They have a responsibility to recognize outside, non-school, factors and understand their importance to the learning process. Good/effective teachers identify, act upon, and mitigate these factors through professional practice.

Interestingly, this person also introduces the idea of teachers as role models. In this regard, teachers serve a larger purpose than delivery of content knowledge. Through their professional practice they shape "…good character, responsibility, and just all the things we want our kids to have…" This is a profound concept since it elevates the purpose, role, behaviors, and impacts of teachers well beyond content. Good/effective teachers shape the individual within a social context: character, integrity, responsibility are all facets of the interpersonal, the social. In this way, the principal connects both the role of teachers and the purpose of education to a much larger community framework.

A fourth principal defines good/effective teaching in terms of commitment, dedication, and passion for ensuring students succeeds, saying:

> A good teacher is someone who has the passion to do whatever it takes for their kids. They will look at each child and not just say, oh, this whole group is this … They are the ones who celebrate the kids' natural excitement about discovering and learning. They care about kids. (Principal 201)

Here, the emphasis is on the individual, "each child." Good/effective teachers celebrate each student, care about the individual, and leverage students' "natural excitement" that connects them to learning. Good/effective teachers possess a "whatever it takes" attitude.

Similarly, a fifth principal describes good/effective teachers in terms of their personal characteristics and commitment to teaching:

> [In addition to the state curriculum] I'm going to be a great person, and I'm going to help kids. I'm going to be positive for them, and I'm going to work with them.

> I'm going to work with their parents. I'm going to show these kids that I love them in tangible ways like going out of my way for them, talking to them about, those types of behaviors are what make kids into "I can do this because someone believes in me. Someone's behind me. Someone cares about me". [Student's] know that there are people that care about them and are in their corner. (Principal 207)

This principal sees the teacher's character as an important attribute. Here, the teacher's personality and commitment directly impact students' success through relationship, caring, and support. For this principal, emotional sincerity seems central to the concept of being a good/effective teacher. That is, student must believe that the teacher values them, is concerned about them, and cares for their welfare. In this way, a good/effective teacher instills the belief that "… Someone's behind me. Someone cares about me …" In this way, the student's self-perception is foundational to learning and good/effective teachers are the chief architects of this perception ("…those types of [teacher] behaviors…") There is active, involved, intervention where the teacher must "…show these kids… in tangible ways … [go] out of my way for them…"

A different principal (below) discusses attributes of good/effective teachers as a combination of both the affective and academic. However, the academic component is not expressed as absolute, but rather as growth, improvement, and progress:

> If they [students] had a good teacher, a kid would feel very positive about their learning. It wouldn't matter if they were low or if they were high [achievement]. They would know that they are doing better, and they would have that positive feeling about themselves and the direction they're going.
>
> I think you are moving every student forward, meaning that wherever a student came to you at, whether it was super high or super low, that that student has gained more knowledge and has a better understanding and more of a positive view on themselves and their learning. That would be a good teacher. (Principal 206)

Here, in achievement terms, direction is more important that amount. Student self-perception is a co-constructed narrative supported by evidence of success and improvement—students "gain knowledge" and have "…positive view[s] on themselves…" Importantly, as discussed earlier by a teacher participant, achievement is partially seen as a tool for impacting the affective. In this context, academic measures become useful when used by a teacher to support, enhance, and improve student self-perception and strengthen the connection to learning. Good teachers may be identified by these metrics: "…That would be a good teacher…"

All of the above discourse exits within the context of the classroom and the actions of the teacher as an instructor, role model, mentor, counselor, and caregiver. In this context, a seventh principal summarizes what a good/effective teacher accomplishes by stating "… [a good/effective teacher] creates an environment where a student can learn, that culture and environment of learning, to be able to create a learning environment…" (Principal 202). The statement succinctly encapsulates the affective attributes with a causal pathway to learning. The environment brings together the student personal/emotional context with motivation, engagement, and desire to learn. In addition, environment is constructed and therefore depends on the capability, actions, and attributes of the teacher, inclusive of his/her affective characteristics.

Principals summary (*What Good Teachers Do*). The narrative provided by principals regarding *What Good/Effective Teachers Do* is aligned to that of teachers, mirroring aspects of *Providing a Safe/Nurturing Environment* and *Mentor, Support, and Motivate*. It emphasizes the affective/personal for both the student and teacher. Good/effective teachers directly impact student's self-perception and the connection they

have to the learning process through action, intent, and process. They develop intimate understandings of the student context and adapt/adjust instruction as a method of connection. Teachers' commitment to ensuring students grow, improve, and ultimately succeed is a core attribute. As in the teacher narrative, relationship is foundational.

District (What Good/Effective Teachers Do). District participants place emphasis on non-achievement attributes of good/effective teaching. As with teachers and principals, district members speak of the learning environment, relationships, motivation, support, and trust—all factors related to how teachers impact a student's self-perception and his/her connection to the learning process. Implicitly, it assumes that achievement is a function of these affective attributes. One district member shares his/her perspective on attributes of good/effective teachers, saying:

> A good teacher is a teacher that can provide an environment where students are/feel safe to be able to do what they need to do. That's it. There isn't [any] negative ramifications - where they feel supported, where they feel safe, where they feel like, "I can take a risk. I can learn something, and if I do it wrong, it's okay. I can learn from that just as I can learn from if I had did it right." … Often we learn more from our mistakes than we do when we do it right. A good teacher creates that kind of an environment. (District 301)

Here, good/effective teacher creates a learning environment through his/her action. It is an outcome of quality professional practice. The environment is purposeful, a catalyst for learning, and directed at the affective. Attributes of environment include safety, low consequence, and support. Students become risk-takers, able to struggle with the process of learning and it is personal ("…do what they need to do…"). In this way, environment is meant to empower students in learning. The same person goes on to add:

> A teacher should be a facilitator. They're not just a deliverer of content. They shouldn't just be standing up in front of students delivering content … They need

> to be able to build relationships and trust with students … Mutual respect between students and teacher have to be in place.
>
> It's not just about academic things. There are problems like character issues, cultural things. I think public education [teachers] play a role in those pieces as well … Teaching is such a much broader endeavor than teaching a math concept to students in a classroom. It encompasses so many things. One good teacher builds relationships in a different way than another good teacher. It could look different in different classrooms. (District 301)

The view describes good/effective teachers as facilitators. They create conditions that permit personal exploration and empowerment. They are not providers or deliverers. They do not grant knowledge. Rather they invite students to engage in learning and then mentor, assist, and support their efforts. The process is founded on relationship. Relationship builds trust and mutual respect. Relationships are key because "…It's not just about academic things…" Finally, different teachers develop relationship in different ways, suggesting the professional practices of good/effective teachers are as uniquely diverse as the students they teach.

Another district member reflects on the meaning of good/effective teachers by offering:

> Certainly teachers have to have knowledge of their curriculum to be able to be effective, but they also have to have knowledge of psychology and how to deal with a variety of students. [Teachers] have to know how to motivate kids, how to make things relevant and how to really compete with everything else that's going on … I think that good teachers create and instill a love for learning. I think good teachers make kids want to explore more on their own. I think good teachers create motivation. They create desire. They create passion for the subjects that they teach … A good teacher might not have as great of scores as another teacher, but yet still be a good teacher. (District 303)

The individual initiates the discussion by stating teachers must "…have knowledge of their curriculum…" This is stated as a basic assumption, a minimum required expectation, a certainty. From there, the requirements move into non-academic realms:

"…knowledge of psychology … how to motivate kids … make things relevant…" Good/effective teachers "…instill a love for learning … create motivation … create desire … create passion…" The perspective is impact on the affective and the personal. Test scores become devalued as a correlated indicator of good/effective teaching.

A third district member also positions teachers as facilitators of the learning process and then describes the context of this facilitation:

> I think our teachers now have to be facilitators. They have to be mentors. They have to help the student design their own learning. Personalizing education. We need to tap into more than just what the student knows or doesn't know. We really need to tap into their interests, their strengths, those pieces that will motivate them, inspire them, so that they can take on those opportunities and those challenges. (District 304)

For this person, good/effective teachers mentor, help, motivate, and inspire. They personalize by understanding the individual's "…interests, their strengths…" Good/effective teachers use this understanding to empower students to "…take on those opportunities and those challenges…" Once again, the teacher is not being described as a purveyor of knowledge, but as a resource for acquiring knowledge. The same individual goes on to share:

> I would like to have a way to really track our students as they go on in life. If students are able to accomplish whatever those career goals are, if they are able to get into those programs that they need to get into to be able to become productive and successful, then we're doing a very good job. (District 304)

This last comment implies that measures of good/effective teaching are personal, long-term, and non-academic. The necessary metrics are currently unavailable ("…I would like to have…") and unique to the individual ("…career goals … those programs ..."). Indeed, in this context, the primary indicator of good/effective teaching is whether

students are "… productive and successful…" in their lives. If so, "…then we're doing a very good job…"

Only one of the four district participants made a direct connection between good/effective teaching and academic indicators stating:

> You [could] define a good teacher on the traditional student response of, "Who's your favorite teacher, and why is that teacher your favorite teacher?" Or, you could judge a teacher, as a good teacher, based on the percent of students who score in the "exceeds" category of an assessment. Probably, the best teacher is the favorite teacher and has the highest scores. (District 302)

However, even here, relationship, the affective, becomes a primary indicator of instructional quality. Arguably, the person may be suggesting a causal connection between student engagement and achievement: that the benefits of building strong student-teacher relationships lead to greater achievement due to increased motivation, participation, and support. Regardless, the importance placed on non-academic attributes is evident in the comment.

District summary (*What Good Teachers Do*). Like teachers, and principals, district members emphasized affective dimensions of within the student-teacher relationship. Good/effective teachers purposefully construct learning environments that are safe and supportive, they build strong personal relationships with student in order to develop trust, motivation, and desire to learn. Good/effective teachers inspire and empower students. They facilitate learning as opposed to deliver knowledge. Academic outcomes are seen, at best, as eventual outcomes of relationship.

State (What Good/Effective Teachers Do). State-level stakeholders expressed similar sentiments regarding good/effective teaching: there is an emphasis on student affective impacts built upon relationship and attention to non-academic facets of

education. Again, dimensions of the affective are seen as foundational and necessary to affect higher academic learning. There is also recognition of time: short-run affective impacts are required to improve learning over the longer run. In this regard, one state member summarizes characteristics of good/effective teachers saying:

> Absolutely key is their [teacher's] ability to form relationships with each student, to engage parents in assisting, and to support the entire education culture of the school … They [great teachers] are passionate about education itself. They care about their students. They are capable of emotionally connecting with their students so that they can unlock them in terms of changing their attitudes and improving their attitude towards learning. That's the number one thing that a teacher has to do is, they have to engage the student before the journey starts. (State 401)

The person is clear with his/her perspective – good/effective teachers "…form relationships with each student…" The narrative uses words/phrase such as engage, assisting, support, passionate, care, and emotionally connecting. There is a common link between what teachers instill in students and the teacher's commitment to the process ("…They are passionate about education itself…" and "…They care about their students…").

In the perspective (above), good/effective teachers impact students by "…changing their attitudes…" Importantly, teachers affect this change by emotionally connecting. The goal is student empowerment: increasing motivation and desire transfers the act of learning from teacher to student. Changing student's attitude toward learning is "…the number one thing…" and it precedes the act of learning: "…before the journey starts…" The same individual goes on to say:

> Some teachers are confronted with such profound challenges that they almost can't teach content. They have to first unlock the student's alienation from the very process of education. In unlocking that alienation, they can permanently change their trajectory [even though] they might not do a very good job of teaching the content that year. (State 401)

Here, good/effective teachers mitigate "profound challenges" that students bring to the classroom. These factors impair learning. The teacher unlocks the student from these conditions. In doing so, they permanently change the trajectory of learning. Here again, the dimension of time is implicit: in the short run, good/effective teachers focus on creating the personal and environmental conditions necessary for learning. Once completed, students are empowered to learn. In this way, near-term, static, measures of achievement become poor proxies for instructional quality because they fail to assess the affective, both in terms of students personal/emotional and the teacher's own professional practice. This focus on the affective constitutes an indirect curriculum where the primary focus (goals, objectives, and desired outcomes) is non-academic. The state-level stakeholder then summarizes:

> It all comes down to motivation. It's profoundly important that they [students] be motivated on that subject matter. A great teacher will not only teach the student the subject matter, but they will make them a permanently better student for all teachers that come after them - because they will move that student to a higher motivation level as it relates to learning. So great teachers move/change the motivation level of students. (State 401)

For this individual, good/effective teachers "…not only teach …" but "…make them a permanently better student…" In this way, the teacher is transformational where impact is realized by the student's subsequent teachers ("…for all teachers that come after them…"). There is a causal pathway in the collective dialog of this individual that begins with relationships, allowing remediation of external limitations and eventual transformation in terms of connection/motivation to learning. This then leads (eventually) to higher achievement in the future. Again, the affective is an immediate and necessary

condition for learning. Arguably, for this person transformation equates to empowerment, where students become personally invested in their education.

A second state participant echoes the importance of this type of transformational timeline:

> A great teacher that will, over a period of time, take a student and demonstrate that they have not just gotten those particular skills, but they've really grown in their capacity to learn and want to learn … [It is] Making a student more aware and more engaged in having a desire to continue to want to learn. (State 403)

Again, the focus is on empowerment through transformation in terms of motivation, engagement, and connection to learning. The outcome, the result, is whether or not students "…want to learn … having a desire…" The assumption is that these conditions are required and foundational. The measure of success is not limited to the academic ("…not just gotten those particular skills…"). Good/effective teachers do more than academic. In this way, defining good/effective teaching is complex, multilayered, and difficult. To this end, the same individual qualifies his/her discussion (below). (Note: The person is reacting to the difficulty in clearly defining/delineating what it means to be a good/effective teacher).

Interviewer:	Okay, all right. How would you put a definition to what a good teacher is?
State 403:	I think that's a little bit problematic because I think a lot of times the state or school districts or even for that matter Congress has tried to do that, right, and tried to say, "Okay, this is a quality teacher." I'm not so sure you could say that simply, just like you can't say what is a quality parent or what is a quality governor. There's different types of teachers that I think are difficult to just pigeonhole.

The stakeholder raises concerns of simplification, reductionism, and generalization. Attempting to articulate a finite set of attributes across all contexts,

situations, and conditions is problematic because "…there's different types of teachers…" The word pigeonhole is pejorative, negative, indicting attempts to construct an archetype of instructional quality.

Complexity is a concept inherent in the narrative shared by a third state member who describes good/effective teaching in terms of the many roles teachers occupy:

> Interviewer: How would you define what it means to be a good or effective teacher?
>
> State 402: A teacher has many roles, so maybe that's what I'm trying to say. The roles take on different emphasis based on the age of the student. I think the teacher, first of all, is a purveyor of learning. Someone who has content to share and presents and develops that content for the learner. I think the teacher is an advisor or a mentor, someone that the person can go to and seek comfort or advice from. I think a teacher models being a member of a community and of society, and that the teacher's behaviors contribute to a broader view of responsibility on the part of individuals.

For this person, teachers are fundamentally "…purveyor[s] of learning…" Teachers have knowledge that they attempt to transfer to students. However, they are also advisors, mentors, and counselors ("… seek comfort or advice from…"). In addition, good effective teachers are role models, exemplifying the character and behaviors necessary for entering into community and society. As role models, good/effective teachers acknowledge this where their "… behaviors contribute to a broader view…" This infuses the non-academic and imposes a larger responsibility of building social character. The context is complex where relative importance of each attribute differs with context ("... different emphasis based on the age of the student…"). This is consistent with other stakeholder views which position affective needs ahead to content learning for selected groups of students. The same individual adds to the discussion, saying:

> I don't think most [teachers] enter the teaching profession with the sense that, "I have content to impart." I think we come into this profession for many other reasons, and some of those reasons are related to the roles that I've described. (State 402)

This comment qualifies the participant's previous statement positioning teachers as "…purveyor[s] of learning…" suggesting it is a too simplistic representation. Here, the motivation to enter the teaching profession is not driven by a simply desire to impart knowledge. Rather, it is more complex and founded in a desire to impact the personal and develop the individual.

State summary (*What Good Teachers Do*). State members emphasized many of the same attributes as other stakeholder groups. The narrative emphasizes the affective, the personal, and the emotional. The role of teachers is complex and extends beyond content knowledge. Teachers are positioned as role models in a broader social/community context. They ensure future learning by forming relationships, building confidence, and developing connection, engagement, and motivation. For state members, good/effective teachers are transformational, permanently changing a student's connection to learning. Attempts to simplify and/or reduce effective teaching to a finite set of attributes is problematic the context of teaching is dynamic and complex.

Summary of findings RQ1E (a). The research question addressed under RQ1E (a) was formulated as *What is the theoretical construct definition held by stakeholders defining good/effective teaching?* The interviewee prompt took the general form *What does it mean to be a "good teacher"?* Analysis of the stakeholder narrative suggests the following perspective:

> Stakeholder narratives (n=22) do not define the construct of good/effective teaching solely in terms of academic outcomes or skills-based pedagogy. Indeed,

these attributes may be viewed as minimal expectations that do not adequately distinguished between poor, average, or exceptional teachers. Here, facets of being a good/effective teacher are discussed more in terms of the affective, defined more by what teachers do for students rather than to students. In this regard, good/effective teachers are seen as facilitators who empower, inspire, and motivate. They are not simply delivers of knowledge. For these participants, the affective dimensions of the student-teacher relationship dominate the construct where the affective relates to the personal, emotional, and self-perceptual.

The narrative posits that good/effective teachers form foundational relationships in order to develop an understanding of the student's personal context. In so doing, they are positioned to individualize and personalize interventions (both academic and non-academic). Through purposeful action, good/effective teachers are transformational, permanently strengthening student's personal connection and responsibility to the learning process. Transformation occurs along a causal pathway beginning with relationship, which permits trust, which facilitates motivation, engagement, and a desire for learning. Academic outcomes are a partial, long-term, by-product of this transformation. Good/effective teachers transform students from passive to active learners. The effects and benefits of this are persistent and sustaining, realized by subsequent teachers.

Good/effective teachers purposefully engineer both the relationship and the learning environment. They are architects of the personal context of the learning environment. To develop environment, good/effective teachers become mentors, counselors, caregivers, and champions. They nurture the individual toward self-empowerment through relationship, trust, and support. These teachers fashion a learning environment that is safe by rewarding effort, difficulty, attempt, failure, and struggle. The environment is risk-adverse built on trust.

Good/effective teachers are passionate, committed, and emotionally invested in the profession. They internalize and take responsibility for the personal welfare of their students. In this context, welfare is not defined strictly in terms of the academic. They also serve as role models in a larger social and/or community context. Here, good/effective teachers develop the citizen, the social person, the member a larger community outside of the educational experience. They do this by modeling behaviors that, as an adult, permit access to the community at large.

Good/effective teachers see a distinction between direct and indirect curriculums. While academic instruction is an example of direct curriculum, indirect curricula refer to the goals, objectives, and outcomes aligned to the affective and involve transformation based on connection, self-perception, and motivation. Importantly, indirect curriculums are tangible outcomes that are reified, purposeful, and acted upon.

Overall, stakeholder narratives expose a construct that is complex, multi-dimensional, and nuanced. What distinguishes good/effective teachers is not dominated by academics or pedagogical skill. Here, delivery of instruction is a necessary, but substantively insufficient, attribute of instructional competency. What distinguishes good/effective teachers is affective transformation in terms of the personal/emotional.

Petite assertions for RQ1E (a) and (b). *What is the theoretical construct definition held by stakeholders regarding high quality teaching? Do the theoretical construct definitions differ by stakeholder group?*

Exemplars of Good/Effective Teachers:

1. Academic outcomes convey a limited and insufficient context to distinguishing good/effect teachers.
2. Exemplars of good/effective teachers are dominated by the transformational, affective, impacts they have on student's personal/emotional connection to the learning process.
3. Exemplars of good/effective teachers include the emotional investment in the (non-academic) welfare and development of students.
4. Good/effective teachers construct "safe" learning environments in which students are free to take risk, struggle, and succeed in the learning process
5. Good/effective teachers attend to both direct (written and tested) and indirect (affective) curricula (attendance, engagement, self-perceptual, desire, enjoyment, etc.)

Differences in Stakeholder Perspectives:

- No substantive differences were revealed between stakeholder groups regarding the construct *What Good/Effective Teachers Do.* Each group emphasized the importance of affective attributes and qualities as the primary, distinguishing, characteristic of good/effective teaching.
- Interestingly, in the narratives aligned to other facets of the evaluation process, principals expressed markedly different views from their colleagues in the three remaining stakeholder groups. Here, principals expressed satisfaction with the evaluation structure/process and emphasized test scores as a primary indicator of instructional efficacy. However, when discussing a more idealized representation of what good/effective teacher do, their perspectives aligned with those of the other groups – emphasizing affective attributes.

RQ1E (c). What are stakeholder perspectives regarding the purpose and intended outcome of teacher evaluation? Do these perspectives differ across stakeholder groups? The approach is semi-structured interview. The measure is coded interview responses.

Introduction. This research question examines stakeholder perspectives regarding the overall purpose of evaluation. It informs on an important aspect of construct validation: the alignment between stated (hypothesized) and perceived (actualized) intent (AERA et al., 1999, 2014; Messick, 1989a; Cronbach & Meehl, 1955). By design, *Purpose of Teacher Evaluation* positions the act of evaluation within the larger global constructs of *Purpose of Education* and the *Role of Teachers*. It presumes a causal

connection between gathering evidence of instructional efficacy to ensure larger societal-level education outcomes.

As discussed in Chapter 1, the policy premise for conducting teacher evaluation is improved instructional quality and higher student achievement (Ariz. Rev. Stat. §15-203A.38, 2010). In this context, the task of capacity building and improvement is presumed by state policy to be facilitated through a process of evaluation and accountability (ADE, 2012a; Task Force on Teacher and Principal Evaluations, 2011). That is, if teachers are evaluated (and then held accountable to the results), their practice will improve along with student learning. Thus, the question is whether or not the theorized purpose of evaluation is actually realized in practiced: Does instructional practice improve under the context of evaluation? It is argued herein that examination of stakeholder perspectives partially informs on this connection between theory and practice: Do stakeholders believe that implementing the state-formulated evaluation structure will result in the outcomes posited by the policy directive? Disconfirming evidence would raise questions as to evaluation efficacy and related inference-based consequences—intended outcomes are not being realized in practice.

The generalized interview prompt for this research question was presented to stakeholders as "*What is the purpose of teacher evaluation?*" The prompt was positioned as part of a broader discussion regarding the idealized purpose of education and associated attributes of good/effective teaching. In this way, the initial set of discussion prompts were purposefully ordered as:

1. What do you think the purpose of public education is?
2. What role do teachers play in that purpose?

3. What is the definition of a good/effective teacher?

4. What's the purpose of teacher evaluation?

The intent of the ordering was to explicitly examine stakeholder perspective outside of the context of the realities imposed by the locally implemented evaluation process. For validation purposes, it is this comparative representation between idealized and realized contexts that inform on construct validation. Interestingly, most interview discussions naturally entwined perspectives of both idealized and realized implementations.

Construct overview. Narrative analysis revealed two primary themes associated with the global construct *Purpose of Teacher Evaluation*: *Policy Intent* and *Operational Intent*. *Policy Intent* characterizes the question of purpose in terms of *Why Evaluate* while *Operational Intent* reflects a context of "*What To Evaluate."*

Each theme is viewed against two main perspectives describing the purpose of evaluation: *Evaluation-As-Accountability* versus *Evaluation-To-Improve-Practice*. That is, a *Policy Intent* may be to hold teachers consequentially accountability to some a priori set of criteria (i.e., *Evaluation-As-Accountability)*, or, to gather non-consequential information upon which to improve instruction (i.e., *Evaluation-To-Improve-Practice)*. In the same way, *Operation Intent* may be purposed to faithfully reify all attributes of good/effective teachers in order to inform on, and ultimately improve, professional practice (i.e., *Evaluation-To-Improve-Practice),* or, to directly assess student mastery of key subject areas based on specified learning targets (i.e., *Evaluation-As-Accountability)*.

Finally, for each permutation, stakeholder narratives may be explored in terms of an idealized versus actualized point of view. An idealized perspective expresses what the

person believes should be while an actualized perspective reflects his/her perception of what actually is. The two may or may not be in agreement within either Policy *Intent* and/or *Operational Intent*. A graphic depicting the construct components is provided in Figure 50.

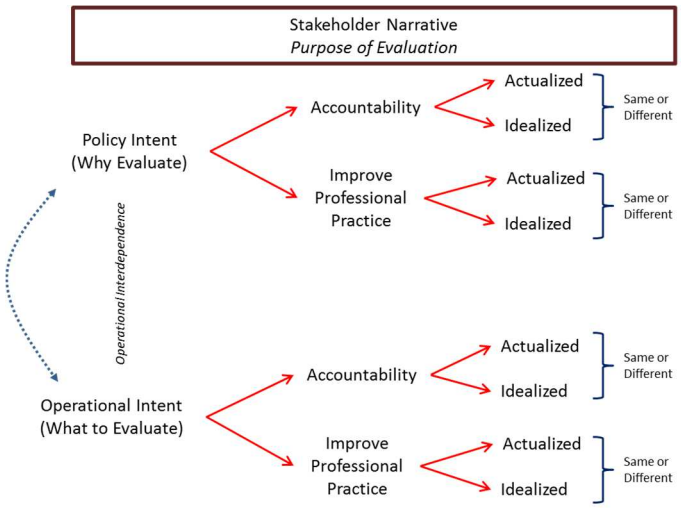


Figure 50. Components for global construct - purpose of education.

Throughout the narratives, conflicts arise between idealized and actualized perspectives of purpose. When conflict arises, actualized systems, structures, and/or processes are more often seen as having been imposed and/or mandated by higher policy levels that are external to an individual's locus of control. Here, actualized evaluation does not align with stakeholder's personal, idealized, belief of how evaluation should be structured, conducted, and/or utilized. Under *Policy Intent*, this conflict is differentiated

by competing purposes ascribed to different policy groups while under *Operational Intent* the conflict is between competing structural designs (i.e, measured components).

Based on narrative analysis, the following construct annotations are posited:

Policy and Operational Intent:

Policy Intent defines the stated purpose for conducting the evaluation activity. It forms the basis from which to operationally construct evaluation systems and assessment of related policy goals. Policy Intent includes an expectation of inferential judgments derived from the evaluation data. It is a statement of authority that supposedly guides the operational.

Operational Intent defines the parameters upon which to construct and implement evaluation systems. The comparison is between the factors reified by the system and those implied by the theoretical policy framework. In this way, Operational Intent reifies the component measures supporting the Policy Intent. Thus, Operational Intent refers to structure, component parts, and processes used to carry out the evaluation activity. Here, the actual conduct of the evaluation may or may not align to the Operational Intent.

Evaluation-to-Improve-Practice:

Evaluation-to-Improve-Practice positions evaluation as a process for purposefully improving the professional practice of classroom teachers. It does so by informing on current behavior. Evaluation data is used to identify targeted areas of intervention and training. In this way, evaluation is meant to inform, identify, and facilitate. The actions associated to these data serve to develop, enhance, and/or improve instructional competence which then presumably translates into improved student learning. In this context, the major benefit derived from evaluation is self-awareness, self-development, and evolving instructional competence. Evaluation-to-Improve-Practice does not connote accountability in terms of consequential action or punitive sanction. The purpose is developmental, positive, and empowering.

Evaluation-to-Improve-Practice is characterized by the following codes and identities:

- Improving professional practice; supportive; not punitive; positive cultural focus; get better at craft; guide teachers toward improved effectiveness; gauge where/how to improve; provide opportunities to improve; growth; target professional development; facilitate efforts to improve practice;
- Inform, provide feedback, provide information; acknowledge competencies, express appreciation, affirmation; constant feedback system; identify strengths/weakness;
- Promote (self) reflection, dialog, communication, collaboration, motivate (to improve); personal, individualized, build relationships;

Figure 51. Codes and identities Evaluation-to-Improve-Practice

Evaluation-as-Accountability:

The *Evaluation-as-Accountability* component is an outcome/efficacy concept. The context focuses on assigning responsibility of outcomes to an individual and assessing conformity of professional practice to establish norms/standards/expectations. The intent is to identify, distinguish, and to rate/rank. It is an assignment of performance and/or quality used to differentiate along a quality scale. Arguably, the connotation of *Evaluation-as-Accountability* is more concerned with reward/punish than to inform/improve. *Evaluation-as-Accountability* is a benchmark, a standard, and an expectation of performance and it implies an expectation of conformity;

Evaluation-as-Accountability is characterized by the following codes and identities (Figure 52).

Evaluate Teacher PP:

- Standards compliance, conformity, adherence; curriculum/instructional fidelity; content delivery; assess/monitor instructional practices; quality instruction, high level instruction; effectiveness; pedagogy; goals, objectives, outcomes; expectation;

Assess Effectiveness:

- Assess, measure, quantify; identify, distinguish, rank, rate; levels of effectiveness; good/poor instruction

Ensure Student Learning:

- Student academic success; academic goals, objectives, outcomes; content mastery, test scores, learning; effective instruction;

Figure 52. Codes and identities Evaluation-as-Accountability.

Stakeholder narrative and analysis.

Teacher narrative. Teacher narratives regarding the *Purpose of Evaluation* incorporated aspects of both *Policy* and *Operational Intent*. Within *Policy Intent*, three out of seven teachers shared distinctions between *Evaluation-As-Accountability* versus *Evaluation-To-Improve-Practice*. Here, accountability is originating externally to their locus of control. It is imposed (actualized) by policy makers at higher administrative levels. It is generally seen as an accepted consequence of employment not fully equated with their own (idealized) personal belief system (i.e., The current use of evaluation versus how they believe it should be used).

Four additional teachers responded to the question of evaluation purpose by highlighting structural distinctions (*Operational Intent)*. That is, they raised concerns that the current system does not adequately assess all of the important attributes of good/effective teaching causing a fundamental disconnect between purported (idealized) intent and what is actually measured. The resulting bias distorts interpretation and raises question as to the efficacy of the policy-stated purpose of evaluation.

Teachers: Policy intent. As mentioned three teachers reacted to the *Purpose of Evaluation* question in terms of *Policy Intent*. Here, conflict originates between evaluation used primarily as a means of accountability, compliance, and punitive consequence versus a more benevolent tool to inform and improve teacher's professional practice. In this regard, one teacher expresses this basic conflict directly:

> What I think it [evaluation] should be is to really look at the whole teacher. But what I think it does is, it doesn't. I think that schools should do teacher evaluations to support a teacher, but personally from my experience, I feel like they do it to get rid of teachers. They want to find out who's gonna make it and who they got to cut. Who's in for the long haul, who's committed to what their belief system is,

> and who is not ready to rise to the occasion according to their belief system. Then you need to hit the road. Maybe you better find another profession, or maybe we shouldn't invite you back (Teacher 104)

Arguably, this individual represents the most pessimistic view of evaluation within the collection of teacher narratives. It is direct, emotional, and internalized. There is an "us versus them" tone to the perspective: "...I feel like they do it … they want to find out … their belief system…" The perspective reveals a basic mistrust in how the evaluation system is being utilized: "…get rid of teachers…who's gonna make… who they got to cut… who is not ready…" Evaluation is a tool for accountability, identification, sanction, and/or reprimand. It is punitive for teachers who do not conform or comply ("…not ready to rise to the occasion…"). This same teacher goes on to discuss what he/she believes evaluation should concern:

> Here's a weakness: how can we help you with the weakness? You need help with a class, how can we help you with classroom management? How can we help you with identifying your curriculum, aligning your curriculum, rigorizing your curriculum to the next level, higher level thinking? What are some of the tools that we can use to help you? … Picking out strengths and weaknesses, then let's focus on individualized professional development aimed to help that group of teachers … We need to be more individualized and specialized to what that teacher needs (Teacher 104)

Here, the desired purpose for evaluation is to help*,* support, and improve the professional practices of individual teachers ("…How can we help you…"). Evaluation serves this role through identification, "…picking out strengths and weaknesses…" followed by intervention, assistance, and training. In addition, evaluation for improvement is personalized, specific, and individualized ("…individualized professional development…"). For this teacher, instructional competence is an evolving construct aided by evaluation. There is an inherent sense of growth, development, improvement, becoming better at the craft of teaching.

Mistrust in the *Policy Intent* of the current evaluation system is also expressed by another teacher:

> Interviewer: In all that we've talked about, what is the purpose of teacher evaluation?
>
> Teacher 105: I think that at this point it has gotten to the point where teachers are evaluated on if we're doing this brand-new evaluation, I think it's getting more to the point of "are the students mastering what they're supposed to be mastering"? If not, then the teacher must not be effective.

Arguably, there are three key phrases in this response: "…gotten to the point…", "…brand-new evaluation…", and "…supposed to be mastering…" However, to examine them, a larger context to the teacher's perspective is first necessary. From narrative provided throughout the interview, a major concern held by this person is the evaluation system's overreliance on test scores, stating "… being the teacher being evaluated on their performance, you have no control over how that student is going to do that [state] test..." (Teacher 105). In addition, this teacher expressed concern on any teacher's ability to fully comply with all of the 22 components contained in the Danielson framework, stating "… I don't think that that [Danielson] is a true component of what a good teacher is. Not every teacher has all those 22 concepts ..." Finally, during part of the interview, the teacher reflected "… I know that I am a fairly decent teacher…" So, this person has a basic mistrust in the ability of the evaluation to process to adequately represent his/her instructional competence either due to issues with test scores or with the appropriateness of the Danielson ratings.

It seems reasonable then to suggest that these general perspectives help shape his/her response (above) regarding the *Purpose of Evaluation*. In this context, the phrase

"…brand-new evaluation…" seems to imply that the current structure is different from before and has been imposed. The phrase "… gotten to the point…" is derogatory in terms of improperly reducing evaluation criteria down to "…are the students mastering what they're supposed to be mastering..." where "…supposed to be mastering…" reflects an externally imposed standard. There is a sense of we have to do this or else. If teachers do not conform to the structure, "…then the teacher must not be effective…" Thus, the collective narrative provide by this teacher suggests that the perceived organizational purpose of evaluation is to measure whether or not students learn content as measured by test scores and/or whether teachers conform/comply to the Danielson criterion—which is seen as unattainable. Regardless, the organizational *Purpose of Evaluation* is not aligned to a more idealized conceptualization (not explicitly expressed by this individual).

The extended dialog provided by a third teacher (below) reveals this conflict between idealized and actualized purpose of evaluation. The person begins his/her reflection by first acknowledging the role of organizational accountability and then qualifying the perspective in terms of professional improvement. The individual begins by responding to the initial question of purpose:

Interviewer:	In that context, what would be the purpose of teacher evaluation?
Teacher 103:	The role of the evaluation, I guess, is to make sure that teachers are doing their job, which is a huge job, but it's not, for me, it's not everybody's in an Exceeds category [on the state test] and now you're the greatest person in the world. You could work at a school where there are a lot of low kids, but they're moving up - growth in lots of different areas. Maybe this year there's less bullying, hey, that's a growth.

The initial response is an acknowledgment of accountability ("… I guess … make sure … doing their job…"). However, the response is qualified by a personal perspective "…but it's not, for me…" suggesting a difference between idealized and actualized purposes. The caveat is based his/her valuation of test scores: "… not everybody's in an Exceeds category…" which, if they were, is inappropriately interpreted as "…now you're the greatest person in the world…" In this way, the individual seems to devalue achievement measures as insufficient, perhaps incomplete, indicators of instructional quality.

For this person, interpretation of quality is contextual ("…You could work at a school where…") where reification of quality is legitimately expressed by non-academic indictors such as "less bullying", justifying "…that's a growth…" In this way, achievement measures (test scores) are devalued, affective aspects of growth emphasized and other, unmeasured, metrics of impact (i.e., bullying) are important (i.e., "… lots of different areas…"). The individual continues his/her discussion of purpose, saying:

Interviewer:	So, the purpose of evaluation then is, in part, to make sure that teachers are doing what they're supposed to do?
Teacher 103:	Well, yeah, to make sure that you're following some sort of a curriculum, following state standards, looking at lesson plans to make sure that … I'm teaching this, and this teacher's teaching this, and we're kind of moving them [students] towards where they need to be.

Here, there is a sense of basic expectation in evaluation: "…following some sort of a curriculum, following state standards, looking at lesson plans…" On this level, evaluation assures conformance to established standards and getting students to "…where they need to be…" However, the teacher gradually expands on this foundation to include other, more affective, attributes:

> Then you need to have a good rapport with the kids and the families and your school. All of that kind of encompasses everything, so yeah, it's—it's so hard to describe because it's like, you're doing your job. You're talking to families, communicating, having a good rapport with kids, being able to disseminate the information to them, and they're able to give it back to you as best they can and be pretty successful at it, [and] if they're not, then reteach. All that together is what the evaluation, I think, is for. (Teacher 103)

The concept of evaluation is extended to relationship ("…rapport with the kids…) inclusive of "families" and a more global "your school" context. Then the teacher continues to extend the context into the concept of professional growth and support:

> I think it's [evaluation] for maybe even putting people on a [improvement] plan, if they need to, so we can grow teachers, especially like a new teacher who may be struggling, so that we can look at and say, "Okay, these people are struggling. Let's help them," not penalize them or make them feel like they have a scarlet letter or something. (Teacher 103)

Importantly, the use of the phrase "…on a [improvement] plan…" is not punitive. Rather, it centers on the improvement in professional practice ("…so we can grow teachers…"). The intent is to identify, help, and/or support "…a new teacher… struggling [teachers]…" as opposed to sanctioning or penalizing ("…help them, not penalize them…"). Indeed, the accountability aspect of evaluation is believed to harm professional identity by making teachers "…feel like they have a scarlet letter…"

To clarify this balance between accountability and improvement, the teacher reacts to an additional probe:

Interviewer:	What do you think's more important in an evaluation, accountability/conformity to the standards or the growth and professional development of the teacher?
Teacher 103:	That's what I would wish that the evaluation would be … your skills are good here, here, and here, but maybe communication with families is something that needs to be addressed more. How can we do that and see if we can move you along?
Interviewer:	Help to develop?

Teacher 103: To help develop the teacher, yeah.

Interviewer: You said "wish". That leads me to believe that you have a sense that it may not be being used like that, but more in an accountability framework than a development framework?

Teacher 103: Yeah, I think it's a numbers game kind of a thing … You just sign off on it. All right, I'm done. Whew, I'm done for the year, kind of thing.

Interviewer: Is there not enough dialogue - continuing dialogue?

Teacher 103: I just don't think there's enough time for the administration to even get around back to it. I just see that they're kind of running around with—chickens with no heads on and trying to get it done maybe?

Interviewer: [The administrator] Comes in, does your evaluation, gets it on paper, signs it, it's okay, move on?

Teacher 103: Yeah.

So with additional discussion, the teacher (above) seems to move from commenting on the reality of organizational evaluation (i.e., the need to assess conformance to established standards/criterion) to a more idealized concept of evaluation (improvement). The individual internalizes a purpose of evaluation as a method of professional improvement, growth, and development saying, "…That's what I would wish that the evaluation would be … to help develop the teacher…" *Evaluation-As-Accountability* is a short-term organizational tool to ensure compliance to a foundational set of performance expectations versus *Evaluation-To-Improve-Practice* utilized to support a higher level, longer term, purpose of improving teachers as practitioners. In this way, the conflict is a concern over application rather than content, policy intent rather than structure/computation.

Finally, this particular teacher (103) identifies lack of time as the specific limitation preventing evaluation from being used as a method of improvement and support ("…there's [not] enough time…"). Here, lack of time prevents meaningful dialog and critical reflection between evaluator and teacher and reduces the entire process to "… a numbers game…" The implication is that without meaningful dialog and reflection, evaluation becomes compliance without the opportunity for improvement.

Teachers: Operational intent. As mentioned, four teachers discussed the *Purpose of Evaluation* primarily in terms of *Operational Intent*. Here, conflict arises between the stated purpose of evaluation and the perceived structural inadequacies of the implemented evaluation process. The construct differentiates between what actually is being measured versus what should have been measured.

The conflict concerns omission of attributes, activities, and/or impacts believed reflective of good/effective teachers. Structural omission leads to bias and inappropriate inference of observed evaluation results. As a result, perspectives regarding loss of control, externality, imposition, and compliance/conformance become embedded in the teacher's conceptualization of purpose. In this way, idealized and actualized intention is misaligned. This reduces evaluation to an accountability system that inadequately assesses instructional competence with the implication that evaluation will not facilitate improve professional practice.

In this regard, one teacher responds to the question of purpose in terms of positionality and component representation:

> Interviewer: Okay. In the context that we just discussed, what role does teacher evaluation play? What is the purpose of teacher evaluation?

Teacher 106: I think the purpose is for accountability to make sure that we are teaching the standards, that we're following what the government says that we need to do, and, also, that we're adhering to district policies and programs that we work into our school, like Character Counts, and that type of thing. I think, kind of, from the global standpoint it's to make sure that we are meeting all of those standards, and taking into consideration all of those things that we need to teach them.

The immediate response to the question of purpose is to characterize evaluation in terms of compliance/conformance to externally imposed standards "…following what the government says… we need to do… adhering to … make sure that ..." Externality is evident from terms "... the government… district policies and programs… global standpoint…. There also is a sense of loss of control, inevitability, and evaluation as part of the job (i.e., "… make sure… teaching the standards…"). Here, evaluation is not a product of the teacher but rather of outside policy, outside the classroom. However, then the same teacher goes on to qualify this initial response:

> But I don't always think that it looks at all those other pieces. It would be nice if there were an added component … I have a little boy, who, probably, won't ever perform very highly on benchmarks or AIMS, and he was giving his speech today about—he's a boy scout, and his speech was how to start a survival fire. I got emotional because I don't get to see that side of him very often because traditional standards-based education when it's pen and paper or pencil is difficult for him. I wish that there was a way to incorporate experiences like that that our kiddos have that are big successes for some that don't always have those same successes. (Teacher 106)

For this teacher, the conflict of purpose is structural: current evaluation components are incomplete and insufficient. There are "…other pieces… added component[s]… experiences like that…" that are missing from the evaluation. Omission creates a distinction between an actualized and idealized context ("... I wish that there was a way…"). The term "big successes" is being used to suggest that some impacts on

students are not quantified by test scores or other evaluation metrics. The overall implication is that the purpose of evaluation is unrealized due to limitations within the current system.

Another teacher initially sees evaluation as accountability, but then struggles to articulate a more qualified reflection:

> Interviewer: What is the purpose of teacher evaluation? Not the organization's definition—your definition?
>
> Teacher 102: The purpose is to be sure that teachers are teaching the standards, I guess. However, but in a way that students can learn them and use them successfully. Can a teacher do that? Can a teacher perform her skills—not skills—job, her job requirements … accurately, efficiently? There's another one I wanted to say but it's like on the tip of my tongue and I can't think of it [giggling].
>
> Interviewer: I think you were grappling with what should be the purpose of teacher evaluation, what should it be?
>
> Teacher 102: What should it be? Well … there's things that are not part of teachers' control, I guess I should say, that they're evaluated on. And it's not necessarily, I don't wanna say fair, cuz, ya know, life's not fair. There's not a lotta control but there's part - these things that are part of our evaluations that are not test scores. Like, a lot of our biggest concerns, test scores can't really, I mean, we do what we can to make sure that they do well on the tests, but somebody's sick or they miss part or you have a gifted cluster or a SPED cluster … Those are just some things that are concerning.

For this teacher (102), as with others, the initial response is to focus on the criteria specified by the existing evaluation framework ("…teaching the standards …") under an environment of accountability ("… be sure that teachers are…"). However, this conviction is uncertain, ending the sentence with "…I guess…" Then the initial statement becomes qualified by "…but in a way that students can … use them successfully…" He/she also struggles to differentiate skills from job requirements. Presumably, the

teacher is making a distinction between teaching standards as a skill from ensuring students can utilize that knowledge to some meaningful end as a job requirement.

Struggling to clarify his/her thoughts, the teacher admits "…there's another one I wanted to say … it's like on the tip of my tongue…" After further query, the teacher suggests there are components of the evaluation metric that are "…not part of teachers' control…" making the evaluation process unfair. In this context, by saying "…life's not fair…" the person exposes a sense of resignation, loss of control, over the evaluation context ("…There's not a lotta control…"). There is a sense of inevitability and compliance in the life's not fair statement. Thus, for this teacher, the fact that evaluation is used to measure compliance to established standards is recognition of reality, not personal agreement.

Further dialog reveals an underlying "… biggest concern…" which involves the use of test scores as an evaluation indicator. Here, the teacher uses the phrase "…test scores can't really…" as if to indict some aspect of test scores as not being fully representative of instructional impact. This is clarified by the phrases "…somebody's sick … gifted cluster … SPED cluster…" and "…we do what we can…" suggesting that score bias a core issue. Saying "…we do what we can…" reveals a frustration with this part of the evaluation structure.

In this way, the teacher (102) distinguishes between what the evaluation system is currently measuring from what it should be measuring. Facets/factors outside of teacher's control, compliance to an externally policy-imposed structure, and concern over score bias distinguish the individual's personal perspective of the evaluation system from its operational context.

A sixth teacher begins by sharing his/her perspective of an idealized evaluation structure:

> Interviewer: In that context, what's the purpose or role of the teacher evaluation?
>
> Teacher 101: I think the role of the evaluation should include where the kids started, … maybe their backgrounds, some sort of data that these students only came to school this amount of time, or these students were low in these areas, and then having some sort of evaluation in place that everybody agrees with that we can say, "The attendance for the student rose, so that was part of you being inspirational, part of you talking with the parents, spending the time, and so you get this amount of points for that."

The immediate response is to focus on what "…the evaluation should include…" suggesting the presence of missing and/or unmeasured elements of good/effective teaching. There is an emphasis placed on growth, change, or development ("…where the kids started…") and a need for "…some sort of data…" not currently utilized by the evaluation process. In addition, the focus is on non-academic aspects of learning such as attendance ("…their backgrounds … students only came to school this amount of time… low in these areas …). To affect change teachers need to be "inspirational" and "… talking with parents … spending time…" In this way, the attributes of effective teaching need to include affective components of professional practice, relationship, and personal connection. The same teacher states:

> That's, I guess, coming from a teacher, so it's like I want to be evaluated on my counseling methods, my motivation, my inspiration, and I also want to be scored on if I'm a good teacher or not. (Teacher 101)

Arguably, it is unclear whether the closing phrase "…I also want to be scored on if I'm a good teacher or not…" is meant to summarize the collection of all missing attributes or a separation between these and another, independent, collective of attributes.

Regardless, the teacher (101) seems to internalize the context of *Purpose of Evaluation* in terms of structural misalignments between idealized and operationalized intent—it cannot measure what it claims to measure because of important missing components. By extension, the inferential context of interpreting evaluation results becomes suspect.

Unlike his/her colleagues, a seventh teacher (below) did not explicitly respond to this discussion prompt with the same clear distinction between actualized/idealized conflict of purpose. Indeed, the initial dialog presented seems to endorse an *Evaluation-as-Accountability* sentiment without much secondary qualification. However, as argued below, this becomes an insufficient, simplistic, interpretation of the teacher's perspective. By reviewing the more extended narrative shared from other parts of the interview, a more nuanced reflection on the *Purpose of Evaluation* is revealed. Reacting to the initial prompt:

Interviewer:	Okay. From your feelings on the purpose of education and the role of the teacher in education, what should be the purpose of teacher evaluation?
Teacher 108:	Well, I think ultimately, we need to make - we, as in the district or the state - needs to make sure the teachers are doing what they're supposed to be doing, that they're not just showing movies every day or [just] handing out a worksheet and saying, "Here, do it. The directions are at the top. Figure it out."
Interviewer:	Okay, so there's an accountability piece and that kind of thing?
Teacher 108:	Yes. Yes.

The initial reflection seems to align directly with an *Evaluation-as-Accountability perspective*, where the legitimate purpose is to ensure that teachers are following established standards and performance criterion. The statement is supportive and indicates agreement with its fundamental purpose ("…we need to make [sure]…

supposed to be doing…"). In addition, there is a perspective that teachers benefit from having explicit evaluation standards by providing clarity in practice: "…not just showing movies every day… handing out a worksheet… Figure it out…" In this context, *Evaluation-as-Accountability* seems to be the primary perspective for this individual.

However, it is also apparent that perspective positions the locus of control, power, and/or decision making for *Evaluation-as-Accountability* external to the classroom and external to the teacher: "…we, as in the district or the state - needs to make sure…" The qualification originates the *Policy Intent* for *Evaluation-as-Accountability* on policy makers at the district or state level. Thus, while the response may suggest accountability is important to external policy makers, it remains unclear whether this individual is responding strictly in terms of compliance or from internalized personal agreement. (Readers Note: Unfortunately, follow-up probes meant to clarify this distinction were not initiated at this part of the interview).

In this regard, other narratives shared by this individual (108) provide additional perspective, specifically comments shared during discussions on the *Purpose of Education* and the *Role of Teachers in Education*. Previously responding to the question on the *Purpose of Education*, the participant shared the following:

Interviewer:	To begin, what would you say is the Purpose of Education?
Teacher 108:	To educate, to promote a society that can continue to flourish because if you don't have an educated society, then it starts to fall apart because it turns into mayhem. [Laughter]. I would say that education is a way of continuing to prospering a society.
Interviewer:	Okay. Could I ask you to elaborate on the word "educate"?

Teacher 108: Mm-hmm. I would say teaching basic skills like adding, counting, that will be life skills, but then also teaching, or having a student learn, how to problem solve or learn those complex things that you aren't born with.

Interviewer: Okay, so a content component and then some sort of life skills component in a social setting?

Teacher 108: Yes. Yes. Yeah, how to interact with others, but then when there's only one milk left, what are we all gonna do when we all want milk? Those kinds of things.

Interviewer: Is that a non-content component? How would you weight those two in the role of education? You started out with the statement of society. Are they equally weighted in the purpose of education?

Teacher 108: I don't feel like they are anymore … I feel like … we're only focusing on very specific skills that may not be skills that are always used … I feel like some of those focuses are being crammed on people.

Interviewer: Okay. Who's doing the cramming, in your mind?

Teacher 108: I feel like it's a top down thing. I do feel like it comes from government, which has forgotten maybe what it was like to be in school … They don't quite get it, especially since so many schools are so different from each other. Even in the same district, schools can be just night and day between parent involvement, what the kids are eating, those kinds of things. I definitely think it's a top down thing, and it just trickles down to, ultimately, to the students.

Interviewer: So, the more sociable skills that a person needs to work in society? Those are imbalanced?

Teacher 108: Yes. We're emphasizing maybe the content or the skill and not the human that's the student.

From the dialog, the individual values the purpose of education as providing a larger set of learnings in addition to academic content. There is a distinction between "…basic skills like adding, counting…" and social learnings that support functioning in society such as problem solving, interact with others and "…when there's only one milk

left, what are we all gonna do when we all want milk? Those kinds of things. …"). There is a perceived imbalance of emphasis, stating "…We're emphasizing maybe the content or the skill and not the human that's the student…" and this imbalance is "…being crammed on people…" The cramming is being proliferated "… top down…" by government who "…don't quite get it…" and who don't recognize the importance of nurturing developmental/life-skills attributes needed by selected groups of students ("…so different … just night and day … parent involvement … what the kids are eating …").

By implication, *Evaluation-As-Accountability* is a perspective originating from policy actors external to the school/classroom. By over-emphasizing academic content and the pedagogical skills and practices to deliver this content, actualized evaluation might be viewed as incomplete and insufficient. Additional dialog shared by this teacher (108) regarding the *Role of Teachers in Education* is also informative:

> Interviewer: In that context, what's the role of a teacher?
>
> Teacher 108: I wanna say that a teacher's supposed to be a facilitator, but sometimes, it is literally just managing and making sure that we're awake and we have our pencil and we're doing those kinds of things.
>
> I kinda feel like the concept of being a teacher has kinda lost it's shape… I feel like sometimes it's making sure they [students] understand that they're in a safe environment, making sure that they understand … and that almost parenting sometimes becomes more important to managing the classroom
>
> Now, it seems a lot of it [the focus of teaching] is that pencil and paper, bubble test. You can have the student in the room do a concept, and they do it wonderfully. Then they take one test, and that ultimately decides if they know it or they don't know it. There's no background. You don't know what

happened to the kid on the way to school. You don't know if he had a fight with his mom on the way to school …

Interviewer: Can you be a good teacher and have students not learn the content?

Teacher 108: Yes. Well, there's that whole concept, you can lead a horse to water, but you can't make them drink. … There's too much—we're humans. The teachers are humans, and the students are humans, and I think there's too much - I don't wanna say - science. There's too much personality sometimes that gets in the way.

The individual (108) is highlighting the complexity of teaching by including student-specific, non-academic, attributes either not measured by test scores and/or are non-academic in nature. In addition to being a content provider, the role of the teachers takes on the form of facilitator, pseudo-parent ("…almost parenting…"), and engineer of a "... safe environment…" where these facets "... sometimes becomes more important to managing the classroom…" He/she seems to argue that in its current form, the measure (accountability?) of teacher quality has been inappropriately reduced to "…that pencil and paper, bubble test…" that provides "…no background…" of the student's personal context ("…what happened to the kid on the way to school … if he had a fight with his mom…"). There is a perceived bias in accountability systems which emphasize test scores since students "…take one test… that ultimately decides…" the collective impact of good/effective teaching.

Arguably, this teacher (108) provides an acknowledgment of *Evaluation-as-Accountability* as a perspective imposed by external power centers. It might be argued that his/her personal sentiment is whether or not the current implementation is sufficient for this purpose (operational intent). From the larger narrative, structural issues including omitted attributes and measurement bias (test scores) suggest a disconnect between

actualized and idealized intent. In this regard, this person may be indirectly acknowledging the reality of an externally-imposed structure versus a more personalized view for an improved system.

Teacher summary. Regarding the *Purpose of Evaluation*, it is argued that teacher narratives are distinguished by two primary themes: *Policy Intent* and *Operational Intent*. Dialog focused on *Policy Intent* concern issues of why evaluate? This is distinguished by two sub-components, *Evaluation-As-Accountability* versus *Evaluation-To-Improve-Practice*. Discussion related to O*perational Intent* concerns issues of what to evaluate. Here, the focus is on structural integrity (i.e., measuring the proper attributes of good/effective teaching) and measurement quality (reliability, bias). Across all dimensions, teacher sentiments are characterized in terms of idealized (i.e., what should be the purpose of evaluation and/or which attributes should be measured by the system?) versus actualized (what is the policy-imposed purpose of evaluation and/or which attributes currently are measured by the system?). A graphic depicting these relationships is provided in Figure 53.

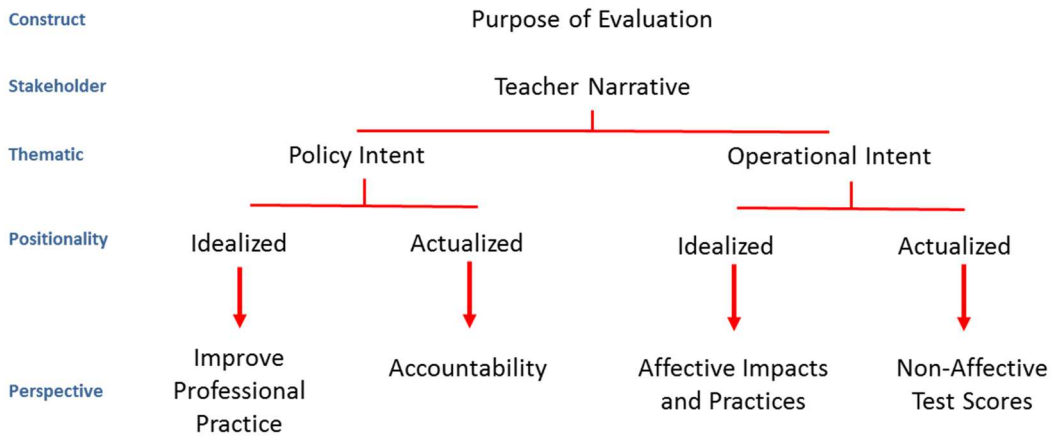


Figure 53. Teacher concept components.

Teacher narratives suggest conflict within both *Policy* and *Operational Intent*. Within *Policy Intent*, teachers view the current (actualized) implementation of evaluation as an externally mandated system of accountability. The purpose is to identify and sanction. *Evaluation-as-Accountability* is primarily seen as punitive for the purpose of identification, compliance, and conformity. Teachers are resigned to *Evaluation-as-Accountability* as a condition of employment. In contrast, teachers believe (idealized) evaluation should be used to support/facilitate improved professional practice.

Under *Operational Intent*, teachers exposed concern over the components currently measured by the evaluation system versus components that should be incorporated. These included affective impacts on students and within professional practice. In addition, they expressed bias/reliability concerns with test scores used as a primary indicator of instructional quality. On the basis of these structural concerns, the stated organizational purpose of evaluation is not manifest in the current implementation.

Principal narrative. As discussed below, all principals conceptualize evaluation as a means for improving professional practice. However, the context of this differs substantively from those shared by teachers. Principals view evaluation primarily as a means of monitoring teacher practices in terms of the measured evaluation components. In this way, *Evaluation-as-Accountability* is a fundamental activity. It is perceived as necessary in order to realize other benefits such as instructional improvement and enhanced student learning.

Principals, unlike teachers, have a fundamental trust in the evaluation structure as a measure of instructional quality. This trust empowers principals to believe that compliance/conformity to the evaluation process/components is beneficial and ensures professional competency. In this way, accountability is not punitive since one must adhere to the established standards/criterion to be a good/effective teacher. Identifying areas of poor performance provides opportunity to become better at his/her craft.

Seven out of eight principals interviewed failed to raise any structural such as omission, bias, or inadequacy. Thus, for this group, *Evaluation-as-Accountability* and *Evaluation-to-Improve-Practice* are equivalent concepts where idealized and actualized implementation is not in conflict.

The narrative shared by principals is provided below. One principal expressed the following regarding *the Purpose of Teacher Evaluation*:

Interviewer: Okay. What's the purpose of teacher evaluation?

Principal 204: It's a process by which two or more people work together to improve and refine instructional practice, to help someone move forward in their professional growth. The purpose is to help everyone grow, and to be honest, I had a two hour post on Friday with my eighth grade science teacher, and I think I grew

as an administrator throughout that two hour conversation. In order for me to help guide and facilitate learning and conversations, I'm continually rethinking … I have to be reflective as an evaluator to make sure that I'm continuing to grow, and refine my conversations as far as giving that feedback.

This principal sees evaluation as a means to improve professional practice: "…work together to improve … refine instructional practice… help someone move forward … help everyone grow…" Here, evaluation provides benefit to both the teacher and evaluator ("…make sure that I'm continuing to grow…"). The principal sees his/her role as providing feedback, "…help guide…" the teacher. By doing so, the principal is taking responsibility to "…facilitate learning…" by improving his/her capacity to provide meaningful feedback and reflection.

Similarly, a second principal (below) situates evaluation primarily as a means of improving practice:

Interviewer: What role does teacher evaluation play in all of this?

Principal 207: Teacher evaluation plays I think a good role in that it's always nice to have someone come in and give an observation and say, "Hey, how do I do that? How do I take the steps? Help me." Then, there's the other piece where it's, you know what, this piece is good for student learning and you're doing it. You're doing a really effective job at it. Keep doing that.

Interviewer: I heard a mentoring activity, and then I heard the opportunity to give positive reinforcement?

Principal 207: To me, they're one and the same as far as growth. I wanna know what I'm doing right. I wanna know what I'm doing wrong. I'm not perfect, and I never will be, but I wanna get better. I wanna keep getting better.

Interviewer: Is that the primary reason we do teacher evaluation? Provide feedback and input into their craft?

Principal 207: Sure. Absolutely.

As before, this principal sees his/her role as providing feedback, reflection, and support ("…You're doing a really effective job…"). Here, the evaluator is serving the role of mentor, providing positive reinforcement and support ("…Keep doing that…"). The same individual then goes on to raise the concept of accountability. However, even here, the focus remains on improving instructional practice:

> Principal 207: Then, there's the, to me, the accountability piece, too. I mean cuz what we're talking about is coaching … As a coach, we practice things. We set up things …. Now, let's execute it and work on it and practice those scenarios.
>
> Then, you've got the game [Reader Note: the principal was using a sports metaphor during the discussion] and there's your formative assessment. Let's see how we do. You use the same approach in the classroom … We got our little assessments, and we got a big one at the end, maybe with the game so to speak, and maybe a big one with a unit test or something like that.
>
> I think that's the piece where having a mentor helps, having a principal that wants to help you helps. It could be another teacher that's experienced that's helping you. Those things all are really, really positive.
>
> Interviewer: It provides useful feedback to the teacher to self-reflect?
>
> Principal 207: Yep.

This principal (above) equates evaluation with support, mentoring, and professional growth. In this way, purpose aligns to *Evaluation-to-Improve-Practice*. However, this conceptualization is also fundamentally embedded within *Evaluation-as-Accountability*. Here, "… the game…" is a sporting metaphor for instructional action/delivery. Game-day preparation is accomplished through a prescribed plan implemented through repeated practice of evaluation components: "…we practice things.

We set up things… Now, let's execute it…" Instructional efficacy is measured by "… little assessments… unit test…a big one at the end… Let's see how we do…" (i.e., the game's outcome/score). Interestingly, (post-game) mentoring serves as an important component ("… having a mentor helps…"). The caveat is that related benefits are ultimately dependent on the individual principal's level of commitment: "…having a principal that wants to help you, helps…"

Overall, *Evaluation-as-Accountability* is cast as a necessary-but-insufficient concept of purpose where post-evaluation mentoring serves as the main catalyst for effecting improvement. For this person, mentoring is extended beyond the principal-as-evaluator context to include "… another teacher that's experienced that's helping you…" Regardless, the core benefit of evaluation becomes mentoring, reflection, and guidance from any source, stating "…Those things all are really, really positive…"

A third principal (below) views the activity of evaluation as a means to gather information/data in order to provide feedback to teachers. In this context, *feedback* references the pedagogical attributes prescribed by the Danielson framework. The participant states:

> I think the purpose of teacher evaluation is so that they can get feedback on what they do, on their practice. I believe that I go in there to be able to give the feedback in all their [Danielson] domains, and that they need to know where they're at. Where am I, as a teacher, and what do I need grow in. I think our teachers need that feedback - good, bad or indifferent. They need to know how they're doing. They need those pats on the back. They need those pushes, so yeah, I think that's the purpose.
>
> Years ago, I was a teacher for 16 years, I got this evaluation. It meant nothing. My principal met with me. I got check marks and I was good, I think. They always said nice things but no one ever really told me what else. What else could I do? Because my kids were well behaved, because I took on leadership roles. I

would be such a better teacher now than I was then because of all the learning I've done. (Principal 205)

From the narrative, the purpose of evaluation is to provide teachers "…feedback on what they do…" This feedback is prescribed by the Danielson components. The fidelity and importance of these components remains unquestioned and unqualified where teachers "…need to know where they're at… Where am I, as a teacher…" and assessing performance on these components answers the question "…what do I need grow in…" In this way, the standards-based performance ratings are emphatic, deliberative, and consequential. Ratings of poor performance indicate need for improvement. The purpose of evaluation becomes the provision of these data.

Importantly, this principal (above) positions teachers as passive recipients of evaluation data. The data provides them structure, clarity, and rating/magnitude. The implication is that without this, teachers would lack focus, guidance, and direction. Evaluation is beneficial because "…teachers need that feedback… need to know how they're doing…" the latter implies that without evaluation teachers would *not know* how they are doing. Arguably, it transfers the responsibility of professional competence from the teacher to the evaluation structure, the measured attributes, and the performance (rubric) criteria. In this way, the pathway between accountability and instructional improvement is deterministic.

In contrast, the next principal is more directly focused on evaluation as a method of accountability:

Interviewer: Okay. Excellent. Given that, what's the purpose of teacher evaluation?

Principal 201: The teacher evaluation piece, for me, is to ensure that teachers are following policy, following procedures, providing a

> consistent education, holding students accountable for learning in an environment where learning is expected.
>
> When I go and I evaluate a teacher, my purpose is to make sure it's not to do what's best for that teacher. It's to make sure that that teacher is doing what's best for those kids, based on a rubric
>
> With this teacher evaluation process, the teachers, if they're doing their due diligence … they know what they should be doing. It should be consistent, and not just this one day. I don't want a dog and pony show. Any time I come in, you know what you're expected to be doing.
>
> [However] It's not to catch you [as in] "you weren't doing this, you weren't doing this, you weren't doing this." My main goal is to show them all the great things they are doing and continue those practices … and let's discuss how we can get you there … If I need to support you, if you need some professional development, if you want to go look at another teacher, watch another teacher who does this very well. It's not only to help them understand, it's to help me know what I need to do to support that teacher.

In this dialog, the evaluation components serve as foundational criterion upon which to assess practice ("…based on a rubric…"). The main context is "…to ensure… following policy, following procedures … consistent education… and the focus is exclusively on "…doing what's best for those kids…" and not "…to do what's best for that teacher…" In this way, the Danielson framework is the core criterion for all activities and teachers "… know what they should be doing…" There is no ambiguity or uncertainty in what is expected both of practice and evaluation. Here, the primary benchmark for collective success is student content mastery, saying "…learning is expected…" and it is the evaluation system which ensures proper practices are being conducted.

At the same time, the *Evaluation-as-Accountability* sentiment is qualified in terms of being supportive of teachers saying "…It's not to catch you…" rather, it is "…to show them…" Regardless, the Danielson criterion, components, and performance requirements remain the standard of quality. As such, is not just" … to help them understand…" it is to ensure that teacher practice is in strict conformance to the rubric because this is the standard for good/effective teaching. Arguably, *not to understand* is equated with non-compliance to the rubric. In this way, accountability (conformity, compliance) and improvement are inextricably linked in purpose.

As before, a fifth principal anchors his/her perspective on the *Purpose of Evaluation* to the Danielson components:

> Interviewer: Okay, okay. Given that context, what's the purpose of teacher evaluation?
>
> Principal 203: Well, I think with the model, the Charlotte Danielson model, it really breaks it down to those pieces that need to be in place in order to ensure that students are learning. If we look at each individual piece, it's almost like a piece to a puzzle of that whole child … it is about pedagogy. Really the teacher needs to be informed in each of those pieces to ensure that they've planned accordingly, all leading to student learning and the data to prove that.
>
> That rubric is our guide to determine what is it that we have to do, and strive to do, to be the best that we can, so it's not a guessing game anymore. It really comes down to what evidence do we provide to ensure that we are working towards having our students master the content.

For this individual, evaluation is about conformance ("…those pieces that need to be in place…"), accountability ("…ensure that students are learning … evidence…"), and ultimately that "…students master the content…" There is an unwavering trust in the evaluation structure - the principal unconditionally accepts the Danielson components as

delineating the primary attributes of good/effective teaching: "…really breaks it down … pieces that need to be in place …" Indeed, these attributes cover all aspects of "…that whole child…" By this authority, evaluation both defines and assesses pedagogy where the causal logic is unquestioned: The Danielson framework articulates instructional practice which, if followed with fidelity, leads to student learning. By implication, deviation harms student learning because "… That rubric is our guide … it's not a guessing game…" This belief permits the individual to further share, "I think [evaluation] helps guide teachers … to work towards becoming that distinguished teacher" (Principal 203).

This latter dialog seems aligned with an *Evaluation-to-Improve-Practice* purpose: "…guide teachers… becoming that distinguished teacher... " However, this interpretation is too restrictive. Arguably, the principal's (above) perspective of purpose fundamentally differs from teachers, who reject the *Evaluation-as-Accountability* premise. For teachers, the purpose of evaluation should be about improvement, and not about accountability. But for this principal, *Evaluation-as-Accountability* is the foundational perspective that permits (authorizes) an *Evaluation-to-Improve-Practice* sentiment. For the principal, teachers cannot improve without fidelity to the evaluation criterion.

A sixth principal also integrates this connection between accountability and improvement by viewing evaluation as a tool:

> Interviewer: Okay. What is the purpose of teacher evaluation?
>
> Principal 208: Teacher evaluation is just that. It's like a check and balance. It's making sure. It's a tool that we use to monitor teacher's progress in that facilitating of learning … I think it's a way for them to use [evaluation] as a professional development tool, as well. As you're checking and monitoring; what are you doing

with that information? … To be able to help that teacher become better themselves … We want to continue to grow and learn to be the best that we can be.

Interviewer: Helping them become better at their craft?

Principal 208: Correct.

Interviewer: Through the process of evaluation?

Principal 208: Yes.

The immediate response from the principal is to characterize evaluation as an accountability tool: "… [it] is just that… check and balance… making sure…" Evaluation is "…a tool…" used to monitor teacher practice in order to assess the "... facilitating of learning…" In addition, evaluation provides information necessary to assess the need for professional development "…as you're checking and monitoring…" Thus, evaluation gathers the data according to established criteria, which then identifies areas of needed improvement facilitated by allocating targeted professional development. Arguably, the perspective of purpose is anchored in a perspective of *Evaluation-as-Accountability* because targeted improvement is determined from the criteria imposed by the evaluation structure. This is a deterministic relationship that is unquestioned—evaluation data provide the means to "…help that teacher become better themselves…" and by following the evaluation structure with fidelity teachers "…learn to be the best [they] can be…"

A seventh principal also depicts evaluation as a tool and a means for teachers to improve practice by following the Danielson performance components/criteria:

> It [is] a tool and a rubric. I think our Danielson's system is really effective in that, because it has to start with the planning and preparation. There's also the professionalism piece. Then in between that is what are you doing in that

> classroom for those kids and your instruction, and so you put all those together. I like that model.
>
> This [Danielson] is the model our school district has decided to use. To me it makes perfect sense, so you explain that to the teachers. An effective teacher is a good planner. They're looking at all the different aspects of planning. They're pulling in resources. They've got a plan.
>
> … It's then the leadership's role to work in conjunction and collaboration with the teacher to say, "Did we meet our goals and the outcomes?" If we're down here on unsatisfactory or basic [on the Danielson performance rubric], how do we then work together to get you [the teacher] moved up to that distinguished or proficient? (Principal 202)

As with other principals, this individual's narrative positions the evaluation structure/process as a de facto standard for good/effective teaching. There is no sense of limitation, omission, bias, or inadequacy: the "…Danielson's system is really effective…" Authority originates in completeness, coverage, and appropriateness of the evaluated attributes: "…the planning and preparation… the professionalism piece… your instruction…" By putting "…all those together…" the system constructs a complete profile of good/effective teaching. There is also a sense of conformance because "... [Danielson] is the model our school district has decided to use…" However, this is not the same type of resignation as expressed by teachers. For this principal, use of the Danielson framework "…makes perfect sense… I like that model…" Here, imposition is a good thing where idealized and actualized perspectives of purpose are the same and not in conflict. *Evaluation-as-Accountability* is used as a primary performance measure: "…Did we meet our goals and the outcomes…" The data informs on quality and serves as the basis "…to get you [the teacher] moved up to that distinguished or proficient [level]…"

Finally, another principal (below) responded to the question of purpose strictly in terms of accountability. However, this individual differed from his/her colleagues by making a distinction between his/her idealized versus actualized perception of the evaluation framework. The participant begins the dialog saying:

> Interviewer: Okay. What's the purpose of teacher evaluation in that context? Why do we do teacher evaluation?
>
> Principal 206: It's supposed to measure whether or not our teachers are meeting the needs of all kids. So, whether or not her instruction or his instruction is incorporating high-level questioning for our kids who are there, small groups for our students who aren't there, and their general instruction is at a high level that's moving our middle kids. That's what it's supposed to do

The perspective of purpose seems unequivocal and direct: evaluation is to ensure "…the needs of all kids…" are met. This is accomplished by exhibiting specific attributes such as "…high-level questioning… small groups… high level [general instruction]…" The context seems to be related to achievement: "…kids who are there… students who aren't there… moving our middle kids…" There is an accountability context of outcomes as the defining criteria of success/failure. However, the participant ends the statement with "…That's what it's [evaluation] supposed to do…" As mentioned, this is the only principal that qualified the context of purpose in terms of conflict between an idealized and actualized perspective. In this regard, the person goes on to say:

> Interviewer: I'll let you go on to what it actually does [laughter], then.
>
> Principal 206: It gives snapshots of how a teacher is doing at that time, and what I don't think can be recorded in a document is the feeling of a classroom, positive or negative, 'cause I will have a teacher who, according to the rubric, is doing everything, but … when I go in her class, I can feel like a - it feels like a dark gloom. If I'm feeling it, obviously, the kids she sees feel it as well. Are they having that positive self-worth? Do they feel like I'm [the teacher]

> excited to come here every day and want to be here and want to do anything for this teacher?

In this dialog, the term snapshot seems to expose concern between the current evaluation process and a more comprehensive, inclusive, approach. The element not captured is "…the feeling of a classroom…" Teachers may be "… doing everything … according to the rubric…" and be rated highly, but still not address important aspects of the learning environment. In this way, the actualized evaluation process is incomplete and produces a less than a full representation of instructional quality. The omission is structural and thus this person's conflict concerns operational rather than *Policy Intent*.

For this participant (above), omission of classroom environment attributes do not merely reflect lapses in professional practice. Consequentially, classroom settings that "… feels like a dark gloom…" negatively impact students because "…the kids … feel it as well…", harming their "…positive self-worth…" raising questions such as do they [students] "…want to be here … want to do anything for this teacher..." From an evaluation perspective, professional attitude, commitment, and passion are the responsibility of the teacher (i.e., are teachers "…excited to come here every day…"). By implication, these attributes should be measured by the evaluation process.

The same individual goes on to say:

> Then right across the hall, I have a teacher who teaches at a very high level. Honestly, her kids would bend over backwards to do anything for her … when they're in her class, just her positive environment, the kids will do anything for her.
>
> I have another teacher who, her kids start off really slow, but they love coming to school every day. They kiss their brains every day. They have a real sense of community. I can try to emulate that as much in the [the Danielson component for] creating a positive environment, but some things can't be recorded, so it's hard. (Principal 206)

The individual feels that the current system does not sufficient opportunity to capture "…creating a positive environment…" because "…some things can't be recorded…" The following exchange clarifies this perspective of omission:

> Interviewer: All right. Would you be saying that there's a missing thing in the current structure that you wish or like to be able to put into that evaluation context that you're feeling that's not in there sufficiently?
>
> Principal 206: Yes.
>
> Interviewer: Okay. That's interesting, 'cause one of my questions would be, "Is there a component of good teaching that's missing in the evaluation activity?" I think you're saying that intangible piece?
>
> Principal 206: Right.

Principal summary. Overall, principals feel that the evaluation system acts to improve instructional practice. It does so through a process of accountability that ensures compliance/conformance to practices believed consistent with good/effective teaching. By identifying poor performance, teachers are provided opportunity to improve practice and increase student learning. Thus, *Evaluation-as-Accountability* and *Evaluation-to-Improve-Practice* are indistinguishable.

Fundamentally, principals hold an inherent trust in the evaluation system's ability to properly identify and distinguish good/effective teachers. The process and underlying structure of the system remain unquestioned. With one exception, there is no conflict between idealized and actualized implementation perspectives; only one out of the eight principals interviewed believed that certain aspects of classroom environment were not adequately captured. Finally, since the system is comprehensive and implemented with

fidelity, there is little/no conflict between the policy and operational intent of the system. These perspectives are distinctly different from those shared by teachers.

District narrative. District participants primarily see the purpose of evaluation as a tool for improving instruction. Evaluation systems provide opportunity to identify, communicate, dialog, and reflect on practice. The narratives also suggest some disconnect/concern with how such systems are actually developed and/or utilized.

One district-level member reacts to the basic question of purpose, saying:

> Interviewer: Okay. In that context, what's the purpose of teacher evaluation?
>
> District 301: I believe the purpose of teacher evaluation is to provide feedback to teachers in order for them to improve practice, to perfect what they're doing, to grow as educators … We should be constantly learning, changing, looking at our practice, so that we can get better at what we're doing. You can always get better. We can always grow. We can always learn.
>
> … [It is] Not only feedback but opportunities for dialogue for their [teachers] own self-reflection … A huge part of evaluation should be about my own opportunities for dialogue on my practice: what I'm doing, why I'm doing what I'm doing.

The narrative expresses an *Evaluation-to-Improve-Practice* sentiment where the main utility of the process is to "… provide feedback…" In this perspective, evaluation information forms the basis from which to "…improve practice, to perfect… to grow…" The main mechanism for improvement is opening "…opportunities for dialogue…" and "self-reflection" Indeed, critical reflection is seen as a "…huge part of evaluation…" The same individual goes on to say:

> There are times when you have to perhaps provide supports for a teacher that they don't think they need, or that they don't want. They need to be built in. They

> need to be in place because they aren't as effective as they should be or need to be. I think evaluation provides you a structure to be able to do that. Then to make decisions on whether or not the teachers are in the right field and really doing what they should be doing. There are dismissal purposes for evaluation. I don't think that should be the focus, or the biggest percentage of why we use evaluation, but it is a necessity, part of the process. (District 301)

The second purpose of evaluation for this individual is accountability. Here, some teachers' self-perception may not align to the evaluation information: "…they don't think they need, or that they don't want…" In these cases evaluation data serves as factual authority: "…because they [teachers] aren't as effective as they should be…" That is, the evaluation process trumps personal sensibility regarding evidence of professional competence. The implication is that teachers who do not (or are not able to) adjust their practice to affect higher performance ratings need to make decisions regarding employment ("…the right field… what they should be doing…"). The caveat is that this person devalues *Evaluation-as-Accountability* saying "…I don't think that should be the focus…" Regardless, accountability seems to be an accepted reality of the evaluation environment.

Overall, for this person, there is no explicit distinction made between an idealized or actualized purpose of evaluation; there is no sense of what should be versus what is. Rather, evaluation properly facilitates instructional improvement. It does so by identifying areas of needed focus, stimulating dialog and critical self-reflection. However, in some instances, it also must be use as an accountability tool when teachers fail to recognize or accept evaluation results as indicators of poor performance.

A second district participant expresses a similar, multi-dimensional, purpose for evaluation, saying:

Interviewer: Given that, what's the purpose of teacher evaluation?

District 303: I think any kind of evaluation is, the purpose of any kind of evaluation is to improve and to continue to refine your craft. I see a teacher evaluation as no different. I think the best teachers in the world have room to improve and I think that we all continue to have room to improve. When I was a principal evaluating teachers, that's the approach that I took - my role in evaluation is to help you get better at what you do. We're going to do that through discussion, and through observation, and in a collaborative process—in a joint process—and not just me dictating to you.

As before, this person's immediate response is to characterize the purpose of "… any kind of evaluation…" in terms of professional improvement: "…to improve… refine your craft…" The role of the principal/evaluator in this process is to ensure that improvement is realized: "…my role … is to help you [the teacher] get better at what you do…" Similar to the narrative provided by his/her colleague above, the mechanism for improvement is primarily "…through discussion, and through observation, and in a collaborative process…" Thus, targeted dialog and reflection are made possible by the information provided from evaluation process.

The same district member (303) adds a second, but related, dimension: goal setting and progress monitoring:

> It's also a measuring stick of how we're doing in terms of goals that we may set … We'd set goals together and then evaluation would be to monitor those goals and see how we're doing. That's what I see as the purpose. You look at just the word evaluation in general. How do we evaluate our kids? Well, we evaluate our kids based upon what goals we want them to accomplish during the course of the year … Teacher evaluation, I don't think, is any different than that. (District 303)

Here, evaluation is equated to "…a measuring stick…" set against the "…goals that we may set…" The presumption is that the evaluation structure specifies performance attributes/criteria that permit teachers to "…set goals…" In this way,

evaluation becomes a method/tool to "… monitor those goals…" However, this narrative is not really about accountability (to the extent expressed by the previous district participant) because there is no sense of punitive consequence other than to inform on general efficacy, "…see how we're doing…" The perspective seems more in line with monitoring and adjusting rather than sanction and high stakes consequence.

As before, this individual (above) does not provide indications of conflict between idealized and actualized intent or between policy/operational intent. *Evaluation-to-Improve-Practice* seems to be the dominant perspective that evaluation currently fulfills. There is no distinction offered between what should be versus what is the purpose of evaluation.

A third district participant continues with this *Evaluation-to-Improve-Practice* perspective initially stating:

> We need to use it to help our teachers to continue to build their skills and their ability to get the job done. Teacher evaluation should be a way to give teachers feedback, a way to design that professional development to meet the individualized needs of the teacher and the personalized needs of the teacher. It should be a way for us to gauge where we, [district name], as an organization, are in having that prepared workforce. (District 304)

As stated, the initial response regarding purpose is improvement, to "…help our teachers… build their skills… design that professional development" The context is both organizational ("…having that prepared workforce…") and personalized ("…individualized needs…"). However, the narrative differs from previous district members by using terms like "…We need to use…" and "… evaluation should be…" This provides a sense of conflict between the idealized and the actualized; that what should be and what is are not the same. The individual explains:

> If we say we want them [students] to be productive, we want them to be skilled, we want them to be mentored, facilitated, to design, we should be doing all those exact same things as part of our teacher evaluation system. Students' evaluation and teacher evaluation should, I do believe, be these two pieces that run parallel but support each other as well. (District 304)

The context of the statement is "…if we say we want…" then "…we should be doing…" implying a distinction between what is and what should be the purpose of evaluation. The conflict seems structural rather policy-derived, suggesting that intent is not in question; rather, it is how evaluation is constructed that is faulty. This qualification becomes clear in the remaining portion of the narrative:

> You know, they really do [that] now, [but] for all the wrong things … we are using that reading test and that math test to say, oh, student A, you fit in that little hole, and student B, you fit in that little hole, and we're doing the same thing to the teachers. We're doing a fine a job in doing that, but unfortunately not using the right tools. (District 304)

Thus, the concern is not with the policy intent but with operational intent. Conflict occurs because of "…not using the right tools…" The criticism is that the system is "…using that reading test… that math [test]…" inappropriately to "…fit [teachers] in that little hole…" The context is one of reduction, restriction, and simplification based on insufficient measure. The phrase "…we're doing a fine job …" is sarcastic in tone, meaning the evaluation process does a good job at imperfectly assessing instruction quality. Thus, for this person, there is a distinction between an idealized and actualized evaluation structure. Arguably, operational intent (an attempt to fully assess professional practice) is compromised by means of omission—not measuring the right attributes. Thus, this person's perspective of purpose diverges from his/her two colleagues (above) because the current system is not described as being fully realized and properly implemented.

A fourth district participant (below) responds to the question of purpose from a different perspective. The narrative presents a clearer conflict between the idealized and actualized, criticizing the disconnect between a larger purpose of education from its reified, practical, form. The person begins the discussion as follows:

Interviewer: All right. What's the purpose of teacher evaluation in that context?

District 302: The teacher evaluation is, probably, not constructed, or intended to be constructed, to address your socio-emotional, your affective natures, of a teacher's daily responsibility. I think the teacher evaluation system talks about lesson preparation; lesson delivery.

The immediate reaction was to focus on what teacher evaluation was "… not constructed, or intended…" to measure the "socio-emotional", the affective, dimensions of professional practice. Thus, the way the reified system is constructed and/or implemented is flawed by omission. Its measures are incomplete, and therefore, present an imperfect representation of instructional quality. Injustice is performed by reducing professional practice down to "…lesson preparation; lesson delivery…" Finally, there a sense of external imposition in the comment. By saying "…evaluation is, probably, not constructed…" suggests that, if empowered to do so, this individual would have constructed evaluation differently. For this individual, the power center of the evaluation structure lies outside the immediate context of the school/teacher. Clarifying, the same person goes on to discuss:

Interviewer: Our teacher evaluation, would you say?

District 302: I'd say, generally, nationally recognized teacher evaluation systems focus on instruction and learning to determine the goodness or the effectiveness of a teacher. A teacher evaluation system puts a numerical value on that realm of a teacher's job performance.

Interviewer: Would you be extending that statement to say, in some ways, that the way we do teacher evaluation as a practical thing in public schools, is incomplete?

District 302: Yes.

Interviewer: Because it doesn't get at these other dimensions of the child's life and what a teacher's actually expected to do, versus what we measure they do?

District 302: If you go back to the beginning of our discussion about the purpose of education, the teacher evaluation system satisfies your globally competitive; your industry demands on education. We do not have a teacher evaluation system that addresses the other part of the purpose of education, that talks about a child's experience, on a daily basis, coming to school. The teacher evaluation system may say certain cultural or environmental parts of the classroom experience, but it really doesn't get into evaluating whether or not a student has a good experience that day. (District 302)

For this district participant, there is disconnect between an idealized and actualized evaluation process. The actualized system attempts to quantify professional practice by placing "…a numerical value…" on "…a teacher's job performance…" The problem is that these quantifications are incomplete. Misrepresentation is manifest when the reified system is compared to the idealized "… purpose of education…" Here, the reified system is being determined by "global" competiveness and "industry demands" while the idealized version incorporates "…the other part of the purpose of education…" which concerns "…a child's experience…"

District summary. District participants present a generally consistent view of the purpose of evaluation which is primarily to improve the instructional practice of classroom teachers. A catalyst for this purpose is increased communication, dialog, and reflection. Three of the four individuals added aspects of accountability, but devalued this

as a primary purpose. Here, accountability was viewed more in terms of setting/attaining goals and informing on areas of poor performance areas to affect improvement.

Two of the district participants expressed disconnect between idealized and actualized accountability systems. One individual reflected on how "…we need to use it…" versus current utilizations. Here, the main issue was over-reliance on standardize test scores that tend to "…fit [teachers] in that little hole…"—a reference of reductionism, and over-simplification of instructional practice. The other individual made very clear distinctions between current evaluation criteria and the types of attributes such systems should be measuring. In this case, current systems are externally constructed and imposed and omit important affective dimensions of practice. In this way, such systems present an incomplete representation of instructional quality. In addition, these externally-developed systems are biased toward economic/industry purposes of teaching and fail to align with a broader sense of purpose which emphasizes socio-emotional aspects of learning.

State narrative. In the narrative presented below, two of three state-level members offered instructional improvement as the primary purpose of evaluation. To do this, evaluation first informs on performance and identifies areas of needed improvement. This permits teachers to focus on targeted skills/behaviors which, in turn, improve instruction. In this way *Evaluation-to-Improve-Practice* and *Evaluation-as-Accountability* are indistinguishable. For the third state-participant, *Evaluation-as-Accountability* becomes the dominant purpose. However, the perspective is qualified by the role of the decision maker. That is, state level decision makers require a narrowly defined accountability measure that serves as a quality check of academic attainment. In contrast, a (local)

school administrator may see the purpose of evaluation as a tool for enabling conversation about daily instructional activity. In this way, the *Purpose of Evaluation* becomes contextually dependent.

One member of the state policy group responds to the question of purpose as follows:

> Interviewer: In that context [the Purpose of Education and the Role of Teachers], what does the evaluation process do?
>
> State 401: The purpose of teacher evaluation is to motivate and inform a teacher so that they can move to higher levels of performance … The [purpose] is also that they're doing a good job, so that's the way that evaluation should end up. It should properly let you know all the things that you're doing well, but then say, "Hey, here's some areas in which you can pick up the pace"

For this person, evaluation is about informing and motivating teachers to "… higher levels of performance…" by indicating whether or not they are "…doing a good job…" In this context, evaluation is not being positioned strictly as accountability. That is, there is no mention of negative consequence, sanction, or punitive action. Rather evaluation itemizes performance (i.e., "…all the things…") in order to target areas of needed improvement (i.e., "… pick up the pace…"). The same individual goes on to say:

> The tragedy of teacher evaluation is that quite often, if it's done without really thinking it through, it can actually deflate teachers and demoralize them, and they can actually move them to lower levels of performance. That's your potential, potential to set teachers on an improvement path and/or to set them on a path in which they're demoralized and the systems grind them down, year by year, until they get ground down to lower levels (State 401)

Here, the quality of evaluation is dependent on the procedures and environment in which it takes place. Good evaluation is purposeful (i.e., "…that's your potential…"). In this way, effective evaluation is dependent on the choices made by those implementing the system (i.e., "… if it's done without really thinking it through…"). Consequentially,

poorly administered evaluation systems "…can actually deflate teachers and demoralize them…" with the effect of moving them "…to lower levels of performance…", the exact opposite of its intended purpose. The implication is that unintended, negative consequences of evaluation may be mitigated and prevented. Negative effects are no longer an artifact of the act of evaluation, but of the persons implementing the process.

A second state participant (below) reflects:

Interviewer: In your mind, what's the purpose of teacher evaluation?

State 402: The purpose of teacher evaluation would be to provide feedback to the individual, the teacher, in terms of how he or she can improve in all of those areas to the benefit of the students they serve.

As before, *Evaluation-to-Improve-Practice* is the primary sentiment. Here, evaluation is a tool to generate information ("provide feedback ") in order to target "… all of those areas…" of needed improvement. When a teacher improves in these areas, it "…benefit[s] of the students they serve…" In this way, evaluation data provides teachers the opportunity to improve. The assumption is that teachers take advantage of this opportunity to the betterment of students. The person then comments:

State 402: Now, obviously, being a member of society and the community might be less important to a kindergarten teacher's role, though I do think there is some responsibility there, than it might be to a middle schoolteacher, but the purpose of the evaluation is to acknowledge the good things that that teacher is doing, and to identify areas of potential improvement.

Interviewer: For improving the profession?

State 402: Right. To give that teacher the feedback that is motivating to that person. It is also to acknowledge the competencies that that person has exhibited and the qualities that that person brings to the work setting, and to express appreciation to that

person, in addition to helping that person grow in his or her role.

Arguably, this narrative continues to devalue the *Evaluation-as-Accountability* perspective in favor of *Evaluation-to-Improve-Practice*. Indeed, evaluation is specifically purposed to provide "…that teacher the feedback that is motivating …", "…acknowledge the competencies…", and "…express appreciation…" These qualities are "…in addition to…" those of "…helping that person grow…" Interestingly, the individual discusses the *Purpose of Evaluation* in terms of attributes not readily reified by the system: that is, the extent to which instruction facilitates "…being a member of society and the community…" (It is noted that these attributes are not explicitly specified within the Danielson framework nor assessed by items on the state achievement tests).

As a follow-up question, the same individual was asked to reflect on the legislator's perspective of evaluation:

Interviewer: In your opinion, what was the perspective of the legislature regarding the purpose of Teacher Evaluation, and is it different than yours?

State 402: Oh, I think it's very different than mine. I don't think that they [legislators] gave much thought to the purpose of evaluation when the legislation was introduced and ultimately passed. I think what their purpose was - was to be consistent with the education initiatives being played out in some other states. What the [legislators] were thinking is, "Let's get this stuff passed because that will put us in line for Race to the Top money." That was the purpose.

For this person, the view of legislative intent is reduced to acquiring funds ("Race to the Top money") and conforming to "… some other states…" evaluation models. The comment seems pointed and meant to be critical.

The narrow/simplistic focus suggest state legislatures did not "… [give] much thought to the purpose of evaluation…" The implication is that the teacher evaluation system has a broader purpose that deserves careful reflection that was not afforded to Arizona's efforts. The person continues:

> The idea was that if we can do away with tenure, … limit, reduce, or eliminate or modify the RIF [Reduction in Force] policies to not speak to seniority, if we can pass legislation about keeping kids in third grade if they can't pass the test, if we can pass legislation that says "we need teacher and principal evaluation systems", then Arizona would be in play to apply for the Race to the Top money because you couldn't even apply unless you had certain things in play (State 402)

This clarification indicates that in order for states to have been considered for additional Race to the Top monies, they had to first implement numerous legislative (education reform) initiatives aligned to federal expectations. This perspective is consistent with interpretations advanced by published public policy literature, noting that Race to the Top constituted a competitive grant program orchestrated by the White House and the U.S. Department of Education (Weiss, 2014; Nicholson-Crotty & Staley, 2012; White House, 2009; Viteritti, 2012). To be considered for the for a grant, applicants were required to show compliance to specific education reform efforts, including legislating structured teacher evaluation systems inclusive of nationally recognized professional practice standards and use of achievement measures (Manna & Ryan, 2011). The individual continues:

Interviewer:	[So] the bill going through the legislature and on to the State Board, and all the implementation events, was heavily driven by the desire to access to the Race to the Top money?
State 402:	Oh, with 99.9 percent.
Interviewer:	If the Race to the Top grant money was not out there, it wouldn't have happened?

State 402: Right … the other one tenth of the percent, I think, really came out of the political structure of the group of very conservative governors [and] ALEC, that group who were pushing legislation to state legislators across the nation, not just Arizona. That "here's a bill, it's already written, just put your state's name on it. Here it is. Now you go introduce that."

I think the initial motivation on the part of state [policy makers] had to do with the dangling of Race to the Top money. The motivation at a more broad level, I think was entrepreneurial, and I think it still is because when you look at who's funding ALEC, and I can't think of the name of the other group right now, it's people who also have a stake in online learning, in evaluation testing companies. I believe that we are selling the profession, in essence, to the private sector.

Interviewer: Did it also feed into an ideological framework of accountability, education reform? Measurement? Test Scores?

State 402: Yeah, I think that sounds good … I don't think that there is the respect for the work that public education has done … I don't believe that they [legislators] actually have thought through this, but the sense is that if we pound people over the head enough, our kids will do better.

Here, the connotation is that the purpose for implementing Arizona's teacher evaluation legislation is multi-faceted including obtaining Race to the Top money (the "initial motivation"), political influences emanating from external public interest groups ("… conservative governors…"), ALEC (the American Legislative Exchange Council: http://www.alec.org/) who were "pushing legislation", and private profit interests (i.e., testing companies). For this person, the collective impacts have had the effect to "… selling the profession [education, teaching], in essence, to the private sector…"

Arguably, from the narrative provided, this individual (above) is making a clear distinction between an idealized and actualized purpose for evaluation. The idealized perspective relates to informing and improving professional practice while the actualized

serves political and economic interests. Indeed, the logic of the legislative perspective is described as "…if we pound people over the head enough, our kids will do better…" That is, accountability leads to improvement.

A third state policy member (below) at first depicts the purpose of evaluation in terms of academic accountability, but then qualifies this perspective suggesting purpose differs by stakeholder group (i.e., state vs. local levels). Reacting to the question of purpose:

> Interviewer: What is the purpose of teacher evaluation?
>
> State 403: It's a way to try and capture how a teacher is able to get that student [to learn] and measure the type of educational attainment that student has gotten under the tutelage of that person.
>
> Interviewer: Okay, so it's primarily to measure the effectiveness of the teacher in the context of what the student has learned?
>
> State 403: Yes. … [However] I think there are different levels of teacher evaluation. I think there's certain teacher evaluation processes that we as a state need to have in order to determine whether or not what the state has viewed as critical areas of educational attainment are being met, for instance mathematics.
>
> At a local level, maybe a principal has different criteria for evaluation that is different from what the state policies are. The principal might want to know if that teacher is also engaging students. Is that teacher somebody that makes the student want to come back to class the next day? … Is that teacher also a good teacher in the context of working with other teachers and being able to be part of a team?" There's different ways to evaluate teacher effectiveness that we should use

At first, the individual (above) is speaking as a state policy representative noting that, at this level, "educational attainment" is the primary metric of interest. In addition, student "educational attainment" is explicitly ascribed to teachers ("… under the tutelage

of that person [the teacher]…"). At the same time, the state member acknowledges that "…there are different levels of teacher evaluation…" and that state policy perspectives may not be the same as those "…at a local level…" because "…a principal [may have] different criteria for evaluation…" This person concludes that there are "…different ways to evaluate…"

Arguably, the narrative (above) suggests that any single, prescribed, method of evaluation is insufficient to serve all stakeholders, whose "criteria" for evaluation differs. It seems reasonable to extend this rationale to conclude that any singularly defined *Purpose of Evaluation* fails to inform on a more universal assessment of what it means to be a good/effective teacher because these definitions become relative to the interest group. That is, a state-level accountability perspective that focuses on math and reading scores might be substantively insufficient if the local focus of instruction is more affective than academic (i.e., increasing attendance, graduation rates, behavioral attributes, and/or the personal/emotional well-being of students).

The same individual continues his/her remarks, exposing a conflict between the need to evaluate teacher efficacy by a clearly delineated academic attainment criteria versus affective impacts good/effective teachers have on students:

Interviewer:	What I infer is that your sense of teacher evaluation is not simply accountability but for some other purpose as well. What would that other purpose be?
State 403:	I think simply the reason for [having] a quality teacher evaluation [system] is to constantly look at are we making academic gains on behalf of those students? Are we making sure that those students are being exposed to a teacher that is allowing them to make those academic gains?

> I'll be totally honest with you. As a parent, I put more stock in my daughter having exposure to those teachers that are the transformative teachers, even if it were at the expense of her academic success, because to me, I would value that more. I think she would be a much better person because of that. Now from the state policy … I have to say that we have the responsibility to show academic attainment against those core academic areas that it's important for us to achieve.

The statements seem to reveal a values struggle: the need to assess "academic gains" versus student's exposure to "transformative teachers." From narrative previously provided by this individual in RQ1E (a) and (b), the phrase "transformative teachers" is being used to describe teachers that effect fundamental change in students' connection to the learning process (i.e., instill a love of learning, desire to learn, self-motivation to learn, etc.). The personal belief is that this measure of teacher effect is superior to simple academic gain, saying "…I would value that more… even if it were at the expense of her academic success…" For this person, transformative teachers make the student "…a much better person…" The conflict is that, "…from the state policy…" decision makers "…have the responsibility… to show academic attainment…"

Arguably, for this state participant, the policy-imposed evaluation focus is incomplete and insufficient because it fails to capture the more important affective, personal, and transformative impacts good/effective teachers have on student's lives. Thus, systems that emphasize academic outcomes serve only a narrow definition of instructional quality desired by a select stakeholder group.

State summary. Narrative provided by state participants suggests the primary purpose of evaluation is to improve instruction. It does so by providing information that can be used to focus improvement efforts. Poorly implemented systems may have a profound negative effect on teacher morale, school climate, the learning environment,

and harm desired outcomes. Properly implemented systems recognize, reward, and motivate teachers to become better at their craft.

There is recognition that different policy groups may have different purposes for evaluation: for example, state policy needs may differ from those of local (school) policy or an individual parent. In this way, any single approach to evaluation fails to address the needs/perspectives of all stakeholders. Most important, academic centered evaluation systems omit important affective aspects of good/effective teaching. That is, emphasizing content metrics devalues transformative impacts: teachers who fundamentally enhance/strengthen student's personal and emotional connection to learning. In this way, a state-directed evaluation system may properly serve one purpose (accountability emphasizing educational attainment) and neglect another (personal, emotional).

Construct summary RQ1E (c). Stakeholder narratives concern two primary attributes of the Purpose of Evaluation: *Policy Intent* and *Operational Intent*. *Policy Intent* concerns questions of "why evaluate" while *Operational Intent* concerns issues of "what to evaluate." For both, the concepts of *Evaluation-to-Improve-Practice* and *Evaluation-as-Accountability* are distinguished. Finally, in discussing evaluation purpose, many stakeholders implicitly framed their reflections in terms of idealized versus actualized systems.

Teachers reflect that the primary purpose of evaluation should be to improve professional practice (Idealized). However, the (actualized) way the system is being utilized is believed to be for accountability purposes. Overemphasis on test scores, omission of important affective elements, and concerns over measure reliability/bias conspire to reduce teacher's trust in the system (i.e., the operational/structural aspects of

the system are questioned). In this way, the stated policy intent of the system is not aligned with its actual implementation and use.

Principals expressed satisfaction with the system as a means to improve instruction. Importantly, this group makes no substantive distinction between policy and operational intent. That is, the intended purpose is to improve instruction, this is the purpose it is actually being used for, and there are no substantive limitations in the operational structure or process for accomplishing this goal. In this way, idealized and actualized activities are aligned.

District participants also agreed that the purpose for evaluation is to improve instruction. This is accomplished via increased communication, dialog, and reflection on teaching practice. However, their idealized and actualized perspectives conflicted. This group cited an over-emphasis on test scores and (affective) structural omissions, both of which lead to incomplete representations of instructional quality. In this way, stated (Policy) and reified (Operational) intent do not align.

State participants reflect an *Evaluation-to-Improve-Practice* sentiment regarding evaluation purpose. However, it was recognized that poorly designed and/or implemented systems may actually harm teacher morale, school climate, and the learning environment. In this way, system implementation is just as important as structural characteristics. If the *Policy Intent* is to improve instruction, both the *Operational Intent* (structure) and system administration are critical factors to consider.

State-level stakeholders also exhibited conflict between idealized and actualized aspects of evaluation. Here, academically-centered systems fail to consider affective impacts on students (i.e., transformative teachers). In addition, it is recognized that

different stakeholder groups may hold uniquely different perspectives on the purpose of evaluation. That is, state policy actors may value *Evaluation-as-Accountability* by focusing on a narrowly defined set of academic-oriented indicators, school principals may value critical dialog/reflection on attributes of classroom instruction, and parents may value the personal/emotional context of education and the student's growth as a well-adjusted human being. In this way, any single structure/process may not reflect a uniform purpose and/or approach to evaluation. Finally, the policy intent of current evaluation systems is dominated by externally sourced economic and political concerns.

Petite assertions for RQ1E(c).

1. Evaluation-to-Improve-Instruction is the dominant perspective regarding the purpose for conducting teacher evaluations.
2. Substantive conflict exits within the teacher, district, and state groups regarding idealized versus actualized implementations of evaluation systems. Omission of affective components and over-emphasis on test scores is the primary reason for this discrepancy. Structural inadequacies lead to inadequate representations of instructional quality.
3. Teachers feel most strongly that the stated policy intent of evaluation (i.e., to improve instruction) substantively deviates from its actual use (i.e., consequential accountability).
4. Different policy groups value evaluation for different purposes. As such, utilization of a single structural framework inadequately serves the needs of all stakeholder groups.

Primary Research Question #2

How does providing an evolving dialog on validity evidence influence organizational policy decisions regarding implementation of the LEA's teacher evaluation system?

RQ2 (a) and RQ2 (b)

To what extent do policy-level stakeholders value the collection and review of validity evidences as an important input to the system's ongoing development? To what extent does validation evidence prompt changes in organizational decisions regarding system implementation? The approach is semi-structured interviews. The measures are coded interview responses.

Introduction

The reader is reminded of this research study's operational title: *Examining the Construct Validity of a State Policy-Directed Framework for Evaluating Teacher Instructional Quality: Informing Policy, Impacting Practice.* Research Question 2, and its supporting examinations, is aligned to the qualifying context of Informing Policy, Impacting Practice.

From the outset of the project, it was hypothesized that providing an ongoing dialog concerning the empirical aspects of the evaluation process would (1) be valued by decisions makers and (2) directly influence policy decisions. In this way, Research Question 2 was structured as follows:

How does providing an evolving dialog on validity evidence influence organizational policy decisions regarding implementation of the LEA's teacher evaluation system?

- RQ2 (a): *To what extent do policy-level stakeholders value the collection and review of validity evidences as an important input to the system's ongoing development?*
- RQ2 (b): *To what extent does validation evidence prompt changes in organizational decisions regarding system implementation?*

As discussed in Chapter 1, this researcher was responsible for designing the empirical structure of the evaluation activity. From the outset, this involved presenting ongoing empirical information to committee members in the form of analytic concepts, rationale, supporting published research, empirical data, and multi-faceted analysis. The role involved leading discussions, encouraging critical reflections, making clarifications, responding to questions (with additional analysis), assisting understanding, and facilitating the connection between the empirical analysis and the decision making process. Importantly, this researcher was not the primary decision maker on the committee. Rather, the role was to assist, facilitate, and support the decision making process. (Reader's Note: The district's Evaluation Committee was comprised of 12 individuals representing the following areas: five teachers, two principals, five central office administrators, and one representative of the teacher education association).

The process of providing empirical information to the committee was facilitated primarily through ongoing dialog (both formal and informal), meetings, and presentations with members of the district's Teacher Evaluation Committee. To lesser extent, similar activities occurred with additional members of senior administration who were not formal members of the committee.

To obtain a policy maker perspective for this research question, follow-up one-on-one interviews were conducted with two key members of the Teacher Evaluation Committee (n = 12). These persons were purposefully selected based on the following considerations: each had been part of the core decision/policy cadre from the outset of the legislated-imposed evaluation process (beginning approximately in SY2010-11); each occupied key decision making positions on the committee; each were responsible for engaging in extensive district-wide communication with teachers and administrators; and each were positioned within a larger (district-level) policy making context in the organization. No other members of the evaluation committee had this range of longevity, historical background/understanding, direct communication with stakeholder groups, decision making authority, or connection to district level policy.

Construct Overview

As mentioned above, Research Question 2 is subdivided into two supporting inquiries. The first is expressed as *RQ2 (a): To what extent do policy-level stakeholders value the collection and review of validity evidences as an important input to the system's ongoing development?* This is essentially a question of *value*, the value that policy makers afford empirical information when formulating perspectives and decisions. It is an attribute of importance, weight, and criticality. *Value* includes a subcomponent related to *Researcher as Information Broker*, the value placed on specialized staff members specifically responsible for providing empirical information and analysis to decision makers. This latter component is inclusive of this researcher and members of the district's research office staff. (Reader's Note: In his discussion of *Communities of*

Practice, Wenger (1998) uses the term knowledge broker to refer to individuals who facilitate the exchange of knowledge across community boundaries.)

The second is expressed as *RQ2 (b): To what extent does validation evidence prompt changes in organizational decisions regarding system implementation?* This is a question of *influence*—the degree to which empirical information affects change and/or modification. That is, has the provision of empirical information resulted in decisions that would not have happened in its absence? In this regard, Research Question 2 is conceptualized across these two components: *Value of Information* and *Influence on Decision Making*. Based on initial analysis of the narratives, operational definitions for each are discussed below.

Value of Information - RQ2 (a):

> Value is conceptualized as the importance placed by policy makers on the need to have access to, and interpretation of, empirical information for the purpose of (1) shaping thinking and (2) making decisions. Value is an attribute held by the policy makers themselves.
>
> Value examines the degree to which empirical information is seen as a necessary means to critically reflect, think, and formulate personal perspectives upon which subsequent policy decisions are based. Sentiments of high value place empirical information as a foundational requirement from which to authorize policy decisions. In contrast, perceptions of low value would position empirical information as secondary to other forms of decision authority (i.e., practicality, political expediency, personal sentiments, external interests, etc.).

Codes and identities delineating *Value* within the narratives include:

- trust
- transparency
- understanding
- important for making decisions
- foundation
- critical
- shaping thinking
- validation
- confirmation
- critical reflection,
- questioning
- high stakes requires high quality data
- consequential decision making
- expectation to use
- researcher as facilitator
- research team
- researcher as information broker
- informed
- data presentations
- data dialogs

Figure 54. Codes and identities delineating *Value*.

Influence on Decision Making – RQ2 (b):

Influence is conceptualized as the causal role that empirical information has in affecting decision making, change, and/or modification. In this way, information acts to justify policy action because such actions are predicated on data, evidence, and/or empirical foundation. There is a causal pathway between having access to information and making decisions based on that information.

Codes and identities delineating *Influence* within the narratives include:

- clearly defined
- active involvement
- data sharing
- evidence
- communicating
- increased focus
- increased accountability
- specific action
- change in training
- focus on rater reliability
- setting cut scores
- policy decisions
- evidence
- impact
- data driven decisions
- extensive review
- confidence
- trust in decision

Figure 55. Codes and identities delineating *Influence.*

Stakeholder Narratives

For the purpose of this analysis, reflections from the two committee members are designated by *Member 1* and *Member 2* so as not to compromise anonymity. Discussion of stakeholder narrative is sectioned into two main components: *Value of Information* and *Influence on Decision Making*. The *Value of Information* discussion is further examined under the concept of *Researcher as Information Broker*. The narrative data/analysis for each are provided below.

***Value of Information* and *Researcher as Information Broker*: RQ2 (a).**

Value of Information. One member of the committee (below) discusses the *Value of Information* in terms of the contribution it made to the evaluation development process, saying:

> Interviewer: What role has the empirical information played in the evolving context of the evaluation process?
>
> Member 1: I think it's played a role that it's supposed to play, in the fact that it's been something that continues. We go back to the data, we look at it, and we reevaluate what we're doing. We relook at it. It causes us to question, and to make decisions where we can improve it, or make it better, or we look at it differently than we looked at it before we got that data. It forces us to look at things maybe through a different lens, and to look at other options, and look at things. I think it's forced us to do those things, and it's caused us to make changes through the process as we've gone. I think as we continue to collect data, that we'll continue to do that; that this isn't a process that's done. It's an ongoing process.

For this member, review of information is an ongoing activity embedded in the evaluation design/decision making process. Saying "…it's played a role that it's supposed to play…" implies something that is fundamental, a "given," where the need for information is an accepted baseline from which to engage in the evaluation process. Data

is seen as a continual input to the process saying "…it's been something that continues… it's an ongoing process…" There is also a sense that the data provokes reflection and shapes thinking in a continuing cycle (i.e., "… we go back to the data, we look at it, and we reevaluate what we're doing…").

For this individual, reflection "…causes us to question…" and it "…forced us … caused us to make changes…" The sentiment is positive because the information directly impacted decisions so that "…we can improve it, or make it better…" In so doing, data is valued as both necessary and influential in decision making: "…it forces us to look at things maybe through a different lens…" and to "…look at it [evaluation issues] differently…" Here, the empirical information challenges preconceived perspectives and guides decision makers to new ways of thinking and questioning. Overall, there is a sense of benefit, of positive contribution, where policy decisions would have been less beneficial in the absence of empirical information.

The same individual goes on to say:

Interviewer:	So the process of gathering data, bringing it back, relooking at it, rethinking of it, you see that as an ongoing thing that's just part of the evaluation system—that needs to continue over and over?
Member 1:	Absolutely.
Interviewer:	Do you think that, from the board, to the cabinet, to the people on the evaluation committee, they valued the fact that we did so much analysis with the data, and reviewing it, and looking at it to shape our thinking? Do you think they valued it, the decision-makers?
Member 1:	Yes, I do. Do I think they felt the same passion as the committee and myself do about it? No, because I think they work from a broader scope … [but] do I think that they valued the fact that they trusted us to do it? The expectation from them and from leadership is that we would do what we did, and

> that we would use the data that we have to create the best practice that we could create here. It's valued as an expectation of us, because I think our cabinet and our governing board have a high level of expectation for—they have high standards, and they expect us to meet those standards.

From this perspective, using empirical information to make decisions is a basic expectation of senior policy makers: "…they trusted us to do it…" It becomes an "…expectation from them and from leadership…" to "…use the data…" to "…create the best…" possible evaluation system. The authority to base decisions on empirical information originates from "…a high level of expectation… high standards…" that the evaluation committee was expected to follow.

The second committee member (below) reflects on the value of the empirical information. This part of the discussion concerned the standard setting process whereby committee members were tasked to establish cut scores designating teacher performance classifications. The member initially comments:

> Interviewer: That whole discussion and analysis [of the empirical data] - that helped shape those decisions?
>
> Member 2: We could not have made a definite - we couldn't have made a decision on that [standard setting] unless we would have had all of that information and all of that analysis.
>
> Interviewer: I'm taking from your comment that you thought that the people on the committee, and the policy makers from district, valued that whole extended process we went through for two-and-a-half years of looking at the data, and that it influenced decision making?
>
> Member 2: I truly do. We had representatives from just about every segment [of the district]. It was important - it was an important investment of time, and everyone around that table, every time we met, I think truly had a perspective other than their own. [They] were able to see other perspectives because of all of that dialogue, conversation, and digging in the data that we were

> able to do, because we weren't - even the teachers saw a more global perspective. Administrators - we definitely put on our teacher hats again to see what it would be like in each category. Okay, well what if this happened, and this? We definitely worked on, I guess 'perspective' is the best word, because we wanted to make sure it was the most fair process that we could possibly produce.

For Member 2, the extended review of the empirical information provided all members of the committee with a similar set of information from which to arrive at a common set of decisions: "…we couldn't have made a decision … unless we would have had all of that information and all of that analysis…" Here, the decision process is facilitated by two components: information and analysis. The implied causal pathway begins with a common set of information which permits community dialog (analysis) leading to consensus decision.

By providing this information, a shared activity of "…dialogue, conversation, and digging in the data…" is facilitated such that "…everyone around that table… were able to see other perspectives…" In doing so, the team was able to reach "…a more global perspective…" Specifically, administrators were able to "…put on our teacher hats again…" Here, the value (power) of data is in presenting common understanding that transcends personal perspective. For this person, the benefit was to ensure the resulting evaluation system was "…the most fair process that we could possibly produce…" Presumably, fairness emanates from decision making within a context of common information, discussion, and analysis.

This same individual goes on to say "…I wish that our circle had been wider so that others could have experienced that [data discussions], but not everyone can give that level of commitment and time…" (Member 2). Thus, value is derived for anyone

participating this this type of activity. The value seems to be in the power of empirical analysis of facilitate understanding, implying that non-participants may not have the same depth of understanding as those involved in the data-driven dialogs.

Member 2 (below) expanded on the value of information during subsequent conversations regarding the decision to devalue the test score component of the evaluation framework. The reader is reminded that legislative policy permitted districts to assign between 33 and 50% to achievement measures and between 50 and 67% to professional practice ratings (Ariz. Rev. Stat. §15-203A.38, 2010).

Interviewer:	Did the analysis of the data help you resolve, or help you think through, the decision to devalue the weight of the test score component? Was that a decision based upon looking at the data?
Member 2:	I do think that all of that analysis helped us look at not just individually but then also on a bigger scale … Yes, having it all, [the] data on a piece of paper …
Interviewer:	Made it clearer, made it more succinct?
Member 2:	By separating it all out, because we weren't all mathematicians around that table, and we all had different levels of understanding of how data works. With those clear explanations of what the numbers meant, and by separating it out into all of the varying components, then we were able to understand how they all played into each other and how the individual factors created - whether it was weighted differently - that number that we eventually had to assign to a teacher.

Within the comment is a sense of equalization: the data brings everyone to a common understanding. At the outset of the policy discussion "…we weren't all mathematicians … we all had different levels of understanding…" but after "…clear explanations…" and by "…separating it out…" then the collective group was "…able to understand…" Here, the value of information is clarity, leading to common

understanding that enabled the team as a whole to determine the performance criteria "…to assign to a teacher…" The perspective is further clarified by the follow-up dialog:

> Interviewer: Right, … so the availability to look at the empirical data - did that help you come to that conclusion that we've weighted it properly because it's based on evidence rather than just how are we going to do it?
>
> Member 2: Yes, in our conversations and through your data analysis, we were able to, I think, at each meeting, become more confident that this was not just an arbitrary number. This is evidence based. This is truly the best we can do or your team can do to make sure that this number and this weight has a direct and true correlation to that teacher.

For this individual, the value of the empirical evidence is that it provides authority for the decision making process: "…this was not just an arbitrary number. This is evidence based…" As a result, the decision reached by the committee was "… truly the best we can do…" In addition, this evidence-based decision meant that "…this weight has a direct and true correlation to that teacher…" Again, the implication is that any weighting decision reached in the absence of empirical analysis would have been less than optimal.

Speaking about the same cut score decision making process, Member 1 shares the following reflection:

> Being able to look at the data, see it, and knowing for a lot of the people here it's going to apply to them directly, and the fact that they were able to look at it, see things, and then make decisions, and then we use the process to say, "All right, we can't just cut". We had the discussion "Are we gonna cut equally [and] the bottom quarter's gonna be ineffective; The middle quarters?" So, having those [conversations] and looking at it [data], and then going through the whole discussions about "where are the majority of them [teachers]?", "where are realistic [cuts] based on what the range of scores are?", and having them [committee members] understand that, because they go out and share with their peers.

In the context of the discussion presented above, Member 1 notes that the committee members are partly composed of classroom teachers who are the recipients of the decisions made by the collective team (i.e., "…it's going to apply to them directly…"). The value of the data is that "…they [teachers on the committee] were able to look at it, see things, and then make decisions…" In this way, the data helps transcend personal interests. In addition, the power of the data lies with the collective discussion (i.e., "…we had the discussion… having those [conversations] and looking at it [data]…") and its ability to equip the group with common understanding. This becomes an important value attribute "…because they go out and share with their peers…" In this way, information empowers everyone with common understanding and common voice that can be widely disseminated.

Member 1 also connected the *Value of Information* to building trust in the decision making process.

Interviewer:	Did it [information] bring a level of trust, did it raise the confidence? What did the data bring to the thinking?
Member 1:	Well, I do believe that it improved? Yes. Did it [information] improve trust? Did it increase trust? Definitely in the committee, because their understanding of the process and of how to include student achievement, and the fact that we continue to look at it [data] … You talked about fairness earlier, and trust. Those kinds of discussions, that's how trust is created. This committee worked on that.

For this person, an added value of information is its facilitation of trust among the team. Responding "…did it [information] increase trust? Definitely…" Here, trust is facilitated through "… their understanding of the process…" Understanding is attained when "… we continue to look at it [data]…" and have "…those kinds of discussions…"

It is concluded that "…that's how trust is created…" and the group holds trust in their decisions because "…This committee worked on that…"

Similarly, Member 2 echoed this aspect of trust associated with the empirical information provided to the committee, saying:

> Interviewer: Did the bringing of that early data engender trust or confidence that as an organization we were moving in the right direction and we were going to have a more accurate measure?
>
> Member 2: Absolutely. And in thinking back on our work over these last three or so years, each time we added another layer of information that was connected to that teacher evaluation piece. Then what we were able to do is look at the very narrow scope of the data, look at the evaluation, but then also look at the big picture. And as the process evolved we almost created this dimensional type of a continuum of where teachers fell. And then by separating out the teacher evaluation components, student achievement, [we were able] to see that alignment or misalignment depending on some teachers … there was some discrepancies.

For Member 2, trust developed over a long period (i.e., "…these last three or so years …") of ongoing information analysis where each step "…added another layer of information…" leading to an ever increasing understanding of the evaluation construct. The empirical information permitted members to "…look at the very narrow scope of the data…" while at the same time see the "big picture." Here, trust was evolutionary, progressive, increasing in proportion to the amount of information provided. Arguably, the narrative suggests that trust is heightened by the fact that the data did not uniformly show alignment between evaluation components (VAM and PP ratings): "…there was some discrepancies…" For this person, the fact that the committee members were able "…to see that alignment or misalignment…" seemed to engender higher levels of trust.

The value of information was also expressed in terms of transparency—that providing ongoing empirical data, followed by review/discussion, engendered a sense of completeness, thoroughness, and rigor in the decision making process. Member 1 speaks to this, saying:

> It was a transparent process. We brought everything [all the data], and we looked at it. It was transparent to staff. We communicated pieces to staff. Did it increase trust in our process district-wide? I hope so, because we've communicated, we communicated, we communicated.

From the narrative (above), transparency and trust are linked. There is a sense that the data presented a complete picture (i.e., "… we brought everything [all the data]…") facilitating analysis and understanding (i.e., "… and we looked at it…"). This link between trust and transparency extends to the general staff population because "…we've communicated, we communicated, we communicated…" The sense is that by widely disseminating the information the empirical context became "… transparent to staff…" and heightened trust in the evaluation decision making process. In this way, trust becomes dependent on three important elements: (1) access to empirical information, (2) review of this information, and (3) dissemination of the information beyond the decision making cadre.

Member 2 also responded to this idea of value within data transparency, responding:

Interviewer:	Do you think that doing all those discussions with the data and adding in the committee reinforced our goal of transparency and our assurance that we were doing the best that we could for teachers?
Member 2:	Most definitely, and I think - I'm not that familiar with everyone else's evaluation instrument, but I truly believe that it would be hard to find another district that put, that invested as

> much time and effort into our process. We did not jump into anything new. We were very careful and methodical in this process … [it] led to a more, a truer, reflection of what their [teacher] performance was … teachers [are] slowly becoming more empowered and more knowledgeable about how that evaluation process is driven. I do think so.

For this member, transparency and rigor of analysis are connected. The district's commitment to analyzing information is seen as unique (i.e., "…hard to find another district that put … as much time and effort…" Here transparency comes from deep analysis of the data (i.e., "…careful and methodical…") that leads to "… a truer, reflection…" of instructional quality. This process results in teachers "…slowly becoming more empowered and more knowledgeable…" presumably because they become more aware of the information used in the analytic process.

The general value of information is also expressed within the context of the high stakes, consequential, environment surrounding the evaluation process. Member 2 weaves this into his/her discussion, saying:

> I think as our conversations progressed we really were looking at our task of measuring the effectiveness of teachers, and that it is very subjective, so we wanted to take as much questionable, as many questionable, factors off the table as possible to get a true and accurate rating, because our job is a very, [it] holds a lot of weight, because putting a label, an effectiveness rating, on a teacher, connecting that to that teacher is daunting. It's huge. That could be life changing especially when we're looking at a RIF time.

The member sees the act of assigning performance labels to teachers as inherently "very subjective." The value of the information lies with its ability to help decision makers "… [take] as many questionable factors off the table as possible …" In this way, determining performance criteria (i.e., "…to get a true and accurate rating…") becomes less subjective and more anchored in evidence. The *Value of Information* is emphasized

because performance ratings "…hold a lot of weight…" making it a "daunting" task. The sentiment suggests that the impact of making a wrong decision is "…huge. That could be life changing…" (Reader's Note: In the narrative, the term 'RIF' refers to Reduction in Force, a procedure to prioritize staff layoffs due to budget reductions). The same individual goes on to reflect:

> I am thankful that I've been able to be on this committee because I have the benefit of knowing all of your data, that most people don't, and how each individual piece of the thousands of pieces all line up and support a rating. That's an ominous task.

Here, value is an attribute of scope, completeness, and rigor where "…the thousands of pieces…" of information collectively empower the decision maker. Member 2 sees himself/herself to be in a position of privilege (i.e., "… I have the benefit of knowing all of your data…") where the collective body of information "…all line up…" to "…support a rating…" This member draws comfort from his/her access to data because the consequential impact of assessing teacher quality is "…an ominous task…" The sentiment implies concern over causing unintended consequences from poorly-informed decision making.

Speaking in this context of consequential decision making, Member 1 (below) incorporates similar reflections (Reader's Note: this discussion concerned establishing the quantitative criteria for assigning teachers to an *Ineffective* performance category):

> Where do our teachers fall [on the scale]? How many people are in this [Ineffective category]? What's the ramifications of that [performance assignment], and the implications if we have this many people [in that category]? What's it gonna do to our staff morale-wise?" We didn't just look at it from the objective, analytical research component as far as mathematically, but then, what's the other side of the house, and how is that gonna impact our teachers, and what's it gonna do to them?

The dialog (above) exposes an issue within the state policy evaluation framework—a lack of operational context of what it means to be an *Ineffective* teacher. For Member 2 (above), labeling a teacher as *Ineffective* is a high stakes, consequential, decision that warrants close reflection. Here, the data offers only partial value. Other considerations such as "… ramifications … [impact on] staff morale … impact our teachers…" become important concerns in the decision making process. The value of information becomes contextualized within these other concerns. The consequential context requires decision makers to extend beyond the "… objective, analytical research component…" and consider "…what's it gonna do to them [teachers]…" Extending this sentiment, Member 1 goes on to say:

> We couldn't just look at it strictly from a objective thing. We had to take some of those other things into consideration. Then I think that shaped our feeling, which ended up in that final decision … Then, the mathematical component supports that as well, cuz we have the data in the majority of the cases that support that. There's a few outliers.

Here, Member 1 acknowledges that quantitative metrics support assigning an *Ineffective* classification "…in the majority of the cases…" but "… there's a few outliers…" Because data does not present a perfect representation of instructional quality "…we couldn't just look at it strictly from an objective thing…" Regardless, the empirical information becomes the foundation upon which to consider "…those other things…" and in so doing adds value to the decision making process. As with Member 2, Member 1 is concerned with making poorly-informed decisions that may have unintended consequences. For the most part, the quantitative measures lead one to correct decisions, but in some cases, additional considerations are required to ensure proper outcomes.

This person (Member 1) closes his/her general discussion regarding the *Value of Information* saying:

> Interviewer: So, your thinking was influenced by, or came out of, the work that we were doing and the data that we were bringing?
>
> Member 1: Well, the data that we were bringing, to the committee especially, I think the difference for me is being able to say in my heart "I know this is the right thing", and "I feel like this is something we should do". Data always is the evidence that supports what you think. The data definitely says, "Yeah, you're on the right track. You are thinking right," because the data is showing that there's evidence to say that that's something that we need to do.

Similarly, Member 2 concludes his/her reflections, saying:

> I wouldn't change the process. We met dozens of times. I have the privilege of being part of all those conversations … I think ours was a true and valid process because it is so steeped in the data. It's not arbitrary. This is truly tested over time.

Summary - Value of Information. For each member, the *Value of Information* originates from its ability to validate decision making. In this way, access, analysis, reflection, and communication (transparency) collectively facilitate trust and confidence in the decision making process. Importantly, evaluation is seen as a consequential activity, impacting teacher's professional identity. Here, the *Value of Information* is heightened by the high stakes nature of the evaluation process and serves as the foundation from which to question and consider intended and unintended outcomes.

Researcher as Information Broker. A prominent component within the *Value of Information* narrative concerned the *Researcher as Information Broker*. Here, value was ascribed to having a dedicated staff person/team responsible for collecting, assembling, and communicating information to committee members, district policy leaders, and to

stakeholders at large. Member 1 (below) shares the value brought by having a dedicated information broker as part of the evaluation design and implementation process:

> Before you joined our team, we had begun looking at what the teacher evaluation [was] going to look like when we had the student achievement component? … The committee was new, and they were researching it, and kind of looking at it. But the knowledge and the expertise that you and your team brought to us regarding data, and how you can analyze data, and [to] look at it fairly and objectively, completely changed the perspective of everyone on the committee.
>
> Where we are now with student achievement, and what we ended up doing with it, is very different than what I think anyone had any idea that we would be able to do when we were looking at it in the first place. I mean, at the very beginning, I think, because we didn't have deeper knowledge … we didn't have the knowledge specifically, nor did we have any of that research … we weren't even thinking that direction, because we didn't have anyone that had that knowledge to bring that to us … it made a significant change

In this context, the role of information broker is viewed as an important addition to the decision making process. The position injected specialized "…knowledge and the expertise…" not held by committee members. As a result, this added expertise "…made a significant change…" in terms of "… data, and how you can analyze data, and [to] look at it fairly and objectively…" Member 1 shares that prior to the addition of an information broker, members of the committee "…weren't even thinking that direction…" The reference concerns early discussions related to evaluation metrics, analytic/computational approaches, and models for estimating student achievement. In this way, the information broker provided value to the decision making process by adding previously unavailable skills, knowledge, expertise, and information such that "… [it] completely changed the perspective of everyone on the committee…"

A primary role of the information broker concerned communication of information to committee members and policy/decision makers. Implied within Member 1's narrative, information value increases when it becomes understandable and connected

to the decision context. Member 1 comments on this (below) when discussing the complexity of the data being considered by the evaluation committee, saying:

> I think that they [the District Governing Board] have a better understanding because not only did we communicate to staff, but we did many presentations to the board as we went through this process, and because you were able to explain it. This is regardless of our process, but, you individually are really good at being able to understand the mathematical components and explain them in terms where people understand them, which is a challenge when you're talking about this level of statistical analysis.
>
> It [the evaluation framework] is very complex, and for people to have a basic understanding of it is a credit to you. I think that that's something that you've done well, and we, as a committee, have done well, is provided opportunities for people to participate in conversations so they understand a very complex situation.

Here, Member 1 recognizes the need for information to be accessible to decision makers. By communicating complex information in understandable ways, the information broker provides "…opportunities for people to participate in conversations…" The implication is that without understanding, individuals cannot legitimately engage in decision making. Importantly, the role of the information broker is not to simplify (i.e., "…It [the evaluation framework] is very complex…"). Rather, the value afforded by the position is communicating complex information in an accessible manner: "…you were able to explain it… in terms where people understand… so they understand a very complex situation…"

Member 2 (below) also commented on the role of the information broker for enhancing the *Value of Information.* Speaking about the evolution of thinking through the evaluation design, Member 2 reflects:

> Three years ago, the information that we used to look at was the historical information that was pretty linear just as far as the baseline of where teachers were performing on the Danielson rubric. Once you brought in data that involved

> student achievement, and then over time … you'd bring in another little nugget of information. But each one of those layers brought out, I think, the different dimensions of teacher evaluation … and then how it changed, how that score or that designation changed, from year to year. But then there was the student achievement component, the value added, and then your thorough explanation of slicing and dicing it down to all of the different student components.

For Member 2, the information broker provided a continual flow of information that informed the evolving context of the evaluation system. Prior to this, understanding of the framework "…was pretty linear... " involving just the Danielson components. Then, the new legislative framework required the addition of achievement measures, which made the system more complex. Member 2 describes the evolving contribution of information and analysis in terms of "nuggets" and "layers" which helped to reveal "…the different dimensions of teacher evaluation…" The implication is that as more information is added, decision makers obtain a deeper understanding of the complex framework. Here, the role of the information broker is to provide a "…thorough explanation of slicing and dicing…" of the component parts. Continuing this discussion, Member 2 adds:

> Then another very important layer was connecting the teacher to every single student because it's very different when it's just okay this class at this time, but really which students did you directly impact? The evolution of connecting teacher with student data really progressed over the last especially the last two years because not only did you connect it to a school set, but then individual students within that grade level, and then what subject areas. The data that was connected to that teacher became more pure each time you did that, in that, it was truly, as closely as possible, an indicator of that teacher's impact on that student's learning for that year and in that context.

The context of the dialog (above) concerns the evolving thinking on to how best to align achievement measures to individual teachers (i.e., "…The evolution of connecting teacher with student data…" taking place "…over the last two years…"). The

information broker supplied published research studies, various computational methods, and statistical data associated with different approaches. Member 2 summarizes this saying "…not only did you connect it to a school set, but then individual students, within that grade level, and then what subject areas…" The value of the contribution was connecting teachers with classroom students such that it "…became more pure each time you did that…"

The implication (above) is that the information broker was able to construct metrics that afforded decision makers greater confidence in the data's representation of teacher instructional impact. This confidence was founded in the member's sense of understanding and comfort with interpreting the empirical outcomes. In this way, Member 2 shares that the information "…was truly, as closely as possible, an indicator of that teacher's impact on that student's learning for that year and in that context…" Thus, the Value of Information was enhanced by having a specialized information *broker* as part of the decision team.

Member 2 completes his/her reflection on the role of the information broker's ability to bring value, saying:

> I think the more we got in our conversations and the more you were able to extrapolate all of that information for us, and be as pure to the student, the process, and the teacher, as possible, and then by adding not just one test but by multiple tests and multiple years, then that validity increased and then the assurance that that was an accurate - as accurate a measure as we could possibly find - then the confidence level increased.

For Member 2, the information broker not only brought empirical information, but also facilitated member's understanding of its context: "…you were able to extrapolate all of that information for us…" In doing so, members' sense of "…validity

increased…" leading to "assurance" that the results were "…as accurate a measure as we could possibly find…" Thus, the *Value of Information* is contingent on understanding. The information broker enhances this value based on a unique set of skills, knowledge and the ability to facilitate this understanding.

Summary - Researcher as Information Broker. The information broker serves as a catalyst for developing and presenting information to the decision making team. This position brings a unique set of skills and knowledge not held by other members of the evaluation committee or other policy-level stakeholders. While the foundational role is the provision of information, this activity is seen as insufficient for adding substantive value to the decision making process. Here, the information broker facilitates understanding and clarity of interpretation by communicating complex information in a way that is accessible to decision makers. The result is heightened trust and confidence in the decision making process and outcomes (Wenger, 1998).

Influence on Decision Making: RQ2(b*).* Primary Research Question 2 examines the role that empirical information had on the decision making process. RQ2 (a) explored this context in terms of the value decision makers placed on empirical information. This included the role of the *Researcher as Information Broker*. Complementing this analysis, RQ2 (b) assesses the influence the empirical information had on implementation decisions.

For RQ2 (b), review of the stakeholder narratives suggest a number of impacts realized from the empirical information supplied to decision makers. These included substantive changes to the type/focus of evaluator training, an increased emphasis on interrater reliability, the decision to devalue achievement as a primary measure of

instructional quality, decisions regarding performance standards (i.e., cut scores), and an increase in critical reflection, communication, and dialog among program decision makers. Stakeholder narratives related to these areas of influence are presented below.

Changes in training focus. Member 1 (below) was asked to reflect on the influence empirical information had on shaping/changing evaluator training:

Interviewer: Okay. What changed in the focus of [evaluation] training over the last two years? Did anything change in how you wanted to target your training to staff?

Member 1: Yes. Significant changes have occurred in training. At first, when I was working with them [evaluators], it was just training on, "This is the rubric. This is our process. Do it." Now we really are training far more specifically on specific areas of the rubric, and how to collect that data … We've done a lot more of that kind of training specifically … Our training has evolved as far as actually doing a rating, where they watch a video, collect evidence, and discuss the evidence they've had.

Interviewer: Did the discussions that we had with some of the data, the early rating data, did those affect you're thinking about the need to do more training on the rating?

Member 1: Absolutely. We would have never gone that direction had we not had those [data] discussions in this committee and moved forward with where we are.

For Member 1, the initial review/analysis of the early professional practice ratings generated concern that evaluators may not have full command, and application, of the Danielson rubric. As a response, "…significant changes have occurred in training…" evolving from simple familiarity (i.e., "…This is the rubric. This is our process. Do it…") to more focused application in a real-world setting. Saying "…our training has evolved…" Member 1 describes the new focus as "… [where] they watch a video, collect evidence, and discuss the evidence…" Here, this change in focus "…would have never gone that direction had we not had those [data] discussions…"

For this person (Member 1), the influence of data also extends to a broader array of activities, stating:

> The kinds of things, the kinds of training that we've done, even with the student achievement component, all of it, the data collection we're able to do, the CES, and [the things] you're doing, and the reports that we've created for teachers that have the explanation of student achievement components, and the explanation of CES data, and how it's created to get their effectiveness - [all] those things would never have happened, either.

In this way, the influence of empirical information and the committee's analysis of the data impacted broader system design decisions (i.e., "…those things would never have happened, either…"). Member 1 (below) goes on to reflect on concerns raised by the empirical analysis, saying:

> Interviewer: Was that being driven by your concern that they [evaluators] may not be trained well enough, and it might be impacting their ratings?
>
> Member 1: Their ratings, exactly. Getting the professional practice ratings and looking at them, some of it through your processes [of data discussions]… The processes have changed over time in order to make it a better process, more relevant, more aligned to the rubric, and more defensible for them as an evaluator

In this way, review of the information lead to speculation that additional training was necessary. This prompted changes in the system "over time" which acted "…to make it a better process… more relevant… more defensible…"

Similarly, Member 2 (below) reflected about how these changes in training impacted his/her ability to become a better evaluator (Reader's Note: this person served as an evaluator of classroom teachers in addition to being a core member of evaluation decision team). Within this context, the individual remarks:

> I can say that as an administrator, 10 years ago, writing evaluations was something - it was just an obligation you had to do. The feedback was more generic. [Today], as an evaluator, I am much more purposeful, and try to be as

> exact as possible, and true to that rubric, and to providing authentic and specific and purposeful feedback to that teacher - to truly guide their individual growth. (Member 2)

For Member 2, changes in training have had a positive impact on the efficacy of the evaluation process. By becoming a better evaluator, a result of improve training, he/she has become "…much more purposeful…" with an ability for "…providing authentic and specific and purposeful feedback…" which helps "…guide [teachers] individual growth…"

Interrater reliability. During the dialog, Member 1 (below) extended the discussion of data impact and training to the issue of evaluator competency and attempts to both measure and improve interrater reliability:

Interviewer:	We made a decision to do an inter-rater reliability study to make sure that we're doing that well. Am I right that I think that that came out of our first look at that rating data and how it didn't vary? That started to raise a question; we better do good rating, right?
Member 1:	Yes … I think some of the things that have come out of our [data] discussions, and our data collection, and that kinda stuff, is the process that we've built in for [assessing] inter-rater reliability, where we have our administrators paired and doing partners, so that we can collect that data and analyze it and see it. We've done significantly more training regarding inter-rater reliability kinds of things, so that we can ensure there is consistency across the district. We're still not where I would like us to be, but we are much better than we were, as far as that goes.

In this statement, Member 1 connects the committee's decision to implement a formal interrater reliability study based on analysis of early rating information. This information was collected during the pilot year SY2012-13. The resultant reliability study had specifically "…come out of our [data] discussions and our data collection…" In

addition, the member acknowledges fundamental changes in the ongoing administrator training process, saying "… we've done significantly more training regarding inter-rater reliability…"

Changing perspective, instilling confidence and clarity. Review of the narrative suggests that ongoing review of empirical information impacted the perspectives held by decision makers. In addition, the commitment to data accuracy and ongoing analysis enhanced both trust and confidence in resultant decision (an attribute also discussed under RQ2 (a)). In this regard, Member 2 shares:

> We also compared teacher's level of experience and non-experience, and we saw sometimes there was a correlation and sometimes not. I think your use of [looking at] a cohort of teachers, and then you pulled data out. By looking at that subset, it really gave us a different perspective … we were able to see it on a smaller scale to make sure that what we were looking at was what we needed to look at … There were many checks and balances in place with the data team here as far as pulling accurate information, and what we would do then as a committee. … I think it increased the trust level, the confidence level…

In this narrative, Member 2 acknowledges that sometimes empirical data presents contradictory information (i.e., "…sometimes there was a correlation and sometimes not…") However, subsequent deconstruction of the information "…really gave us a different perspective…"

Arguably, the implication is that data integrity is an important attribute of information-informed decision making. This attribute of integrity originates from the "…many checks and balances…" put in place by the "data team." If one has confidence in the information, data ambiguity creates a need to dig deeper, to a more granular level (i.e., "…then you pulled data out…" and "…we were able to see it on a smaller scale…). Such action enables new perspectives born from critical reflection of prior sensibilities.

The narrative suggests that integrity authorizes decision makers to feel that "…what we were looking at was what we needed to look at…"

Similarly, Member 1 contributes the following regarding changes in perspective and the influence of empirical analysis on decision making:

> Going in, running data, looking at it, re-talking about it, getting the input from people, getting your input, and just those discussions - I think we evolved over time into where we are right now … [As a result] the process is more clearly defined as far as a district-wide evaluation process.

Seemingly, Member 1 sees the process of empirical analysis is cyclical, reoccurring, and ongoing. Here, the process involves "…running data…", engaged analysis (i.e., "…looking at it…"), reflection ("… re-talking about it… those discussions…"), and ensuring the information is representative and complete (i.e., "… input from people, getting your input…"). The influence is transformative as a decision making body saying "…we evolved over time into where we are right now…" The result is an evaluation system that is "…more clearly defined…" than it would have been in the absence of the empirical reflections. Member 1 goes on to reflect:

> There's been active involvement, an active committee, that's been working on looking at it [the information] and changing processes and ensuring that the process is well-communicated … Those things weren't occurring, really, at all before.

This narrative (above) suggests that empirical information helps facilitate "…active involvement…" encouraging decision makers to become "… active committee [members] …" These members respond to information by formulating new perspectives that engineer "…changing processes…" In this way, the empirical information influences changes in perspective leading to new policy decisions.

Post conference emphasis. Member 1 (below) also shared his/her perspective on the influence some early evaluator interview information had on identifying future training activities. The context of the narrative concerns the need to incorporate post-conference teacher mentoring skills as part of the general training program provided to evaluators. The empirical information was obtained from early evaluator (principal) interviews (obtained as part of this research study). Member 1 responds:

Interviewer:	Okay. A while ago, I shared with you, that some principals felt the training on the Danielson rubric was going well, and it was comprehensive, but one or two were uncomfortable with post-process mentoring. Has any of that come into your training and your thinking as a result of that comment?
Member 1:	Yes. In fact, through several of those comments, the post-conference, the coaching piece, that's the piece that I feel like we're missing, to a large degree … It is on the books. We've done some, but not nearly as much as we need to do … that is our next step that we need to take it to. I have done some minimal training on that, but that's really the direction we need to go. I'm so glad that you brought it up, because it's been on the back burner now, but it is the next thing that we need to do.
Interviewer:	That initial feedback helped start that thinking on that?
Member 1:	Absolutely.

In the narrative, Member 1 acknowledges the need to incorporate post-conference teacher mentoring skills as part of the training activity, saying "…that's the piece that I feel like we're missing…" The feedback acts to remind and reinforce this need for change (i.e., "…that's really the direction we need to go…") and in so doing heightens the priority of such change. Up to this point in time, the training focus had been exclusively on rubric understanding, application, and scoring consistency (reliability). The new interview information suggested a need to fundamentally change the focus and

priority of training. The implication is that this adjustment will occur at some point in the near future "…it is the next thing that we need to do…" (Reader's Note: The initial conversation regarding this evaluator feedback took place during the 2013-14 school year. The narrative presented above took place in December 2014. Thus, the comment "…I'm so glad that you brought it up, because it's been on the back burner now…" is a reference to the previous conversation and the reminder that this new direction is necessary.)

SPA initiative. The SPA (Support Plan of Action) initiative is targeted to teachers whose evaluation ratings are generally low, but not sufficient to warrant placement on a formal improvement plan. It was designed and implemented by the evaluation committee for use during SY2014-15 as an effort to provide teachers timely feedback and guidance in targeted areas of their practice. The goal was to help struggling teachers improve their instruction and avoid future consequential action. Member 1 discusses the origins of this initiative:

Interviewer:	Okay. We did a support plan separate from the "ineffective [classification]." Talk to me about how that may, or may not, have risen from our [data] conversations and our review of the information, analysis, and the need for that [Support Plan].
Member 1:	That SPA, the support plan, our SPA, it actually did generate from this committee, the committee that worked on it, because, as we discussed teacher performance, and we looked at the coaching model, and the requirements for plans of improvement … we know that sometimes teachers don't need that [formal improvement] plan, but they need some support … That came out of experience with teachers, the data that was collected here, the fact that we looked at the ratings, [and] the law change. The committee worked to say, "We know we need that [informal support plan] piece in place because it's part of a coaching model."… That is our purpose for teacher evaluation, to improve practice.

Interviewer: You think your thinking was influenced or came out of the work that we were doing, and the data that we were bringing?

Member 1: Well, the data that we're bringing to the committee especially. I think the difference for me is being able to say, in my heart I know this is the right thing, and I feel like this is something we should do. Data always is the, okay, this is the evidence that supports what you think [laughter].

The narrative (above) connects the support plan initiative with the empirical rating information initially reviewed by the committee members saying "…That came out of experience with teachers, the data that was collected here, the fact that we looked at the ratings,…" The sense is that the data both suggested the need for the SPA and authorized it as being "…the right thing… something we should do…" Here, the data interpretations, decisions, and evaluation purpose are all aligned (i.e., "…That is our purpose for teacher evaluation, to improve practice…").

Setting performance cut scores (Ineffective). A major decision focus of the evaluation committee concerned setting performance standards distinguishing levels of instructional quality. The reader will recall that Arizona legislation requires districts to assign one of four classifications: *Ineffective, Developing, Effective,* and *Highly Effective.* These performance standards are to be based on the empirical information generated by the evaluation process. To this end, much of the committee's data discussions concerned the evidences and integrity supporting these decisions. During the interview, Member 2 (below) reflects on this decision process and the influence the empirical data had on his/her thinking:

Interviewer: The standard setting process, the process of setting those cut marks, do you feel good about that process of looking at the data and coming to the conclusion of where we set that? Did the data help, or do you still feel that maybe that processes could be improved upon?

Member 2: My initial thoughts probably changed the most in that calibration part because initially, when you look at those numbers … and I remember at one point there were only four points difference between one category and the other. Then there was like 17 over here or 12 over here, and I'm thinking, "Oh my gosh, what is this? Did we not stretch it out? What did we do?" But then when you go and look at [the data], when you shared all of the analysis, there might have been 500 people or 300 people in those four points. That was one of the hardest things in explaining that to teachers is that when you fall between here and here, it's this rating, but between 62 and 65, and then 65 and 69, or whatever, those cutoff points are, then that bell curve really comes into play.

Member 2 is speaking about the difficulty of establishing the performance scales and related cut-scores that distinguish instructional performance classifications. The difficulty involves understanding distinctions between the scale of measure, the distributional characteristics of the underlying data (i.e., the range and variation of persons across the scale), and the decisions of where/how to differentiate meaningful performance criteria. The narrative reflects an evolution in this understanding (i.e., "…because initially, when you look at those numbers…") and the influence the data discussions had on decision making (i.e., "…when you go and look at [the data], when you shared all of the analysis…"). The discussion also reveals the complexity of the standard setting process and the difficulty in communicating the decision rational to external stakeholders: "… that was one of the hardest things in explaining that to teachers…" Member 2 goes on to reflect on the team conversations, saying:

> I remember, sometimes, I thought they were painstaking conversations or discussions, because we were really trying to decide "okay, should it be one point here, or one point here, and how many people would that impact," because we truly wanted it to be as representative as possible.

> I think in moving forward, and we did have also conversations about how this is normed - we have to keep it this way. We can't keep changing our ruler because then it's not an accurate measure.

Here, Member 2 positions the "…painstaking conversations…" in the context of making the best possible decision "…because we truly wanted it to be as representative as possible…" This exposes a recognition of the high stakes, consequential, setting in which decision making was taking place. The implication is that the committee members were fully aware of the task's importance and the responsibility they had to their colleagues (i.e., "…how many people would that [decision] impact …" In this way they struggled to understand and interpret the data: "…should it be one point here, or one point here…" Member 2 comments on the overall process, stating:

> The time investment was necessary because not everyone had that same level of understanding until all of those conversations were finished. We all had to understand and accept what the findings said and accept what cutoffs were made because those were heavy decisions.

The comment (above) speaks to the value of the data, the importance of the data discussions, and the influence the empirical analysis had on decision making. In this person's mind, "…the time investment was necessary…" in part because of complexity and the need for decision makers to have a full command of the information (i.e., "…that same level of understanding… we all had to understand…"). As used here, the phrase "…accept what cutoffs…") is not a sentiment of compliance. Rather, it is a statement of derived consensus, of agreement reached through critical reflection and collaborative decision making. Again, the weight of responsibility is revealed in the final phrase "…because those were heavy decisions…"

Similarly, Member 1 (below) discusses the influence the data discussions had specifically on establishing the *Ineffective* performance criteria:

> Interviewer: What about the "ineffective" rating; the process that we went through, our thinking process of how we arrived at finalizing what it means, and how we calculate or determine an ineffective rating? Did the data discussions and the information that we brought here shape your thinking about how we got to that final decision?
>
> Member 1: Yes. Yeah. Looking at the data that we - no one wants to be called ineffective [laughter], so the fact that we are forced to use those in the first place, I mean, you're looking at, talking about, something that's emotional, especially for people that are in the classroom teaching. A lot of people on our committee are actually recipients of those classifications.
>
> Interviewer: So by having that data, that helped you put this concept of "ineffective" into a context that led to our final decision? Would that be fair to say?
>
> Member 1: Absolutely.

Member 2's initial response connects the weight (i.e., importance, consequence, etc.) of the decision process to the review of the factual data saying "…Looking at the data …you're looking at, talking about, something that's emotional…" With regard to the role that data played in arriving at the Ineffective decision, Member 2 responds "Absolutely." Interestingly, there is recognition that many committee members were also classroom teachers. Thus, the decision process had personal relevance because some were "…actually recipients of those classifications…"

Correlation between achievement and Professional Practice scores. One important observation apparent in empirical information provided to the committee was the weak correlation existing between achievement scores and the professional practice ratings. This finding garnered considerable discussion and reflection among committee

members since it implicated the efficacy of the information to accurately discriminate instructional quality. Member 1 reflected on the impact this had on his/her perspective, initially saying:

> Interviewer: Okay. We struggled with putting together the value-added and the professional practice data. What did you think when that data revealed there wasn't really a very tight correlation between the two? What did that do to your thinking?
>
> Member 1: Well, it's one of those things where, I think, you hope that that's not gonna happen, but you think that it probably will [laughter] … Then you think, what are things that you can do to change that, to get it to the correlation that you think it should have, instead of the reality that it doesn't have?

For Member 1, the lack of correlation was counter to his/her personal perspective (i.e., "…you hope that that's not gonna happen…") while at the same time validating the fear that the two measures may not strongly associate ("…but you think that it probably will [not correlate]…). However, there is an inherent belief that "…it should have…" leading to the question "…what are things that you can do to change that…" In this way, the empirical data challenged the premise of coherence and initiated new thinking concerning cause and solution. Regarding causal factors, Member 1 hypothesizes:

> … you can't control student achievement data … What happens with each one of those individual students on a given day is - the teacher has no control over that, and so, what kinds of things can you do? When you see that [lack of correlation], and when you look at the data and it confirms, yes, how are you gonna defend this to teachers, because it raises the doubts. It raises questions, and so then what are you gonna do?

For this person, lack of teacher control over test scores is one reason the two measures did not align, saying "…you can't control student achievement data… the teacher has no control over that…" The reason posited for this lack of control is the influence external factors impose on test results. These influences may impact students

"…on a given day…" The implication is that test scores are an unreliable measure of student learning because they become biased by non-instructional factors. The problem then becomes one of decision integrity (validity): "…how are you gonna defend this?" For the committee members "…it raises the doubts. It raises questions …" of making accurate inferences of instructional competency. Member 1 adds:

> How do I trust this as a process when, clearly, the data shows that, in some cases, they don't match? That I have really high student achievement scores, but my professional practice scores are low?

Based on the narrative, prior assumptions are challenged, initiating the need for a solution. Member 1 then posits:

> So is that a problem in training for principals? Are we not collecting the correct evidence? Are we not analyzing teaching performance correctly, or, what's wrong with the student achievement component? (Member 1)

And then concludes:

> I think it's easier to identify that student achievement may be unstable, and we have less control over it. Because of that, collecting professional practice evidence, that helped drive decisions to really focus on inter-rater reliability, because one way to tighten it [the correlations] is to ensure we have [high] inter-rater reliability, and that we are effectively rating consistently all of our teachers against that rubric. The better we get at that, maybe there's a closer alignment, cuz it's the one thing we can, I think, better control. (Member 1)

So, for Member 1, the solution to the problem of poorly correlated measures is to improve the rating accuracy of evaluators, reinforcing the committee's decision to "… to really focus on inter-rater reliability…" Since "… we have less control over [test scores]…", this increased focus may yield "closer alignment" between the two measures. In this way, the empirical data fundamentally influenced the decision pathway of the committee.

In contrast, the apparent lack of association reinforced Member 2's concern over potential bias within the test score component. Unlike Member 1, this individual did not automatically assume that the two components would display a strong association. Reacting to the topic, Member 2 reflects:

> Interviewer: The one-to-one matching between Danielson scores and value added scores wasn't really as strong as we would like it to have been or we thought it would have been? Did that make you uncomfortable at all or did you feel that that analysis was okay?
>
> Member 2: I think initially I didn't know if I fully trusted how strong that correlation was [going to be]. I remember having those conversations of we want to put as little percent as possible connected to the student achievement because there were so many extenuating circumstances that we couldn't measure on the day of that test. When that child took the AIMS, I remember a conversation "do we know if their dog died that morning, or if Mom and Dad had an argument the night before, or they had breakfast that day."

Here, Member 2's perspective of association was not as well formed as Member 1, saying "… I didn't know if I fully trusted how strong that correlation was…" He/she acknowledges "…having those [committee] conversations…" about external influences and the sentiment "… to put as little percent as possible…" on the achievement indicator. Comparatively, the correlation data seemed to shift Member 1's thinking while solidifying the perspective held by Member 2. Member 2 goes on to share:

> There were so many different factors, and I think that was part of the struggle - is we are putting so much weight on this one day in time and how the students performed here?

Arguably, the narrative provided by Member 2 suggests that weak component associations forced discussion on "…so many different factors…" that potentially impact the integrity of the achievement measures. As a result, he/she suggests that the committee

was forced to ask "…are putting so [i.e. too] much weight…?" on achievement as a primary determinant of instructional competency? This "…was part of the struggle…" that decision makers grappled with concerning construction the final evaluation formula.

Devaluation of test scores. The narrative presented above regarding the *Correlation between Achievement and Professional Practice Scores* was partially embedded in a larger discussion regarding the committee's decision to devalue the weight assigned to the achievement component of the evaluation formula. The reader is reminded that the committee decided to assign a weight of 67% to the professional practice (Danielson) ratings and 33% (the minimum allowed under state legislation) to the achievement component. As discussed above, the weak correlations initiated reflective discussions within the committee that ultimately influenced this decision.

Member 1 shared his/her perspective of the decision to devalue the achievement measure, saying:

Interviewer:	Do you feel better that we devalued the value-added scores after this two-year process of looking at the data, and the relationships, and the lack of correlations - that we devalued it and we weighted the professional practice as much as we did?
Member 1:	I feel good about that, and actually, the longer we're in the process, I feel better and better about the fact that we did that. I would never want to move that to a higher level of the student achievement, to back to 50 percent or something, like we could do. I think that it's better for the teachers. I know it was a hard sell for our [Governing] Board, because the board really saw student achievement as something that they wanted at a higher level … but we did many presentations to the board as we went through this process, and because you were able to explain it. (Readers Note: this last statement was previously referenced under the *Value of Information* section discussed above)

Member 1 responds to the question of devaluing achievement saying "… I feel good about that…" and reflects that he/she feels increasingly comfortable with the decision over time (i.e., "…I feel better and better about the fact that we did that…"). In addition, the decision was "… better for the teachers…" and "…I would never want to move that to a higher level…" The implication is that integrity issues are enduring, preventing any future change in perspective. From the collective narrative provided by Member 1, the pathway for arriving at this perspective began with the initial empirical information, leading to reflection and discussion, resulting in a re-evaluation of personal beliefs, and ending with the decision to devalue the component.

Interestingly, Member 1 reflects that the decision to devalue test scores "… was a hard sell for our [Governing] board…" However, just as with the committee experience, access to, and reflection on, the empirical data acted to influence a change in perspective. Here, the Board "…really saw student achievement as something that they wanted at a higher level …" However, presentation of the empirical information precipitated understanding and a change in thinking (i.e., "… we did many presentations [and] you were able to explain it…"). In this way, the influence of the empirical information and related data discussions represented a powerful catalyst for changing thinking and reaching decisions both within the committee and at the highest policy levels.

Similarly, Member 2 (below) echoed confidence in the committee's decision to devalue the achievement component of the evaluation formula. (Reader's Note: Some of the dialog presented below was also referenced in the earlier section regarding the *Value of Information*).

Interviewer: Did the analysis of the data help you resolve or help you think through the decision to devalue the weight of the test score component? Was that a decision based upon looking at the data?

Member 2: I do think that all of that analysis helped us look at not just individually but then also on a bigger scale … Yes, having it all, [the] data on a piece of paper …

Interviewer: Made it clearer, made it more succinct?

Member 2: By separating it all out … those clear explanations … and by separating it out into all of the varying components …

Interviewer: Right, … so the availability to look at the empirical data - did that help you come to that conclusion that we've weighted it properly because it's based on evidence rather than just how are we going to do it?

Member 2: Yes, in our conversations and through your data analysis, we were able to, I think, at each meeting, become more confident that this was not just an arbitrary number. This is evidence based.

In this series of dialog, Member 2 places confidence in the decision to devalue the achievement component primarily based on having access to the information (i.e., "…[the] data on a piece of paper…"), engaging in extensive discussion (i.e., "…that analysis helped us…"), and developing understanding through "…those clear explanations…" The empirical information helped members "…become more confident…" in their decision because the decision process was "…evidence based…"

Member 2 also shared comments regarding adoption of the value-added approach to measuring instructional effectiveness:

Interviewer: Did the decision to go with the particular type of value added model with these [student background] adjustments in it, did you feel at the time that that was an appropriate approach to the issue of using test scores for evaluation?

Member 2: I did, and that was something the committee looked at very carefully and thoughtfully because we wanted to make sure that we had lots of assessment options, but we wanted to make sure that we only included those that were valid and true to the measure, because I know at one time we considered [other models] in there, and we considered putting other assessments in, and we just kept going back to what is the most accurate measure of student growth.

Interviewer: So, you felt good about that decision?

Member 2: Absolutely. I wish more people had had the opportunity to participate in all of our in-depth conversations, but I think that the members were able to communicate all of the thought, and just the serious process, that was followed in our overall evaluation process.

Here, Member 2 reflects that "…the committee looked … very carefully and thoughtfully…" at the issue of adopting a particular approach to measuring the achievement aspect of instructional effectiveness. The importance of the task is reflected in his/her comment, stating "… we wanted to make sure…" that the method was "…valid and true to the measure…" Member 2 notes that the committee reviewed "…other assessments…", but after review and analysis they "…kept going back…" to modeling approach that was eventually adopted. In this way, the information influenced thinking, shaped perspective, and lead to final decisions. Member 2 reflects in the importance of this process by saying "…I wish more people had had the opportunity to participate in all of our in-depth conversations…" because these data-informed conversations reflected a "…serious process…" for decision making.

Summary of RQ2 (b). RQ2 (b) assesses the influence empirical information had on the decision making process. It is evidenced by the reflections from two key evaluation committee members who were uniquely positioned within the decision process

for its entire (approximately) three-plus year period. From the narratives provided by these decision makers, the act of providing ongoing empirical information effected the decision environment in the following areas:

- *Changes to the Training Focus for Evaluators*: A shift of emphasis away from evaluation structure and process onto (Danielson) rubric understanding, application, and empirical foundation.
- *Increased Attention on Ensuring Evaluator Interrater Reliability*: An increased focus in hands-on practical application of rubric criteria to actual instructional settings for the purpose of verification and improvement of scoring precision.
- *Conduct of a Formal Interrater Reliability Verification Study*: Design and implementation of an interrater reliability study to assess the degree of rater consistency and precision in the evaluation process.
- *Facilitating Changes in Perspective*: Empirical information facilitated changes in a priori perspectives held by decision makers by providing the basis for critical reflection, communication, and discussion.
- *Facilitating an Increase in Clarity and Confidence*: Empirical information provided the authority for committee members to have confidence in decisions impacting evaluation structure and process.
- *Articulating the Need for Future Emphasis on Post-Conference Teacher Mentoring*: Empirical information revealed a need to incorporate post-conference mentoring and support activities into future evaluator training.

- *Implementation of the SPA (Support Plan) Initiative*: Review of empirical information assisted in articulating the need to add a support plan initiative to the evaluation model that focuses on struggling teacher.
- *Guiding the Standard Setting Decision Process*: Review of empirical information was critical to determining performance standards and related instructional classifications.
- *Devaluation of Test Scores in the Evaluation Formula*: Critical reflection of disconfirming empirical information raised questions, dialog, and critical reflection on the suitability of test scores to serve as a heavily weighted component in the evaluation framework. As a result, decision makers devalued to importance of achievement indicators to the minimum allowed under state legislation. The empirical information provided foundational support for this decision.

Overall Summary of RQ2

RQ2 was formulated as follows: *How does providing an evolving dialog on validity evidence influence organizational policy decisions regarding implementation of the LEA's teacher evaluation system?* It is informed by two supporting questions: RQ2 (a): *To what extent do policy-level stakeholders value the collection and review of validity evidences as an important input to the system's ongoing development?* And RQ2 (b): *To what extent does validation evidence prompt changes in organizational decisions regarding system implementation?* These supporting questions differentiate the topic in terms of *Value* (RQ2 (a)) and *Influence* (RQ2 (b)) of information in the decision making process.

Based on the narratives, the *Value of Information* originates from its ability to inform and validate decision making. In this way, attributes of access, analysis, reflection, and communication (transparency) collectively engender trust and confidence in the decision process. Importantly, evaluation is seen a consequential activity, impacting teacher's professional identity. Here, the *Value of Information* is heightened due to the high stakes nature of the evaluation process. In this regard, empirical information serves as the foundation from which to question and consider intended and unintended outcomes.

An attribute of the *Value of Information* is the information broker. Information brokers serve as a catalyst for developing and presenting empirical information. In the context of decision making, this position brings a unique set of skills, knowledge, and expertise not held by other members of the evaluation committee or other policy-level stakeholders. While a foundational role of the information Broker is the delivery of information, this activity is seen as insufficient for adding substantive value. To do so, the information broker must facilitate greater levels of understanding, clarity, and interpretation by communicating complex information in a way that is accessible to decision makers. Importantly, this does not imply simplification, but rather deconstruction for the purpose of approachability. The result of this service is enhanced trust and confidence in the decision making process.

Regarding the *Influence of Information*, stakeholder narratives outline numerous areas of impact including changes in training focus, changes in personal perspectives, the addition of new components to the evaluation model, and critical decisions regarding standard setting and establishing component importance for the purpose of quantifying

instructional competence. In this way, provision of empirical information was a central and necessary activity to the design and implementation of the evaluation activity.

Petite Assertions: RQ2

1. The provision of empirical information had a profound and lasting effect on the design and implementation of the evaluation system: challenging a priori perspectives, facilitating critical reflection and dialog, and providing authority and foundation to decisions defining the organization's evaluation environment.
2. Information brokers serve an important role in the decision making process, bringing specialized skills, knowledge, and expertise into the policy environment. Information brokers facilitate understanding of complex information; act as important arbiters of a priori perspectives held by stakeholders; serve as a catalyst for critical reflection; and help enable trust, confidence, and authority in the decision process.
3. The provision of accurate, complete, unbiased information into the decision making process is a critical condition for establishing authority and confidence in the final design and application of the evaluation activity. The focus and type of empirical information must align with the purpose of the evaluation process and offer account on any a priori perspectives held by decision makers.
4. Communication, dialog, and critical reflection is a necessary activity in the decision making process. Empirical information is the foundation of this activity.

Primary Research Question 3

RQ3 (a)

To what degree have the perspectives and concerns of stakeholders been incorporated into, and influence, the system's implementation plan. The approach is semi-structured interviews. The measures are coded interview responses.

Introduction

Implementation of high-stakes evaluation systems impact teachers across the organization. How policy makers design and implement these systems determines how well the innovation is accepted and ultimately its ability to affect sustainable change in professional practice **(**Hargreaves & Fink, 2004, 2006; Kotter, 1996; Rogers, 2003). Theorists posit that successful models of sustained innovation directly incorporate stakeholders as agents of change (Kotter, 1996, 2010, 2011; Hord, Rutherford, Huling-Austin, & Hall, 1987; Hall & Hord, 2011; Marzano, Waters, & McNulty, 2005). In this way, perceptions of inclusion, empowerment, and commitment are intimately linked to assessments of innovation success.

Stakeholders who view innovation as external, imposed, and/or irrelevant fail to associate with the effort's goals and objectives: they lack a shared vision between the innovation and their own personal beliefs/interests (Stumbo & McWalters, 2010; Hargreaves & Fink, 2004; Marzano et al., 2005). The consequential effects may be to limit the efficacy of the innovation and possibly create unintended outcomes. In this way, validity studies benefit from inclusion of stakeholder voice to examine/understand aspects of consequential outcomes, both intended and unintended (Messick, 1998, 1989a; Cronbach, 1971).

For this study, the research question discussed herein examines stakeholder voice in terms of perceptions of inclusion and empowerment in the design, implementation, and application of the district's evaluation system. The discussions were initiated by the following generalized interview protocol: *How much input do you feel teachers (yourself) have had in the design, implementation, and use of the evaluation system*? Variations of the prompt were accommodated to fit each of the four stakeholder groups (Teachers, n = 7; Principals, n = 8; District Policy, n = 4; and State Policy, n = 3).

Construct Overview

Voice (generic use of the term): As used in this study, "voice" is equated with "influence." In this way, voice is conceptualized as the degree to which stakeholder perspectives influence decision making with regard to the design, implementation, and utilization of the evaluation system. In addition, voice implies ownership of consequential decisions.

Stakeholder voice may be manifest through two means: feedback/input provided to external decision makers or by direct responsibility. In the former, stakeholders provide input, suggestion, and/or feedback that is then considered and acted upon by external decision makers. To the extent decision makers use stakeholder feedback to form decisions, stakeholder perspectives are seen as influencing the system (i.e., they have a voice in the system). For the latter, stakeholders are directly empowered with the responsibility to make design decisions (i.e., teachers as members of the decision committee, decision by vote, etc.).

Initial analysis of the narratives differentiates stakeholder perspectives by two dominant sentiments: *Lack of Voice (Negative)* and *Voice as Opportunity (Positive).*

Inclusive of these is a distinction between teachers afforded opportunities to provide feedback and reaction to the evaluation framework/process versus inclusion into the decision making process. Arguably, the two perspectives are very different: one is reactionary (provide feedback) the other is empowering (decision making). Finally, an underlying conflict of *Feedback vs. Decision Making* is embedded throughout the narrative. Each of the construct conceptualizations are summarized below:

Lack of Voice (Negative):

> Within the "Lack of Voice (Negative)" component, stakeholder influence in decision making is viewed as limited or absent. Here, opportunity to provide feedback, reaction, and/or input is viewed as inconsequential – having no real impact on system design. Information flow is one-way, originating from external decision makers. The intent of communication is to inform, train, and/or convey understanding for the sake of compliance, and not to substantially influence design decisions.
>
> The "Lack of Voice (Negative)" construct positions evaluation as externally formulated/constructed, imposed from the top-down. As such, there is a political/policy power-center aspect where teachers are recipients, not designers, of the system. Because evaluation is externally imposed, teachers are afforded little opportunity to change/modify its structure. Thus, expressions of concern are seen as problematic.

Codes and Identities delineating the *Lack of Voice* component are provided below:

No Voice

- Negative, zero, none; no voice, limited voice, never been given voice to say "that's not gonna fly"; "just work here, that's how it is"; no opportunity; no input, , "haven't heard of any teachers involved"; not much; little or no input (design); no teacher voice in development & implementation (just info and opportunity to learn); teachers are recipients not developers;

Top Down/Imposed

- top down; mandated; rolled out; state requirements; legislative mandate; state framework; state driven design; came from Florida; TEval would be very different if influenced by teachers/educators;

Top Down/Political

- Political, Race to the Top (RTT);(SB1040); not represented by educators; educators wouldn't design this type of evaluation system; educators had no formal input to legislation in design or components; outside political actors; business interests – believed educators required an better evaluation framework/accountability;

Structured Process

- FFT is a structured process, can't change it or it wouldn't be the FFT; FFT is a framework - fixed; Can't change framework;

TEval Committee

- teachers on TEval Committee could only work within state framework, not change it to fit local perspectives or needs; TEval committee had limited power to change framework;

Not Taking Advantage of

- Teachers had numerous opportunities to give input but little input was actually given (videos, meetings, discussions, staff meetings); District provided opportunity for input; communication, discussion, meetings, presentations, info sharing, and surveys but for some teachers that won't be seen as enough; Teachers will respond with 'very limited', 'no control', 'no real impact';

Figure 56. Codes and identities related to *Lack of Voice.*

Voice as Opportunity (Positive):

The "Voice as Opportunity" component is characterized as the opportunity for teachers to provide input, reaction, and feedback on the evaluation structure and process. This is an organizational construct in which teachers are afforded (granted) avenues of expression regardless of whether or not they choose to share/contribute. Here, the act of informing is equated with influence in a positive way. That is, external decision makers take stakeholder input into account when making decisions and/or formulating policy. Thus, the perspective is based on trust in organizational decision making. In addition, Voice as Opportunity may also reflect a representational perspective of influence: where a small, select, group of teachers are entrusted to make decisions for the collective good (i.e., members of a steering committee). It is assumed that the empowered few echo the perspectives of the larger group.

Codes and Identities delineating the *Input/Positive* component are provided below:

Opportunity

- Opportunity to express reactions; organization held principal meetings, principals shared info with staff; teachers were invited to attend informational meetings – gave teacher opportunity to share concerns); "Yes, teachers had opportunity for input, many opportunities";

Informed

- Well informed; Teacher campus representatives came to informational/feedback meetings – these teachers felt empowered (being invited to district meetings); (Note: not the same as being engaged in the decision making process); principals kept teachers informed; lots of discussion, information, but little input; lots of time spent on info;

Trust

- Trust in administration to implement tool/system we use; FFT is research-based & therefore must have had had teacher-input in its original design; Principals view that teachers are generally positive about TEval process; FFT good, good training, no guessing game, it is what it is; sufficient input – teachers had "big: role in development;

Policy

- Distinction between legislation and SBE process: no input to legislation but lots of educator's input to State Board of Education & Governor's Evaluation Task Force;

TEval Committee

- Has been careful, scrutinized, revised, cycles of review, positive revisions; TEval Committee had representation; TEval Committee composed of well-respected members;

Figure 57. Codes and identities related to *Input/Positive*.

Feedback vs. Decision Making:

Feedback vs. Decision Making distinctions are embedded throughout both the "Lack of Voice" and "Voice as Opportunity" components. It distinguishes expression (i.e., feedback, input, perspective) from influence (i.e., active change in decisions and/or policy). It posits that input/feedback is not equated with empowerment. By distinguishing feedback from decision making, the sentiments of "provision of information" and "opportunity to comment" are devalued. It presumes that empowerment comes from influence and/or an active role in decision making.

Codes and Identities delineating the *Feedback vs. Decision Making* component are provided below:

- Expression and influence are different; "voice" not same as "influence"; discussion and information not same as decision making; stakeholder being informed not the same as stakeholder makes decisions; training is good but not same as having input;
- Opportunity to inquire and learn but not change, design, or shape; opportunity to discuss but not change; Input not same as deep reflection, discussion, understanding; Training, communication, information not equated to "voice" (design, impact, change decisions) ;

Figure 58. Codes and identities related to *Feedback vs. Decision Making.*

Stakeholder Narratives

Teachers (Voice). As a group, teachers reflected mostly negative sentiments regarding their influence on the design and implementation of the evaluation system. One teacher (below) reacting to the question of voice: responds:

Interviewer: Do you feel that you or your colleagues, as teachers in the district over the last year, have had a voice or a say into how the evaluation system has been designed and implemented?

Teacher 102: No. We've been given all the information but we've never been given that voice to say, "That's not gonna fly."

Interviewer: Your sense would be that you work here and this is the evaluation?

Teacher 102: I work here. This is the evaluation and this is how it is.

This individual makes a distinction between being afforded "… all the information…" and being able to influence the system (i.e., "…never been given that voice…"). Arguably, there is frustration evident in the comment and a lack of personal empowerment (i.e., "…I work here… this is how it is…"). A different teacher (below) echoes this same sentiment saying:

Interviewer: Do you think you and your colleagues have had a voice, some influence, in how the district's implemented and designed its evaluation system?

Teacher 103: No, I don't. No, I think we just kind of showed up in August and they went, "This is how it's gonna be." We went, "Okay, sounds good."

Interviewer: Do you have an opinion on whether the district has provided opportunity for your colleagues to give input?

Teacher 103: I don't know anybody who has.

Here, the person views the evaluation system as being implemented top down by policy actors imposing the system on teachers: "…we just kind of showed up …and they went 'this is how it's gonna be'…" For this individual, teachers are recipients of policy, not designers. Evaluation is being done to them, not by them. In addition, the person is unaware of other colleagues being afforded opportunities to provide input saying "…I don't know anybody who has…" Still a different teacher reacts:

Interviewer: Do you think that you and your colleagues have had a voice or an influence in how the teacher evaluation system has been developed?

Teacher 104: No, no. Yeah, and in our last staff meeting, the whole staff was there and my administrator, the principal, is talking about … some cadre, 'cause it was on a form that the data person sent out … I read the form that the data person sent out and it mentioned a cadre. What, who's, that cadre? And my administrator goes, "Well, this person from district, that person from district, this person from the school, this person from the, whatever, state. This person from whatever, whatever."

I said, "No teachers? There's no teachers on that?" I could see that it kind of "No, there's no teachers on that." "Oh, just wondering." But yet, then I looked at everyone else around me, they were all thinking it. Everybody wanted to know, but nobody had the guts to say anything. It's terrifying.

This individual acknowledges that the district had provided some information on the evaluation system (i.e., "…our last staff meeting … the whole staff was there…"), but reacted to the lack of teacher representation in decision making process (i.e., "…there's no teachers on that [cadre]..."). For him/her, the cadre of decision makers is external, primarily composed of district/state policy makers. In addition, the perception is that his/her colleagues had the same reaction, but were reluctant to express their concern saying "…nobody had the guts to say anything…" and concluding "… it's terrifying…"

Interestingly, this person was unaware (uniformed) of the Evaluation Committee member composition, composed of a mix of elementary, high school, and special area teaches, school administrators, and instruction coaches. Thus, the source of the frustration, lack of trust, lack of empowerment, seems to originate from a basic lack of factual information. It is arguable whether he/she would have reacted similarly if equipped this understanding. The same teacher (104) goes on to share:

Interviewer: So the view that there was no teacher input?

Teacher 104: On this cadre that was to represent us.

Interviewer: No teachers there?

Teacher 104: No teachers, and that was just one example of all these things that are happening. Then all I get is told what to do. Yet do I have any power at all? Do I have any say in anything? Hence, well, maybe I shouldn't be a teacher 'cause I'm pretty, I'm not worthy. I can't teach. You got to have a tutor in here. You got to pull my kids out for this. You got to pull my kids out for that. What am I doing here? I should not even be a teacher. My self-esteem is going down lower and lower and lower.

For this person (above), the lack of input/representation is simply an extension of "…all these things…" that are being imposed on teachers. There is a clear perception of imposition (i.e., "…all I get is told what to do…"), a fundamental lack of empowerment (i.e., "…do I have any power?…Do I have any say in anything?"), harm to personal value and self-esteem (i.e., "… I'm not worthy … my self-esteem is going down lower and lower …"), and questioning of professional competency (i.e., "…I can't teach… What am I doing here…"). For this individual, the context of voice in evaluation is just one more troubling artifact of not feeling valued or empowered in his/her professional identity.

A different teacher (below) shares the sentiment that teachers lack representation in the evaluation system decision making process, saying:

Interviewer: Okay. That's fine. My last question is about 'voice'. How much input or "voice', if you will, do you think you and your colleagues have had in the implementation of the evaluation system to date?

Teacher 108: I'm not sure if I'm just not aware that there was an ability to have input, but I would say next [laughter] to none. But again, I don't know if that's just because I didn't see it, or I don't know if it was offered and I missed it. But yeah, I will have to say that the district is doing this, and having the random interviews, so then there is input on that side, too.

This individual responds by saying "…next to none…" but acknowledges that his/she may not have all the information. In addition, there is appreciation that the district

is making the effort to interview teachers regarding their perspectives and perceptions (Readers Note: All persons interviewed for this study were made aware that the information might be used to inform both the immediate research activity and as input to the organization's evaluation efforts).

Another teacher (below) continues this impression that teachers might not be represented in the decision making process and agrees that the evaluation effort is a top down implementation:

> Interviewer: Okay. How much voice do you think teachers have had in the district in helping shape the evaluation system, if at all, or do you feel that the evaluation system has, pretty much, been a top-down implementation?
>
> Teacher 106: In my opinion, I would probably say, "Top-down." I just haven't heard of anybody being involved in committees or anything like that. It just was kind of rolled out to us in staff meetings that, "Here's the new evaluation process based on Arizona's requirements."

As with his/her colleagues, this individual is not aware of any teacher involvement (i.e., "…I just haven't heard of anybody…") and that the system was "…kind of rolled out to us in staff meetings…" Recognition of a top-down approach was also qualified as originating from the state-level (i.e., "…based on Arizona's requirements…"). A colleague (below) acknowledges this state-level influence saying, "I think it's top down. I really do. I think it's top down. I think, like you said, it's legislature. This is something that's mandated that we have to do" (Teacher 105).

For each of these teachers (106 and 105), the evaluation system is externally imposed, implemented by the district according to state-level mandates and requirements. The implication is (again) that teachers had little input and/or control over the system. In

addition, acknowledging the role of state-level actors implies political considerations in its design and utilization, again, outside the control/influence of classroom teachers.

Finally, in contrast to his/her colleagues, one of the seven individuals interviewed expressed confidence that local (district) teachers have been included in the evaluation's decision making process, stating:

> I know for sure there's been at least two requests for teachers to go to meetings at the district to talk about it. I'm sure it's been discussed with state officials, with admin. I know here at my school, we've talked about it, too, together as a staff to our principal and then our principal would go to the principal's meeting and share that. I believe, based on what I know, that teachers have had a definite opportunity to share their concerns and to voice whatever they needed to voice. I definitely think teachers have been included in the process. (Teacher 101)

This individual expresses trust in his/hers principal's willingness to convey teacher feedback to district level decision makers. Interestingly, there is also a reference to sharing the feedback with state-level decision makers (i.e., "…I'm sure it's been discussed with state officials…"). Because of this pathway of communication (staff to principal to district to state), the person concludes that "…teachers have had a definite opportunity to share their …voice…" In this way, teachers "… have been included in the process…" Arguably, this person is equating input and inclusion in the generalized process with material influence/decision making. That is, teachers do not necessarily have to be decision makers to effect design of the system.

Teacher (Summary). Teachers reflect negative views of their inclusion in the evaluation decision making process. The dominant perspective is that evaluation is externally imposed, lacking substantive influence from classroom teachers. There is some acknowledgment of opportunities to provide feedback/input. Nevertheless, the sentiments reflect a general lack of empowerment in the decision making process.

Principal (Voice). Reacting to the initial question of voice in developing the evaluation system, one principal responds:

> Interviewer: Okay. Okay. We've gone through a rather long development process in our district to implement the process. A natural question is: how much voice do you think you've had in the design of this new evaluation environment? Then I'd like to ask you that same question in terms of your thinking of how your teachers would respond to that.
>
> Principal 202: I felt like we [principals] had some input for the new version that came in. It also, on the flip side, it's kinda "it is what it is". When we're using the Danielson rubric we don't go and change it all up, because it is, this is her rubric, it's been through research to be effective as a tool for districts' use.

The principal acknowledges that they have had "some input" into the new version of the evaluation system, but for the most part "…it is what it is…" There is a rationale for minimizing concern for any lack of influence stating "…this is her rubric, it's been through research…" The comment places trust in the belief that the Danielson model is the product of prior research and development and changing it devalues its legitimacy: "… we don't go and change it all up…" The individual goes on, saying:

> It's a little more training for new administrators, really, but as far as input, the structure itself doesn't allow a whole bunch. [If] we said, "Well, we wanna change all of this", well, then we wouldn't be using Danielson anymore. We'd be using something that we made up, and so I don't think it's for me [to change]. It's almost, like, neither here nor there. It's like, I trust the administration to say this is the tool that we wanna use. (Principal 202)

For this principal, the sentiment is one of trust and acceptance. There is comfort in adopting a research-based evaluation system. The presumption is that it (Danielson) was rigorously developed by experts and adequately assesses instructional quality. As such, there isn't concern over relinquishing decision-making influence/authority because "…then we wouldn't be using Danielson anymore…" That is, altering it harms its

inherent legitimacy to "…using something that we [just] made up…" Regarding opportunity for input, the same principal comments:

> I felt like I'd been able to address, if I do have a question about it, or how might we use this, or what does this mean. They've [questions] been answered. We've been given an opportunity to go through with the new system, and look at some of those videos and interrater reliability of how we're all rating. I felt like we've been given that. (Principal 202)

From the comment, this principal feels the district has provided opportunity to ask questions stating "…They've [questions] been answered…" Importantly, the opportunity to clarify questions is seen as helping in the execution of the system as opposed to changing the system itself (i.e., structure, content, design). Regarding teachers' perceptions, the individual continues:

Interviewer:	Okay, alright. How do you think your teachers would respond? Do you think that the teacher evaluation system has reflected the concerns or the needs or the inputs of teachers?
Principal 202:	The teachers probably, I would think, would say they have little input at all, like, "No, I was just told this is how it's gonna be." Again I don't know if there was teachers outside that were on committees I may have forgotten about, but pretty much I think it was just like, "Hey, here's what we're gonna do." We've been directed by the legislature that we've gotta come up with something. We've always used Danielson. We've incorporated that in our AIMS scores and the growth [measures] … Again, I think there's been good discussion, and there's been good answering questions and concerns.

The principal believes that teachers will respond that they have had no meaningful input: "…I [a teacher] was just told this is how it's gonna be…" The power center for the directive is the external policy/politics (i.e., "…directed by the legislature…") There is a sense of conformance in the phrase "…we've gotta come up with something…" because

it is used in the context of meeting legislative directive. There is no sense of influencing the decision making process.

A different principal (below) bases his/her perspective of voice on the fact that the Danielson model is grounded in established research:

> Interviewer: Okay, all right. How well do you think the implementation of our system has taken account, or tried to include, the voice of the teachers and of the administrators? Do you think your teachers and your colleagues feel that they've had some voice, or some impact, or some contribution to it?
>
> Principal 203: Well, I think we [principals] know that the rubric piece of it, we weren't really going, I mean, it's the Danielson model. It's based on research. I've not really heard a whole lot, but I think the appreciation comes in that they [principals] were trained. They have the information. Like I said, there is no guessing game, so they're well informed about the process, I think, in that regard.

This person discusses voice in terms of information, training, and understanding of the evaluation system: "…they [principals] were trained… have the information… well informed…" Apparently, structure brings clarity (i.e., "…no guessing game…"), clarity brings confidence (i.e., "…it's based on research…") that the framework does not require substantive change and/or improvement. As a result, intrusion into the decision making process is not necessary because the system is already well developed. The only needed activity is training, knowledge, and understanding of what/how to properly apply the system. The same individual then discusses his/her belief on teacher's perceptions:

> Even most recently, us sending one of our teachers to an information session, some other questions came about. She [the teacher] felt empowered to be able to ask those, to be able to represent the teachers on our campus. When that information is shared, again, it's just going to validate once again that voice was considered. (Principal 203)

Again, utility is derived not from participation in the decision making process, but by providing information and clarification. In the principal's view, teachers "…felt empowered…" by attending "… information session(s)…" and the "… [ability] to ask…" questions. In this way, voice (i.e., empowerment) is equated with opportunity to be informed, ask questions, and share concerns. To clarify this perspective, the dialog continued as:

> Interviewer: Do you actually think that your teachers feel that the district has tried to provide them with a lot of information, transparency, at least explanation?
>
> Principal 203: I believe so, and then the other piece we're starting to hear more and more is teachers… telling teachers, "Well, haven't you looked on line? Haven't you looked at the rubric? This is how we can get to this piece." If something comes up during PLCs, they do make reference to the Charlotte Danielson rubric … I feel in that regard, like I said, it's not something new. We're getting to the point where it is going to become our common vocabulary and something that is just a part of our practice, which I like.

For this principal, the evidence of voice, in part, becomes teachers' buy-in of the system as evidenced by "…teachers… telling teachers…" and the use of a "…common vocabulary…" Arguably, these observations permit the principal to posit that teachers have been afforded input into the system (i.e., because teachers are accepting the framework and actively promoting it among colleagues).

The next principal (below) had been a member of the district's Teacher Evaluation Committee. The dialog regarding teacher voice transpired as follows:

> Interviewer: Okay. My last question is really about the implementation and the design of the teacher evaluation system. How well do you feel that the district has included, and listened to, input from teachers and stakeholders in how we've developed it and implemented it? Have we done a good job at that? Do we need to do a better job at it?

Principal 204: Remember my comment when I said that educators are fantastic about just doing what they're told, and they don't take the initiative or effort to do anything about it? This would be a good example of "we have given teachers various ways in which to provide input, to be part of committees or conversations about the teacher evaluation system" … [but] very little feedback has been given. I'd like to say it's because the principals and district is doing a fantastic job of explaining the process and walking them through, whether it's with the amazing [evaluation] videos or with the meetings we've had, but district has done a very adequate job of providing opportunities for it.

I think a few teachers have provided input. I think the committee that's in place has looked at it very carefully, and have scrutinized what's in place, and has made revisions. We've been through several cycles of revisions, some major and some minor, but I think the revisions that have been initiated are positive.

Here, he/she provides qualified support for the idea that "… a few teachers have provided input …" which has informed the decision making process (i.e., decision makers have "…looked at it very carefully…"). The caveat is that while the district "…has done a very adequate job of providing opportunities …" for involvement, "…very little feedback has been given…" However, from the input provided, committee members "… [have] made revisions…" to the evaluation system. In this regard, this principal deviates from his/her colleagues by suggesting teacher voice has influenced decision making. The concern is that not enough teachers took advantage of the opportunity to become more involved (by providing feedback/input).

Importantly, this perspective is being provided by an "insider" to the decision process (i.e., a member of the evaluation committee). Thus, it becomes a different question regarding whether teachers believe they had any impact. This principal clarifies his/her thinking on this by responding:

Interviewer: If I were to ask your teachers, do they feel that they've had a voice in the development and implementation of the teacher evaluation system, how do you think they would answer that question? It's one thing to accept it; it's another thing to ask them "do you feel that you've had a voice in it".

Principal 204: Honestly, I would have to say no, but if you change one word in that question—if you had an "opportunity" to have a voice, they would all say yes.

From the response, this principal (above) posits teachers hold negative perceptions concerning involvement and empowerment. This is in spite of the view that the evaluation committee actually considered, and acted upon, the minimal input provided by some teachers. Thus, it is reasoned that teachers will feel they have had little influence on the system, but that the district has given "opportunity" to provide input. Another principal (below) agrees with this sentiment:

Interviewer: Okay. I have a question about voice, the voice that teachers might have been given in the development of the teacher evaluation system. How well have we done that? Or has it been just, sort of, "this is the way we're gonna do it"?

Principal 206: My understanding is that you guys [district] have provided a lot of opportunities for the teachers to [provide their] voice.

Interviewer: Your understanding?

Principal 206: Yes. That is my understanding. [Laughter]. My understanding is that, 'cause, I do know there's been opportunities for teachers to come and do that kinda stuff. It's just that they [teachers] don't [provide feedback].

Here again, this principal supports previous perspectives that teachers have been given "… lot[s] of opportunities…" to provide input, but have generally not taken advantage of it (i.e., "…It's just that they don't…"). There seems to be an assumption that if they did, decision makers would use that information to make changes/improvements to the evaluation system. The position also seems to argue that

teacher apathy is their own construction; they have the opportunity to actively influence decision making but they choose to ignore it. As mentioned earlier, this perspective equates providing input with being an active member of the decision making process. Arguably, this is similar to voters being considered members of a legislature: if you don't vote, you can't influence law making.

A fifth principal (below) supports that teachers have been given information but have not been involved (given voice*)* in the decision making process, stating:

Interviewer:	Alright. How well do you think we've done at incorporating feedback and teachers' voices, your voices, in the design and the implementation of the evaluation system? Or has it been just, "look, this is the policy. This is what we have to do. Here's your training"?
Principal 207:	I think that they [teachers] would say, "We've done training in a big lump. Then, we've done it in parts and pieces, and we still do that … I would say that teachers are not involved in the creation of that, but are involved in knowing what it's about.
Interviewer:	The process?
Principal 207:	Knowing how to do well, knowing what isn't good …
Interviewer:	So they [teachers] may not feel that they were instrumental in shaping the policy and shaping the system?
Principal 207:	Correct.
Interviewer:	But they may feel that they've received lots of discussion and lots of outreach and information and training and experience?
Principal 207:	Yeah. I think so. Yeah.

From the narrative, this principal believes teachers do not feel empowered in the decision making process. The distinction is made between being informed (i.e., "…We've done training…"), which results in teachers "…knowing how … knowing what…", but not being influential (i.e., "…the creation of [the system]…").

A sixth principal (below) sees the evaluation system from a top-down policy perspective which restricts the ability for teachers and administrators to have much influence in the decision making process:

Interviewer: Do you think that we've done a good job at trying to obtain feedback and input and voice from teachers and principals in the process? The contrast would be the district office saying "Here's the evaluation system, here's what you're going to do. This is the way it is."

Principal 208: Right. There is definitely that part where this is coming down from the state. This is what we're doing. We gotta do it. I do think there is those little committees that look at, bring teachers in, and go over that kind of, and bringing them back and having them understand the process better or understanding those components better. … [But] When it comes right down, I think, this is the way it is. Yeah, we didn't really have much…

Interviewer: Much input?

Principal 208: No. Right.

For this individual there is recognition that "… those little committees…" are present, but they serve more of an informing function rather than actually empowering decision making. In this way, the committees "…bring teachers in…" in order to "…hav[e] them understand…" the process and components of the evaluation system. But the bottom line is that "…we [teachers and administrators] didn't really have much [input]…" The same individual continues:

Interviewer: If I ask your teachers "has the implementation of this system listened to, and accounted for, the voice of teachers or has it been pretty much the district" what'd you think they would say?

Principal 208: I would think they [teachers] would agree [it's been all district]. That it was, just because I remember, yeah, they just kind of, "This is what we're doing." Yeah, and then "This is the training for it." … If there was a little side committee

> looking at the state, what the state was bringing down, I think it was very limited probably.

The principal devalues the ability of the "…little side committee…" to provide teachers an influencing voice stating "…This is what we're doing… This is the training…" This is because the evaluation directive emanates from state policy (i.e., "…the state was bringing down…"). The result is a "very limited" ability for local staff (administrators as well as teachers) to affect policy.

Principal (Summary). Principals reflect a generally negative perspective concerning the ability for teachers to be involved in, or influence, decision making. For this group, voice is equated to the opportunity to provide feedback and input as opposed to influencing or being directly involved in decision making. Principals acknowledge opportunities to provide feedback, but suggest that most teachers choose not to participate. Principals generally see the evaluation system as being well-developed, and therefore, in no need of substantive change or modification. Principals also recognize the top-down (state policy) aspects of the evaluation activity, which limits local stakeholder opportunity to change/impact structure and/or process. Finally, there is little recognition of the many micro-decisions required to implement the system prescribed by state-policy mandate (i.e., how classification ratings are determined, setting performance standards, finalizing component weights, deciding on what/how growth measures are derived, areas of training/focus, etc.).

District (Voice). This section reviews comments provided by the district policy group (n = 4). It is noted that analysis provided below includes only three of the four district participants interviewed. For one individual it was necessary to cut the discussion

time short. Unfortunately, the question of teacher voice was positioned at the end of the discussion and was not addressed.

One district participant (below) shares his/her perspective on teacher voice, saying:

> Interviewer: How much input do you feel teachers have had in the design and the implementation of districts' evaluation system?
>
> District 301: Yeah, we have 1,500 teachers, and we have a small committee when you compare the committee of teachers. Now there has been representation from all areas of teaching on that committee. How much back and forth have those particular teachers gone back and shared with their particular group that they represent? I don't think it's been a lot.
>
> Whether the people on their campus feel like they've had any input into what those people bring back to the committee, that if you asked teachers, it's gonna be very limited. They don't feel like they've had a lot of control. The committee I believe has gone out to people. We've shared. We've tried to provide information and opportunities through surveys to provide feedback, so that we can input that in there. I don't think they'll see that they've had any real impact on what the decisions have been.

This district member acknowledges the presence of teachers on the organization's evaluation committee but that this is "… a small committee..." The amount of information sharing, dialog, and reflection is therefore dependent on the extent to which this small group of teachers inform their colleagues. The conclusion is that this is not happening to any great extent (i.e., "…I don't think it's been a lot…"). The individual is generally negative on the prospect that teachers feel empowered to influence the system saying "…it's gonna be very limited…" because teachers feel they do not "…[have] a lot of control…" There is recognition that the committee has provided opportunities for staff to provide input (i.e.,"…gone out to people… shared… provide information…

surveys…"), but that this effort is not viewed as permitting teachers to impact/alter policy (i.e., "… [no] real impact on what the decisions have been…").

A second district participant (below) reflects a more nuanced view of teacher perceptions:

> Interviewer: How much input did teachers have with how this evaluation system played out. You just made a statement that you think we've done a reasonable job of outreach and information. Do you think the teachers will feel that way?
>
> District 304: I think so, although I'm going to tell you, I believe there will be pockets, so it's gonna depend on who you speak with. I think we probably have some folks that just fear. Think about when you're in a situation and someone's explaining something to you that you're so fearful of, you can't even process it.
>
> … There'll be others that will, and it's probably because they'll have looked at it as the glass half full and not half empty … [Some] will get so caught up in, "I'm gonna get a number, I'm gonna get a number," that they won't be able to really think about the big picture. (District 304)

For this participant, teachers may express a mix of perspectives. Some may feel their voices (concerns, ideas, etc.) were being heard but "…it's gonna depend on who you speak with…" For some, fear is believed to shade their perspective: they fixate on "…I'm gonna get a number…" to such extent that they "…can't even process it …" and be unable to "…think about the big picture…" While it is not explicit in this particular narrative, the impression is that teachers who fear the evaluation process see it as an imposed, and potentially a punitive, activity. It is reasoned that these teachers may not feel empowered to raise concerns, and therefore forgo being able to influence the system.

In contrast to the two district members discussed above, a third participant projects a much more positive sentiment regarding teacher influence in the system. This person comments:

> Interviewer: Do you think teachers have had a lot of input into the design of our system? Has the input been sufficient? What's the voice that teachers have had?
>
> District 303: I think that the approach that's been taken here is very positive; a lot of teachers were involved in the creation of this. Teachers that are respected, not only are seen as good teachers, but respected on their campus and throughout the district and certainly that's always going to help from the perspective of being able to get buy-in from the staff.

For this individual, the initial response is positive (i.e., teachers provided input and their voices influenced system design). However, the perspective is based on the confidence placed in the evaluation committee's representative membership. It is reasoned that "…a lot of teachers were involved…" and that these teachers were "… respected on their campus and throughout the district…" and were generally seen by peers as "good teachers." The rationale is based on the quality of the representative group of teacher decision makers. Thus, it is the teacher "representatives" that are entrusted to make good decision, not the faculty at large. Indeed, the perspective is that all teachers must feel empowered in the process because their representative teachers are highly respected as being good teachers. The same individual continues:

> I would say from our [district] perspective, I think that it has been sufficient [teacher input]. Ya know we mentioned earlier that key teachers had a big role in the development of this piece. This committee went out and rolled this out multiple times to schools to get input. I will tell you that comparatively, based upon what I know, I think that we did substantially more than other school districts did to involve our teachers. [However] From the teacher perspective, I would say it's probably never enough. The reason that I would say that is that I think that this is such a high stakes, high pressure piece from teachers that I don't

> know that they could perceive that you do enough to be able to really roll that out. (District 303)

The individual believes that the committee "…did substantially more than other school districts did to involve our teachers…" and this effort "…has been sufficient…" to gather teacher input. However, because the system is "…high stakes, high pressure…" some teachers will always feel that "…it's probably never enough…" For this person, the combination of "…key teachers…" and numerous attempts to obtain input from the field is evidence that the evaluation process reflected the interest of most staff. Interestingly, the person closes the conversation by making the following analogy:

> [It's] no different than when we go into meet and for [budget] negotiations with folks. The group talks for the whole and then individuals don't think that they've had any input whatsoever [laughing]. Ya know it's the kinda' way it works. (District 303)

The conclusion is that no matter how well representative decision making works, there will be those who feel disenfranchised from the process because they were not actually involved. Regardless, this individual believes teacher were involved in the decision making process by virtue of the committee representation and the attempts by the committee to obtain staff input.

District (Summary). Two of the three district members interviewed were generally negative regarding teacher's sense of influence in the evaluation system. In these cases, the evaluation committee is seen as being composed of a small number of teachers; teachers at large fail to provide input to the process despite numerous opportunities; and many teachers do not feel empowered to influence decision making (some out of fear of the evaluation activity). In contrast, one member reflected more positive views, believing teachers are empowered due to the high quality of the

committee's teacher representatives. However, there will always be individual who feel disenfranchised.

State (Voice). The discussion related to voice was asked to two of the three state members. One member of the state policy group responds to the question emphatically:

> Interviewer: One more question, and it has to do with 'voice'. How much 'voice' do you think educators had in the design of Arizona's evaluation framework?
>
> State 402: I can't quantify it, but I guess I'd say close to zero. Because again, if you go back and trace, a lot of this stuff came out of Florida, came out of the politicians. It came out of the federal government. It came out of getting ready to be able to apply for Race to the Top [grant monies]. Then the groups that are actually developing putting pen to paper and proposing legislation are not represented by educators.

This individual depicts educator input into the design and development of the state's teacher evaluation system as being"…close to zero…" This is because the architectural origins of the evaluation framework were external to Arizona (i.e., "…out of Florida … the federal government…"). Additionally, the driving motivation for state policy makers was to implement a system that meet the structural standards necessary to qualify for federal grant monies (i.e., "…to apply for Race to the Top…") Since the federal criteria was a priori established, and suitable evaluation templates were available, little additional stakeholder (i.e., educator) input was needed. Indeed, the individual indicts the policy process by stating the writers of the evaluation legislation "… [were] not represented by educators…"

A second state policy representative reacts to the same prompt:

> Interviewer: Do you think that educators had a lot of voice or input into this state policy framework? Where did it come from? Who was the driving force?

State 403: I can tell you where it came from. There was an organization called Stanford Children that I think was the driving force. I'll be honest with you. I think they were very effective in arguing that we [Arizona] needed to have a teacher-principal evaluation system. I think there were enough members of the legislature that agreed. I think the business community agreed. I think there were folks on both sides of the aisle that would say, "Yeah, we need to have a better system of evaluating what teachers are doing."

As before, this person agrees that the originating source for the evaluation policy was external to Arizona (Reader's Note: There is no recognized in-state organization called Stanford Children). Here, outside interests were "…were very effective in arguing…" that Arizona required a new "… teacher-principal evaluation system …" The process was political, being determined by "…enough members of the legislature…" and "…the business community…" The underlying premise was that the system currently utilized by educators was insufficient and what was needed was "…a better system…" for assessing "…what teachers are doing…" Arguably, there is a sentiment that educator input was not considered because the system currently utilized was presumed inadequate. The same individual then clarifies the role of educators in the system design, stating:

State 403: To answer your question whether teachers, no, I wouldn't say that there was a formalized process that allowed for the people within the certified teachers to say, "Okay, this is how we think it should be done."

Interviewer: More state-level & political? But the political processes coming from some interest groups, some desire by some to do it?

State 403: Yes. Now when the law got enacted, and the state board had to design the policy around the law, I think there was significant input. I laud the individual and I won't say the name. I would laud the individual that was on the state board at the time that put that together, because it was significant and a lot of time. There was a lot of meetings. It was a statewide effort that

> really put the structure to this that I think made it a much better system once it was implemented.

The distinction being made is between the formulation of the policy (legislative) framework and the subsequent implementation details (rules, process, etc.) that was carried out by the State Board of Education. For the latter, this state member believes that educators provided "… significant input …" into the evaluation that "…made it a much better system…"

State (Summary). Both state members indicated that educators had little input into the legislative formulation of the state's teacher evaluation system. In each case, the origins of the basic framework were external to Arizona and politically/monetarily motivated. One state participant did feel that subsequent State Board of Education efforts to obtain input stakeholder input effectively incorporated teacher/educator perspectives when determining implementation details.

Voice summary. Overall, stakeholder reflections depict a generally negative view regarding teacher/educator inclusion in the decision making process. The narratives expose a general lack of empowerment, viewing the system as being externally imposed/directed in a top-down fashion. There is a distinction made between having an opportunity to provide feedback compared to actually influencing design, implementation, and application decisions. Principal perspectives differed slightly from teacher and district groups. For principals, the evaluation system is seen as well-structured, and therefore not requiring additional modification or change. There is a sense by both principal and district members that many teachers did not take advantage of the opportunities afforded them to provide input. State members agreed that educators had

little influence in developing the state legislative framework. Here, the primary actors were external. In addition, the main state policy motivations were political and monetary in origin.

Petite Assertions for RQ3 (a)

Do stakeholders perceive that policy decisions have been made inclusive of their perspectives?

1. Stakeholder narratives reflect a general negative perspective on the degree of influence teachers (educators) have had on the design, implementation, and/or utilization of the state's evaluation system.
2. The system is generally perceived at each successive stakeholder level as being externally imposed and administered in a top-down fashion with little opportunity for substantive modification.
3. Organizational efforts to provide opportunity for input is not equated to empowering stakeholders to influence decision making.

Chapter 4 Concluding Comments.

Chapter 4 attempted to present a comprehensive analysis of the data collected for Primary Research Questions 1 through 3. These analytic results set the context for Chapter 5 that provides an executive summary of the findings. These findings are placed into a broader policy perspective annotated with interpretations, discussion, and reflections. This then leads into Chapter 6 which encapsulates the findings in the form of recommendations at both the LEA and public policy levels, significance of findings, and limitations of the study. In addition, Chapter 6 contains a discussion of Primary Research

Question 4 since this question concerned personal reflections of the research process and its effect on this researcher.

Chapter 5: Findings

Introduction

In this chapter, Chapter 5, a summary of analytic findings is presented for each of the first three primary research questions posed for this study. Each is addressed sequentially, augmented with salient observations from supporting areas of evidence. An effort is made to connect the empirical results to the larger policy context under which they were derived. At the close of each section, culminating assertions and reflections are provided. The fourth primary research question (RQ4) examined personal reflections made by this researcher regarding important experiences, challenges, and learnings realized during conduct of the study. For this reason, it is addressed within Chapter 6.

The analysis previously presented in Chapter 4 examined multi-faceted validity evidence associated with implementing a state-policy imposed framework for evaluating the instructional competency of classroom teachers (Messick, 1989a). The primary intent of the analysis was to assess the warrant by which the evaluation framework permits claims of instructional quality (AERA et al., 1999, 2014; Danielson, 2010). The core of this analysis was directed by Research Question 1: *RQ1: To what degree does the validity evidence generated by the LEA's policy-directed teacher evaluation system support inferences of Teacher Instructional Quality (TIQ)?*

As mentioned, this initial question (RQ1) was sub-divided into five supporting areas of investigation: *Criterion Evidences (RQ1A), Content Evidences (RQ1B), Consequential Evidences (RQ1C), Reliability Evidences (RQ1D),* and *Construct Articulation Evidences (RQ1E)*. In turn, each component was composed of numerous supporting queries aligned to specific forms of evidence and analytic methods. In this

regard, Chapter 5 begins by summarizing the findings within these five components and how they collectively information on the primary question of interest.

Two additional study questions were also examined in Chapter 4. Research Question 2 evaluated the influence empirical information had on the localized decision making environment (Citroen, 2011). This primary question was specified as *RQ2: How does providing an evolving dialog on validity evidence influence organizational policy decisions regarding implementation of the LEA's teacher evaluation system?* RQ2 took the form of an action research study that assessed the role of researcher-as-information-broker during the design phase of the system's development (Wenger 1989; Lave & Wenger, 1991).

The third research question concerned stakeholder influence (i.e., voice) in the design and implementation of the evaluation framework (Kotter, 1996, 2010, 2011; Hord et. al., 1987; Hall & Hord, 2011). Here, the primary question was formatted as *RQ3: Third, to what degree have the perspectives and concerns of stakeholders been incorporated into, and influence, the system's implementation?*

As with RQ1, both RQ2 and RQ3 were further characterized by multiple supporting investigations. With this in mind, the next section addresses each of these questions independently followed by a collective reflection on their contributions to understanding efficacy of the evaluation activity.

Summary of Findings

Research Question 1 (RQ1)

To what degree does the validity evidence generated by the LEA's policy-directed teacher evaluation system support inferences of Teacher Instructional Quality (TIQ)?

RQ1A: Criterion evidence. RQ1A examines the statistical associations existing between the two primary components measured by the evaluation framework: ratings of professional practice (PP) and value-added (VAM) estimates of academic achievement. Table 63 summarizes the collective findings obtained for these criterion areas of evidence.

Table 63

Summary of Criteria Evidence

VAM x PP Score: Correlation	r = .254	(p < .05)
VAM x Subject (Within Year.): Correlation	r = .259 to .457	(p < .05)
VAM x Subject (Cross Years.): Correlation	r = Not Sig. to .16	(p < .05)
VAM x Year (2012-to-2013): Correlation	r = Not Sig.	(p > .05)
VAM x FFT Growth Group: ANOVA Tukey HSD (Test of Mean Difference; n =)	Low (10th %'ile):	Sig. to Mid and High (p < .05)
	Mid (45th-55th %'ile):	Sig. to Low; Not Sig. to High (p< .05)
	High (90th %'ile):	Sig. to Low; Not Sig. to Mid (p < .05)

Based on the information provided in Chapter 4, a significant, positive, correlation (r = .254, n = 238, p < .001) was observed between professional practice (PP) and Value-Added (VAM) scores, accounting for approximately 6% of common variance. Within-year VAM-subject correlations were weak-to-moderate (r = .259 to .457, n = 233 to 48, p < .05). However, cross-year VAM-subject and between-year composite VAM correlations were generally not significant. Albeit an exception to this finding was found between the Math 2012 to Math 2013 correlation (r = .163, n = 176, p < .05). Between-

year composite VAM scores were insignificant ($p < .05$). Finally, the ANOVA analysis distinguished teacher's professional practice (PP) ratings for only the 10th percentile VAM group. No PP differentiation was observed between teachers falling at the Middle (i.e., 45th to 55th VAM percentile) or High (i.e., 90th VAM percentile) achievement groups.

These criterion-related results suggest the following observations:

1. The overall relationship between achievement and observational components used within the evaluation framework is substantively weak (Cohen, 1988). This weak association (6% of common variation between PP and composite VAM scores) directly challenges the premise that each component substantively contributes to an understanding of the instructional quality construct. Indeed, the near absence of association suggests that the PP and VAM measures inferentially disagree for the majority of teachers. That is, most teachers will report relatively high PP ratings along with relatively low VAM outcomes or vice versa. Thus the interpretive meaning of the composite evaluation measure becomes suspect.
2. Within year, cross subject correlations are weak to moderate. Since the premise underlying the instructional quality construct is subject independent, weak to moderate subject correlations are counter intuitive. Why would a teacher at any given level of instructional quality report substantively different achievement growth (efficacy) measures for reading and mathematics? Lack of agreement indicts the inferential premise of the instructional efficacy component: that is, good/effective teaching is independent of the instructional

context. Indeed, there are no content-specific delineators identified in the evaluation's policy context or within the Danielson framework.

3. The general lack of correlation between the cross-year VAM subject measures indicates that the achievement components are unstable over time (Berliner, 2014; Amrein-Beardsley, 2008, 2009; Haertel, 2008; Darling-Hammond, Amrein-Beardsley et. al., 2012). That is, a teacher's Math/Reading VAM score will change substantively from one year to the next. This presents an interesting problem in conceptualizing instructional efficacy based on achievement metrics. Can a teacher's instruction competency vary substantively year to year, and if so, what is the causal instructional mechanism directing this variation?

   This question becomes more interesting when covariate-adjusted VAM models are utilized (McCaffrey et al., 2003). These models attempt to control for changes in student populations, prior achievement, and various background characteristics in order to isolate the instructional effects. Conceivably, unstable achievement measures lead to situations where a teacher's instructional ability is deemed highly competent one year and incompetent the next. If so, what is being measured, isolated, or evaluated? Can someone attain a high level of ability, then lose that ability, only to gain it back again at some point in the future? If so, should a teacher be sanctioned or terminated if assessed as incompetent in one evaluation period or over many evaluation periods? How many is enough? What if the long run trend in evaluation outcomes is also highly variable? Regardless, instability of the

component metrics raises considerable concern on the inferential rationality of the evaluation system.

4. If the VAM and PP components both inform on the construct of instructional competence, placements along the two component scales should reasonably align. Correlational evidence presented in Chapter 4 suggests this is not generally the case.

    To extend the analysis, an ANOVA procedure was used to examine this association using less stringent criteria. That is, do teachers located at highest, middle, and lowest VAM level report similar placements on the PP scale? The data found that VAM measures distinguish the PP scores only for teachers located in the lowest 10th percentile of achievement outcomes. Thus, only if a teacher's VAM score is extremely low will there be a statistically significant probability that his/hers PP scores will also be relatively low. For everyone else, knowing the VAM scores provides little information on the location of the PP rating.

    Thus, an argument might be made that, at the very least, the evaluation metrics are consistent in measure for distinguishing teachers at the lowest levels (i.e., the 10^{th} percentile) of instructional competence, notwithstanding any other issues regarding the system's inferential soundness.

Criterion evidence summary. Arguably, the criterion evidence raises substantive concerns on the inferential rationality derived from the component metrics. Evidence indicates that the component measures lack substantive association. Indeed aggregating these uncorrelated measures forces movement toward the mean of the two scales, diluting

the inferential utility of either. For example, consider a teacher who displays a high PP rating but a low VAM score. Necessarily, combining (i.e., averaging) both into a composite measure reduces the inferential significance of the PP measure and increase the importance of the achievement component.

RQ1A: Criterion assertions.

1. Correlations observed between evaluation components are generally weak: Weak associations exist between VAM & FFT.
2. VAM measures appear unstable over time.
3. Weak component associations raise questions regarding coherence of the policy-imposed empirical construct. Questions arise of whether each component (VAM and PP) is similarly informing on the hypothesized latent construct (i.e., Teacher Instructional Quality).

RQ1B: Content evidence. Table 64 summarizes the collective findings obtained for the content areas of evidence investigated as part of Research Question 1 as reported in Chapter 4.

Table 64

Summary of Content Evidences

Analysis Category	Method	Finding
FFT Representation (Danielson Framework for Teaching)	CVR (SME: *n* = 33)	23% (5 out of 22) FFT components are not substantive indicators of TIQ; 0% FFT components are inappropriate indicators. 50% (3 out of 6) elements of Domain 4 (Professional Practice) were deemed to be non-essential. *Criteria for Significance*: (*n* = 21, *p* < .05, one tail) = .359
	Confirmatory Factor Analysis	*Uncorrelated Factor Model*: Poor Fit Indices; High Factor Score Correlations (i.e. Low Discriminant Inference): *r* = .70 to .86, *n* = 238, *p* < .05 *Correlated Factor Model*: Mixed/Improve Fit Indices; High Factor Score Correlations (i.e. Low Discriminant Inference): *r* = .82 to .92
	Exploratory Factor Analysis (PAF, Oblique Rotation)	Substantive number of extracted factors = 1 to 2; Presence of a single dominant factor; Number of factors inconsistent with theoretical FFT framework; Within factor loadings inconsistent with theoretical framework with exception of FFT Domain 4 (Professional Responsibilities); Suggestion of a single dominant latent factor; Factor correlations: *r* = .80 (*P* < .001); Factor loading structures differ between experienced & less experienced teachers
	Qualitative: Stakeholder Interviews (n = 22)	Teacher, District, State (*n* = 15): Concerns: Omitted/underrepresented attributes of good/effective teaching Principals (*n* = 8): Limited substantive concerns

Note: (CVR) Construct Validity Ratio Questionnaire; (SME) Subject Matter Experts; (TIQ) Teacher Instructional Quality; (PAF) Principal Axis Factoring; (FFT) Danielson Framework for Teaching

Procedures for evaluating content evidences included use of the Construct Validity Ratio (CVR) questionnaire (content adequacy), application of confirmatory and exploratory factor analytic techniques (content cohesion and structure), and reflections provided by stakeholders (construct representation). The collection of evidences inform

on the degree to which attributes measured under the evaluation framework adequately represent the theorized teacher instructional quality construct (Lawshe, 1975; Messick, 1989a; Cronbach & Meehl, 1955).

CVR. The content validity ratio assesses the degree to which the content representation on a measurement instrument (i.e., achievement test, work performance inventories, etc.) contributes to evaluating the hypothesized construct (Lawshe, 1975; Wilson, 2012). In the context of this study, the Danielson Framework for Teaching (FFT) was the instrument being evaluated since it served as one of the primary evaluation metrics used to determine instructional competency (Danielson, 2011). The CVR approach utilized subject matter experts (SMEs) to provide feedback on each of the 22 items present on the FFT performance inventory (Lawshe, 1975). A total of 23 SMEs provided feedback including members of the district's teacher evaluation committee (n = 9) and persons employed as instructional growth teachers (n = 14). Both these groups had specialized knowledge of the Danielson framework and the evaluation process overall.

Feedback from the SMEs indicated that five (23%) of the 22 FFT rating components were not substantive (strong) indicators of good/effective teaching. Importantly, none of the components were flagged as being inappropriate or harmful to the evaluation activity. This suggests that some items might be removed from the activity, saving time, without harming the overall assessment of instructional quality. Interestingly, out of the five non-essential indicators, three were from Domain 4 (Professional Responsibilities). Since Domain 4 is composed of six rating elements, the SMEs felt that 50% of this domain was unnecessary for identifying good/effective teaching.

Factor analysis. Multiple confirmatory (CFA) and exploratory (EFA) factor analytic models were estimated to evaluate the representational coherence of the Danielson FFT (Sullivan, 1979; Kim & Mueller, 1978; Ferguson & Takane, 1989). Collectively, these CFA/EFA model specifications indicated a generally poor correspondence between the theoretical FFT framework and observed rating data. Correlated CFA models uncovered strong covariance between FFT subdomains (r = .82 to .92, n = 238, $p < .001$) indicating a general lack of discriminant inference. EFA results did not support the hypothesized four-factor model posited by the Danielson framework, favoring a one-to-two factor model.

The data suggest that the practice of deconstructing the FFT ratings by domain or component for the purpose of identifying strong/weak practices and/or areas of needed professional improvement may be inappropriate. In addition, distinguishing effectiveness based on a mix of individual component scores seems unwarranted. In the context of strong component covariance, the factor analytic evidence favors the presence of a single dominant construct indistinguishable from its component items. This suggests inferences of instructional quality be limited to one's placement along the summated rating scale.

Stakeholder feedback. Narratives regarding content representation were obtained from teachers (n = 7), principals (n = 8), district policy (n = 4), and state policy (n = 3) participants. From the narratives, teacher, district, and state stakeholders raised concerns over the lack of sufficient content representation regarding measures of instructional quality – specifically the absence of student affective impacts, affective attributes of quality teaching, and measures of long-term student outcomes (both academic and non-academic). Overall, there is a theme of omission including the assessment of non-

academic goals, objectives, and impacts. In addition, many of these stakeholders expressed concern that the state's standardized tests no longer adequately align to the new Common Core curriculum currently taught in classrooms. These factors lead participants to generally conclude that academic (standardized test) measures provided an inadequate representation of instructed content.

In contrast, principals expressed more confidence in the Danielson FFT as a representation of good/effecting teaching. Principals were also more likely to view standardized test scores as an important, objective, and reliable measure of instructional efficacy. In their view, the fundamental goal of instruction is for students to learn the state-prescribed curriculum. Since that content is directly measured by the state assessment, they value test outcomes as a legitimate indicator of this goal. Because high quality instruction is necessary for students to fully learn the curriculum, combining measures of PP and test scores provides accurate representations of instructional competence.

RQ1B: Content assertions.

1. The empirical factor structure for professional practice (PP) rating data is inconsistent with the latent factor structure posited for the Danielson Framework for Teaching (FFT).
2. Empirical PP rating information report strong factor covariances indicating a general lack of discriminant inference (variability) between behavioral domains.
3. Stakeholder narratives reflect confidence in the measured FFT components as providing a partial representation of good/effective teaching. However, it is

seen as incomplete, excluding important affective dimensions of both teaching and learning.

4. With the exception of principals, stakeholders view test scores as providing an inadequate representation of the TIQ construct.
5. Principals generally hold a much more positive view of the content adequacy of the evaluation framework than do colleagues in the teacher, district policy, or state policy groups.

RQ1C: Consequential evidence. The impact of the evaluation system was examined from a stakeholder perspective (n = 22) both in terms of intended and unintended outcomes. All four participant groups (teachers, principals, district policy, and state policy) were asked to reflect on how the evaluation system impacted instruction and student learning. Additional perspectives were explored with regard to school climate, impact on teachers and the teaching profession, and consequences associated with the framework's designed and implementation.

Figure 59 is an outline of main findings obtained from the stakeholder reflections. The presentation divides the findings into positive and negative affirmations. For each, stakeholder group representation is identified as a header to the bulleted items.

Negative Affirmations	Positive Affirmations
[Teachers, District, and State: *n* = 14] o *Test Scores:* Test scores are seen as generally inadequate/incomplete primary measures of instructional quality o *Purpose:* Participants expressed generally negative sentiments regarding the potential for the evaluation system to substantively improve instruction and/or increase student learning o *Omission/Reductionism:* Participants believe attributes of omission and/or reductionism raises stress/fear, lowers trust, and generally harms the professional identity of classroom teachers o *Clarity/Focus/Structure:* The perspective is that important non-tested, affective, content/learnings are crowded out due to the simplified/reduced-form characterization of instruction and emphasis of standardized test scores	[Principals: *n* = 8] o Principals express substantively positive views of evaluation impact on both instruction and learning; ▪ *Impact*: For principals, nearly all impacts of the evaluation system are viewed as positive. Teacher stress/fear is viewed as a positive consequence. ▪ *Clarity/Focus/Structure*: Leads to increased opportunity for communication, dialog, reflection on measured attributes (Teachers, Principals, State)
[Teachers, District: *n* = 11] o *Top Down/Imposed*: Participants viewed the evaluation system as being externally imposed and primarily top-down. This suggested an environment of compliance rather than acceptance. o *Clarity/Focus/Structure*: The prescriptive structure of the framework leads to reductionism, narrowing of curriculum and instruction, and a reduction of instructional creativity	[Teachers: *n* = 7] o *Communication/Dialog*: To the extent that time/effort is afforded to the evaluation process, increased communication/dialog on instructional practices is a positive consequence. o *Clarity/Focus/Structure*: The FFT provides clarity on the specific pedagogical skills/behaviors being evaluated. This permits staff to focus on improving performance in those areas.
[District, State: *n* = 7] o *Impact on Climate/Profession:* Increased difficulty in attracting/retaining teachers, harm to school climate, administrator's loose of focus on other important administrative/instructional leadership duties	[State: *n* = 3] o *Foundational Accountability*: The measured components represent a foundational set of indicators that should be part of the evaluation process.

Figure 59. Summary of consequential reflections by stakeholder attribution.

Teacher, district, and state participants (n = 14) expressed concern over the increased levels of stress attributable to the evaluation system. This was due, in part, to a lack of trust in the measures, concern over the high stakes, consequential, nature of evaluation outcomes, and an increased emphasis on test scores required by the state-policy framework. For many of these participants, the increased emphasis placed on test scores acts to narrow instructed content and restrict instructional flexibility and/or creativity. As a result, students are deprived from learning experiences not explicitly measured by state assessments.

These same stakeholders also suggested that the evaluation process harms teacher's professional identity and morale, as well as damaging school culture. Some state representatives felt that the excessive amount of time required to conduct school-wide evaluations takes away from other important leadership responsibilities.

Recognizing the excessive time restrictions placed on administrators, many teachers expressed concern over the lack of time evaluators have to obtain a deep understanding of their day-to-day practices. *Lack of Time* limits the teacher-mentor relationship, which is seen as critical for communication, dialog, and critical reflection. As a result, evaluators may lack sufficient grounding to evaluate the individual adequately. Finally, teachers tended to discuss their perspectives in terms of adherence and/or compliance. Here, lack of empowerment leads to resignation and compliance as opposed to acceptance and internalization of the evaluation system's stated policy purpose and objective.

Overall, teacher, district, and state participants felt that the evaluation activity has limited impact on changing instructional practice and will not substantively impact

student learning. Nuances within the perspectives shared by these stakeholders seem generally inconsistent with policy intent, act to indict claims of evaluation purpose, and raise questions of the framework's fidelity to theoretical construct.

In contrast to the above, principals (n = 8) reflected more positively on the impact of the evaluation process. In general, they believe the system adequately represents attributes of good/effective teaching. By requiring compliance with its components, instructional competency will improve, resulting in greater learning. In their view, stress is a natural consequence of evaluation. Any act of compliance has the effect of focusing attention on important attributes of professional practice. Thus, for principals, conformance/compliance is seen as a benefit. The result is that the evaluation structure/process is having a positive impact on both students and the instructional capacity of teachers.

RQ1C: Consequential assertions.

1. Representatives from the teacher, district, and state membership groups express generally negative sentiments regarding the evaluation system's ability to facilitate improve instructional practice and/or student learning. In contrast, members of the principal group uniquely believe the evaluation structure is having a net positive impact on instruction and learning. State participants were more likely to see evaluation as a necessary and worthwhile activity, while at the same time acknowledging its structural, procedural, and inferential limitations.

2. Some problematic issues expressed by stakeholders include:
    - The system is externally imposed and inconsistent with an evaluation structure that would be developed by local educators; external imposition leads to a context of compliance rather than acceptance;
    - The system omits important affective attributes/outcomes associated with good/effective teaching;
    - Structural and measurement issues lead to increased levels of stress/fear;
    - The system lowers trust, negatively impacts school culture/climate, and harms the morale and professional identity of classroom teachers.
3. An increased emphasis on test scores (compared to prior evaluation efforts) acts to narrow areas of instructional focus, restricts instructional creativity and generally limits student learning.
4. Issues of time and effort may divert school administrators from other important school/instructional activities.
5. The evaluation framework brings *clarity/focus/structure* to the evaluation process. This has both positive and negative consequences.
    - Positive: Focuses efforts on pedagogical behaviors believed to be important for effective teaching; clarifies the standards and expectations that instructional staff must adhere; establishes a consistent set of metrics for all stakeholders.
    - Negative: Framework is simplistic, reducing the complex dynamic nature of teaching down to a narrow set of pedagogical behaviors and limited

academic learnings. Inadequate representation harms trust and acceptance of the system.

6. Implementation of the framework exposes substantive perceptual differences between principals (evaluators) and teachers (recipients) regarding system efficacy, purpose, and application.

RQ1D: Reliability evidence. In general, reliability is a measure of score precision. Relatively high reliability indicates low error variance (i.e., error due to random chance, factors unrelated to the examinees true ability), while low reliability indicates the opposite, implying the presence of larger amounts of random error (Traub & Rowley, 1991; Allen & Yen, 1979; Crocker & Algina, 1986). Low reliability is a direct threat to validity in that the empirical information fails to adequately reflect the trait of interest, obscuring authority of inferences made from such measures (AERA et al., 1999, 2014; Wainer & Brown 1988).

Numerous types of reliability analyses were conducted on the professional practice (PP) and value-added (VAM) measurement scales including forms of Coefficient Alpha, Coefficient Theta, Interclass Correlation Coefficients, standard errors of measure, and stakeholder perceptions of evaluation accuracy (Crocker & Algina, 1986; Carmines & Zeller, 1979; Zumbo et al., 2007; Gadermann et al., 2012; Bonanomi et al., 2013). Table 65 summarizes the collective reliability evidences outlined in Chapter 4.

Table 65

Summary of Reliability Evidences

Reliability	PP Composite	α, Ord. α, Ord. θ	α = .95 to .98	
	PP Sub-Domain (4)	α, Ord. α, Ord. θ	α = .81 to .94	
	VAM (Level-1)	Psuedo-r^2	r^2 = .71 to .77	(p < .05)
	VAM (Level-2)	ICC	ICC = .05	
	True Score Item-Scale Range	SEM	22% to 37%	
	VAM/PP Reliability/Bias	Qualitative: Stakeholder Interviews (n = 22)	Teacher, District, State (n = 15): *Concerns*: Rater Consistency, Bias, Construct-Irrelevant Variance (External Influences); *Causal factors*: insufficient observation time, attribute omissions, test scores as inadequate Principals (n = 8): Limited substantive concerns	

Professional Practice (PP) scale reliabilities. Cronbach Alpha, Ordinal Alpha, and Ordinal Theta reliability indices were computed for the four PP subscales (i.e., evaluator ratings on Domains one through four). These measures reported generally high levels of internal consistency (.81 and above): values ranging between .810 (Cronbach alpha) to .940 (Ordinal theta). Reliability measures for the composite PP scale ranged from .949 (Cronbach alpha) to .976 (Ordinal theta).

For some authors, these values indicate good-to-excellent levels of reliability, generally suitable for use within many testing and measurement activities (Gadermann et al., 2012; George & Mallery, 2003; Rudner & Schafer, 2001). However, it is debatable if

alpha/theta reliabilities between .80 and .90 are sufficient to warrant use within high stakes, consequential evaluation settings (Rudner & Schafer, 2001).

Arguably, reliability levels need to be commensurate with the inferential judgments and consequential outcomes derived from its measures (Mason, 2007, Baird & Black, 2013). That is, as score reliability declines, the probability of misclassification increases, and if such classifications have direct and lasting impact on the professional and personal welfare of individuals, the need to minimized incorrect classification becomes a paramount concern (Baird & Black, 2013). However, Mason (2007), in his review of state testing programs, notes that while test manufacturers routinely report empirical analysis concerning the reliability and validity of their products, the information is often incomplete and/or difficult to understand, saying:

> … in general, information about reliability and validity in all states was incomplete; too technical for teachers or parents to understand, or complete and detailed enough to show that further study was necessary before validity could be assumed. (p. 39)

He observes that "… the sufficiency of this evidence to support validity of the purposes of the high stakes test application was less clear…" (p. 39) and concludes

> …if the purpose is to be the only source of information on which critical life-changing decisions about people who work and learn in the schools are based, then the tests must be shown to be valid for that purpose. There is still work to be done. The stakes are high! (p. 44)

The issue raised is what constitutes sufficient (reliability) evidence to warrant making high stakes claims and decisions. In addition, Catano et al. (2012) contend that different assessment contexts require different reliability constructions which can substantively impact related questions of validity. Regardless, there seems consensus that reliabilities need reflect the context in which the scores are utilized (AERA et al., 1999, 2014; APA, 2001) and that more research is needed in this area (Mason, 2007; APA, 2001). But the question remains

how precise is precise enough? In the context of teacher evaluation, where judgments of professional efficacy are reduced to empirical measures, and such judgments have the potential to impose life-changing consequences, requiring scale reliabilities in excess of .90 does not seem unreasonable. Thus, inferences made off the summated PP scale seem more reasonable those derived from the subscales.

VAM model precision. Reliability is an indicator of measurement precision. Arguably, measurement precision is critical for making inferences in high stakes settings. In traditional regression contexts, dispersion indices such as the Coefficient of Determination (aka: r-squared) and standard errors of measurement provide useful reflections regarding statistical precision. However, this study utilized multi-level value-added model specifications of academic achievement. As such, strict forms of r-squared are not defined and construction of pseudo-r-square indices serve as suitable proxies (Raudenbush & Byrk, 2002; Snijders and Bosker, 1999; McCoach & O'Connell, 2012). In addition, another useful indicator of variation is the Interclass correlation coefficient, which indicates the amount of variation in the dependent variable ascribed to higher level grouping variables (Luke, 2004; McCoach & O'Connell, 2012).

As reported in Chapter 4, VAM model level-1 variance explained (psudoe-r^2) indices ranged from .71 to .77. The question becomes whether these levels are suitable for high stakes decision making. In practice it is not uncommon for authors to reference prediction models with r-square values of .70 to .80 (i.e. Goldberger & Hansen, 2010; Harris, Ingle, & Rutledge, 2014) or lower depending on context and purpose of the modeling activity (Plonsky, Frederick, & Oswald, 2014; LeCroy & Krysik, 2007). Others insist that under high stakes contexts such models need to display precision levels in

excess of .80 – that is, equivalent to Pearson correlation values above .90 (Gadermann et al., 2012). Applying this criteria, VAM model estimates fail to reach this threshold.

Review of the VAM model Interclass Correlation Coefficients (ICC) indicated approximately 5% of explained variance remains associated with level-2 school effects, suggesting model fit may be further improved (McCoach & O'Connell, 2012). It is noted that improving model fit translates into more precise achievement estimates and smaller true-score confidence intervals permitting greater confidence in score inferences (and lower rates of misclassification).

Utilizing the VAM-model standard errors of measure, the assembled collection of 95% confidence intervals may be expressed in terms of the number of tested items. In this regard, the range of standard error estimates represents between 22 to 37% of the total items on the referenced achievement tests. That is, the *practical* error evident in the predictive modeling process applied to a (for example) 50-item multiple choice achievement test equates to between 11-to-19 items. Arguably, this is a substantive range of practical error for estimating the true academic abilities of students or for inferring the instructional effectiveness of individual classroom teachers. Thus, discriminating small distinctions between levels of instructional efficacy using point estimates from the VAM models may not be warranted because of the substantive error around those locations.

Qualitative reliability measures. Qualitative interviews with teacher, district, and state stakeholders (n = 15) documented concerns related to rating consistency, bias, and the impact of non-instructional factors on both PP measures and VAM scores. These perceptions of evaluation/rating accuracy were closely tied to the concept of time (lack thereof) and evaluator's inadequate familiarity of a teacher's day-to-day professional

practice. For both classroom observation and the single-event testing experience, *lack of time* increases the potential for external, uncommon, non-instructionally related events to distort the measured outcomes resulting in inaccurate representations of instructional quality. Finally, for these three stakeholder groups, unobserved/unmeasured attributes of good/effective teaching exposed a potential for measurement bias. From the narrative, the omitted attributes are dominated by the affective, non-academic, impacts good/effective teachers are believed to have on students.

In contrast to the sentiments offered by teachers, district, and state participants, principals (n = 8) provided a more positive perspective on score reliability. They expressed confidence in their training as evaluators, the adequacy of time allocated to instructional observation, and the importance/relevance of content-area tests to provide an objective assessment of instructional efficacy. Similarly, principals expressed confidence that evaluation criteria are fairly and consistently applied and the resulting outcomes provide an accurate and unbiased representation of teacher competence.

Collectively, the reliability evidence assembled for the different evaluation components (VAM, PP, and Stakeholders) suggest inadequate levels of precision to warrant their use in consequential decision making. The qualifying context of this assertion is the high stakes nature of the evaluation activity. Arguably, the consequences derived from evaluation inference are substantive: impact on individual's professional and personal identity, professional standing in his/her organization, and current/future employment status. These consequences necessitate higher standards of reliability (Rudner & Schafer, 2001; Harris, 2013).

In this regard, VAM error intervals equate to large intervals in tested items and substantive true-score uncertainty. Reliability measures on PP sub-scales approach, but do not reach, desired thresholds. And narratives provided by a variety of stakeholders raise concern over the precision of both achievement measures and observational assessments of instructional quality.

RQ1D: Reliability assertions.

1. VAM Model reliability indices suggest inadequate levels of precision for making high stakes consequential inferences of instructional efficacy.
2. PP sub-scale reliabilities approach, but do not attain, requisite levels of precision to warrant high stakes judgments of instructional quality at the domain level.
3. Stakeholder perceptions of evaluation accuracy reflect substantive concerns in the system's ability to accurately identify and distinguish instructional competency.

RQ1E: Construct articulation. Validation studies examine the appropriateness of inferences made from empirical information with regard to some specified latent construct (AERA et al., 1999, 2014). Such inferences derive their warrant from two conditions: the clarity of the theoretical (latent) construct and the alignment of the empirical framework used to reify that construct (Messick, 1989a). In this context, without a well-articulated theoretical construct, alignment of empirical measures cannot be established rendering claims of inferential authority suspect.

The theoretical construct posited by the evaluation policy framework is referred to as Teacher Instructional Quality (TIQ). Importantly, no operational definition of TIQ is

provided in the legislative context (ADE, 2012a; Ariz. Rev. Stat. §15-203A.38, 2010; Task Force on Teacher and Principal Evaluations, 2011). By default, the policy environment indirectly operationalizes TIQ through its mandated measurement framework. That is, TIQ is reified by the behavioral components contained within the Danielson FFT and standardized tests scores in reading, mathematics, and science. As a result, validation questions concern whether this reified structure is consistent with stakeholders' sentiments of what it means to be a good/effective teacher. Substantive alignments would lend authority to claims of instructional competence made on the basis of the measured components, while insufficient connections raise questions of inferential fidelity.

To this end, RQ1E examined the perspectives held by stakeholders regarding the characteristics exemplifying good/effective teaching. The intent was to develop an operational context of the construct. To this end, stakeholders (n = 22) were asked to reflect on attributes which identify and distinguish good/effective teachers including influences teachers have on students and their own professional practice. Figure 60 summarizes the main points evident from this narrative.

Limitation of Pedagogical Competence:

- *Non-Distinguishing:* As a professional requirement, most teachers possess pedagogical competence in the areas of instructional delivery, methods, skills, and content knowledge. However, pedagogical competence is insufficient for representing the TIQ domain. As such, these attributes alone are insufficient to substantively identify and distinguish good/effective teachers.

Content Learning/Test Scores as an Insufficient Attribute:

- *Test Scores*: Test scores convey a limited and insufficient context for identifying and distinguishing good/effective teachers. High test scores may be attributed to minimally competent teachers due to characteristics already present in the student population (external learning, preparation, and support; self-motivation, interest, commitment, perseverance; etc.). In this way, test scores insufficiently represent and discriminant across the domain of instructional competence.

- *Academic Outcomes*: Similar to test scores, for some students, academic outcomes (graduation, attendance, college enrollment, course completion, etc.) may be attributed to factors unrelated to the pedagogical competence of teachers.

- *Content learning* is only one of many longer-term educational outcomes realized by good/effective teachers.

Affective as a Dominant Attribute:

- Affective attributes represent an important characteristic for identifying and distinguishing good/effective teachers separate from pedagogical competence. These include affective impacts on students (personal, emotional, self-perceptual, etc.) and affective dimensions of professional practice (personal/emotional professional connection, commitment, dedication, identity, etc.) .

Indirect Curriculum:

- *Non-Academic Goals and Objectives:* Good/effective teachers recognize non-content centered goals and objectives as legitimate and necessary educational outcomes. This takes the form of an *indirect curriculum* linking the affective to the academic. Here, teachers elevate the importance of the short-term affective influence to ensure long-run academic success and personal well-being. In this way, good/effective teachers attend to both direct (written and tested) and indirect (non-academic) curriculums (i.e. attendance, engagement, self-perceptual, etc.)

Professional Affective:

- *Emotional Investment*: Good/effective teachers are exemplified by their emotional investment in the non-academic welfare and development of students

- *Professional Affective Descriptors:* Affective attributes related to teacher professional practice may be characterized by the following descriptors: passionate, attitude, motivation, commitment, emotionally invested, facilitate learning, build/nurture relationships, architects of a personalized learning environment; role models (teacher/classroom as a microcosm of society; to develop citizens); concern with the individual, personal, and emotional well-being.

Teacher as Architect/Engineer:

- *Learning Environment:* Good/effective teachers purposefully engineer the environment, conditions, and context of learning. This concept is primarily concerned with the affective not the physical: that is, the personal, emotional, relational, and experiential. Learning environments are constructed to be 'safe', allowing students to take personal risks and engage in struggle in order to succeed in the learning process.

- *Student Transformation*: Attention to student transformational influences is equally purposeful, goal oriented, and actively pursued.

Student Affective:

- *Transformational*: Good/effective teachers are characterized by the transformational influences they have on student's personal/emotional connection to the learning process. This includes changes in student's sense of the self (value, confidence, belief). The distinguishing attribute is that these transformations would not have occurred without the purposeful effort of the teacher.

- *Relationship*: Good effective teachers focus on the individual student, building understanding of the personal context of learning through the development of relationship. The focus is on the individual and the contribution made from strengthening the teacher-student connection.

- *Student Affective Descriptors*: Student affective transformations are characterized by the following types of descriptors: inspire, motivate, transform, instill desire, trust, self-worth, confidence, future, hope, etc.

Figure 60. Attributes of the Teacher Instructional Quality Construct.

All stakeholders (n = 22) were asked to reflect on what it means to be a good/effective teacher. Across the four groups (teachers, principals, district policy, and state policy participants), narratives uniformly emphasized affective, non-academic, aspects of instructional pedagogy. In all narratives, the focus was on affective impacts and the foundational importance of the teacher-student relationship. Importantly, the domain characterizing good/effective teaching was only minimally defined in terms of test scores or related academic outcomes. This was true even for the principal group.

Most participants alluded to a pathway of causation, purposefully engineered through a teacher's professional practice. In this pathway, good/effective teachers create relationships, relationships establish trust, and trust facilitates connection to the learning process. Content learning is seen as a longer-term outcome as opposed to a short-term goal. In this way teachers are positioned more as facilitators of learning rather than deliverers of knowledge.

Stakeholder narratives also reflected a hierarchy of need that is anchored in the affective. Here, good/effective teachers actively transform students from passive to active learners as a prerequisite to the learning activity. This hierarchy of need is time dependent leading to recognition that short-term content mastery may be secondary to developing a personal connection to learning. In this way, there is a concept of an indirect curriculum, one that is non-academic but equally important, emphasizing development of motivation, engagement, desire, inspiration, and connection.

The narratives also suggest that good/effective teachers display high levels of personal commitment, passion, and dedication to their students. Here, professional identity is substantively defined by elements of the affective and the desire to be

transformative. That is, student mastery of tested content is not the driving force for entering the profession (i.e., to simply teach math so students might pass a test). Rather, it is the broader belief in the importance of education to shape the individual that provides identity and distinguishes a good/effective teacher.

In this broader sense, stakeholders see good/effective teachers as role models within a personal, social, and community context. Here, aspects of character, behavior, and responsibility hold equal/more value than subject area knowledge. There is an overarching purpose for education, for teaching, and the role/impact teachers may have that extends beyond academic content mastery.

The narrative provided by stakeholders regarding characteristics of good/effective teaching collectively emphasize affective attributes of the instructional environment and devalue elements of pedagogical competence and academic achievement. It is not that these latter elements are unimportant or irrelevant. Rather, they are insufficient for suitably representing the many-faceted dimensions of the TIQ construct.

Some qualification to these findings might be found in the nuanced narratives provided by principals. Distinct from other groups, principals were more likely to view the current evaluation structure as producing legitimate measures of good/effective teaching. Nevertheless, their narrative regarding a more idealized construct of instructional quality includes the same affective attributes shared by their colleagues. It may be that this group recognizes the importance of these nuanced affective attributes but accepts that they are difficult to measure.

Comment. It seems evident from the collective narratives that evaluation systems restricted to standards-based pedagogical competences (i.e. instructional actions and

behaviors) and standardized test scores fail to capture important aspects of the education experience. These omitted attributes are affective in nature and relate to desired outcomes that transcend the academic. In this way, such systems provide an inadequate foundation for identifying and distinguishing instructional quality. Importantly, the stakeholder narratives are very much aligned to views offered by theorists such a Dewy (1900, 2009), Good (1999), and McGee-Banks (1997) as discussed in-depth within Chapter 3 and to those advanced by Labaree (1997), Sloan (2012), and Cohen (2006).

For these authors, the concept/purpose of school is interconnected with larger personal and societal goals. For Dewey (1900, 2009), any distinction between the purpose of school and the needs of society is artificial and undesirable. Good (1999) and Labaree (1997) submit that the importance of school is to instill a connection between learning and the larger community in terms of relevance, application, and contribution. In this context, a requisite measure of instructional quality would be whether a student is able to integrate knowledge into a larger social context with meaning and purpose. To this end, Cohen (2006) argues that it is a student's "…social-emotional skills, knowledge, and dispositions [that] provide the foundation for participation in a democracy and improved quality of life…" (p. 201). This suggests that metrics of school (or instructional) quality involve assessing student's understanding (and importance) of academic knowledge within a social context. That is, it is not *what* is learned but rather *why* and *how* these learnings may be applied to social experience. In this way, Dewey (1900, 2009) and others view schools as serving the community with the goal of instilling connection between the individual and society. It is argued herein that the stakeholder narratives align closely with this perspective.

In a similar context, Sloan (2012) contends that "… parents and teachers want schooling to support children's ability to become life-long learners who are able to love, work, and act as responsible members of the community" (p. 2). Again, these attributes represent the same affective dimensions advocated within the study narratives regarding characteristics/impacts of good/effective teaching. Interestingly, Peifer (2014) reports that a 2010 National School Boards Association survey of school board members identified two dominant goals for education: To "*Help students fulfill their potential (43%)* and *Prepare students for satisfying and productive life (32%)* (p. 1). Indeed, the next closest goal category was to *Prepare student for the workforce* at only 8%. Arguably, goals involving *fulfilling potential* and having a *satisfying and productive life* concern an individual's emotional well-being. And these attributes are influenced by aspects of self-awareness, confidence, belief in the future, social engagement, responsibility, etc. – all components identified as either omitted or underrepresented under the current policy-imposed evaluation context.

Similarly, Cohen (2006) speaks about the need for "… pedagogy informed by social-emotional and ethical concerns… " rather than ones restricted to content mastery and skill attainment (p.201). As the stakeholder narrative suggests, schools and teachers should be positioned to empower students to fully integrate into society by developing the moral/ethical values, social/civic awareness, critical self-reflection, and other facets of *the person* which permit students to become productive citizens in a global society. In this regard, McGee-Banks (1997) councils that "… educators must not only educate the mind, they must also educate the heart and create a sense of hope, commitment, and

possibility…" (p. 188), aspects of evaluation that stakeholders suggest are left unmeasured and/or underrepresented.

Arguably, the findings presented herein suggest stakeholders believe the current evaluation environment fails to incorporate critical facets of the school/classroom experience. In addition, they content these facets represent important instructional effects consistent with good/effective teaching. Importantly, their concerns align with perspectives offered by education philosophers dating back over a century. At their origin, study participants see the purpose of education as holistic, part of the larger personal and social context. In this way, the analysis presented brings relevance and immediacy to the perspectives of theorists such a Dewey, Good, Labaree, Sloan, and McGee-Banks and renders authority to the concerns expressed by study participants.

Summary. In summary, stakeholder narrative provided in Chapter 4 suggested problems of omission and/or underrepresentation of affective elements. In addition, measures of academic attainment are viewed as providing an incomplete and biased perspective of instructional competency. These expressions are consistent with the "idealized" profile of good/effective teachers documented under RQ1E. Collectively, these evidences suggest areas of disconnect between the reified evaluation framework and the theorized TIQ construct it is posited to represent. Arguably, omission is the primary issue and the affective is the dominant artifact excluded. With this omission, it is argued that the current evaluation framework fails to fully capture the dynamic, complex, nature of teaching.

RQ1E: Construct articulation assertions.

1. Stakeholders conceptualize good/effective teaching primarily in terms of affective attributes, including non-academic impacts on students and the personal/emotional attributes of professional practice.
2. Importantly, academic achievement (i.e., test scores) and pedagogical competence are not viewed as the defining/distinguishing characteristics of instructional quality. Rather, these attributes are viewed as component pieces within a more complex context.
3. To the extent that evaluation measures focus on academic achievement and standardized skills/behaviors of professional practice, important affective attributes of quality teaching are being omitted and/or under-represented. In this way, a fundamental misalignment exits between the reified evaluation framework and conceptualizations what it means to be a good/effective teacher.

Concluding remarks regarding Research Question 1. Research Question 1 examined multi-faceted, multi-method evidence regarding the assessment of Teacher Instructional Quality (TIQ). By design, the policy framework required use of quantifiable measures of professional practice and student achievement to identify and distinguish teacher performance within a high stakes, consequential environment. The consequential setting of the activity substantively raises the burden of evidence necessary to validate inferential judgments of instructional efficacy.

At best, the collection of evidence presented herein suggests that the evaluation framework presents an incomplete representation of instructional competence. Problems of attribute omission and underrepresentation in the instructional domain reveal a

substantive misalignment between the theoretical and reified frameworks. This omission places undue emphasis on the measured pedagogical competence and academic learnings at the expense of unmeasured affective instructional influences and non-academic outcomes. In addition, the framework forces emphasis on short-term outcomes that may be easily measured, leaving important long-term educational and personal outcomes unaccounted.

At worst, the framework presents an unreliable and/or biased view of instructional quality. Weak component correlations, instability of value-added (VAM) metrics, lack of discriminate inference between professional practice (PP) subscales and items, each represent substantive measurement problems for the context in which the information is being utilized. Issues related to insufficient observation time, suspect rater reliability, and the influences of non-instructional artifacts out of the control of teachers are worrisome. Each indicts the efficacy of inferential decisions made under the authority of the policy-imposed evaluation activity.

Taken collectively, the evidence suggests that high stakes, consequential decisions that substantively impact the professional standing of classroom teachers are unwarranted and should not take place until improvements in the evaluation process are realized. Suggested areas of improvement include the need to:

1. Develop a clearly defined construct of the Teacher Instructional Quality (TIQ) domain including the full range of attributes, influences, and outcomes valued by the educational process. Without such articulation, inference on teacher performance made from a misaligned or incomplete reified framework lacks authority.

2. Develop a reified evaluation framework directly aligned to the articulated TIQ construct, not the other way around.
3. Provide additional evidence of measurement quality (i.e., interrater reliability, bias, precision). This requires integration of appropriate analytic research designs and methods directly into the evaluation activity.
4. Revise the type and methods by which academic content learning is conceptualized and measured. Efforts must address biasing influences of construct-irrelevant factors and expand upon the collective representation of instructional impact. Single event tests of content mastery, even when expressed in terms of growth, are insufficient. Longitudinal aggregation of single-event measures fails to address inherent biases introduced by construct-irrelevant factors.
5. Multiple independent forms of academic assessment should be incorporated that de-emphasizes single-event, single source methods.
6. Incorporate long-term educational outcomes as important artifacts of good/effective teaching. Short term measures of academic outcomes fail to represent these impacts.

RQ1 closing assertions.

1. The state policy-imposed teacher evaluation framework presents an incomplete representation of the teacher instructional quality (TIQ) construct. Attribute omission and underrepresentation render the primary measures specified by the policy framework insufficient for making high stakes consequential decisions of instructional competence.

2. Lack of an articulated representation of the latent TIQ construct constrains claims of instructional competence to those components specifically measured by the current system (i.e., the 22 pedagogical components reified by the Danielson Framework for Teacher (FFT) and academic outcomes assessed by the state's standardized achievement test).
3. Lack of strong associations between professional practice and academic achievement measures indict the proposition that each component independently informs on the posited TIQ construct. Weak associations suggest that the two measures present contradictory inferences of instructional competence for the majority of evaluated teachers. Finally, weak component associations suggest an inadequate representation of the posited instructional framework.
4. The scale characteristics of some measures used in the evaluation framework render them insufficient for supporting high stakes, consequential inferences of instructional quality. Less than adequate scale reliabilities, weak criterion associations, plus concerns over construct-irrelevant influence collectively degrade the inferential authority of the evaluation outcomes.
5. The empirical data do not support the hypothesized four-domain (factor) structure posited by the Danielson FFT. Strong factor covariances reduce inferential discrimination between behavioral components, rendering high stakes, consequential decisions based on subscale ratings suspect.
6. Study data indicate a substantive divergence of views held between principals and the three remaining stakeholder groups included in the study. Teacher,

district, and state participants provide generally negative sentiments regarding the potential for the evaluation system to substantively improve instructional practice and/or student learning. This contrasts with the perspectives offered by the district's site principal. Overall, principals reflect more positively on most aspects of the evaluation process compared to other stakeholder groups.

Research Question 2 (RQ2)

How does providing an evolving dialog on validity evidence influence organizational policy decisions regarding implementation of the LEA's teacher evaluation system?

Research Question 2 (RQ2) examined the following general question: Does injecting empirical information into the evaluation development activity substantively influence the decision making process? To this end, policy leaders were asked whether they valued the infusion of information and if it helped shape and influence their decisions. Inherent in this activity was the role played by the information broker. To this end, the analysis addressed the following: the value of information, the influence of information, and the role of information broker within the decision making context. Data was gathered from personal interviews with two key policy-level decision makers, uniquely positioned to reflect on these contexts due to their long involvement in the evaluation design activity and positions held in the organization.

Summary: Value of information. For participants, the value of information lay in its ability to help validate decision making. Access to detailed and objective information raised member confidence, believing they were fully informed and learned in the topic areas. Access to the data also permitted more comprehensive analysis, critical

reflection, and dialog among committee members and policy leaders. Collectively, this facilitated added trust in the decision making process.

In addition, the availability of information and related analysis permitted ongoing communication with the broader population of stakeholders (teacher and administrators), satisfying the committee's commitment to transparency and open communication.

Importantly, committee members viewed the evaluation process as a high stakes, consequential activity, potentially impacting teacher's professional and personal identity. Here, the value of information was enhanced by the high stakes nature of the evaluation process because it served as the basis to question and reflect on intended and unintended outcomes. In summary, participants afforded great value to the information made available to the evaluation committee, policy leaders, and organizational stakeholders.

Summary: Researcher as information broker. Participants acknowledged the importance of the information broker within the decision making process. The position served as a catalyst for developing and presenting information to the decision making team. Value was afforded to the unique set of skills and knowledge not held by other members of the evaluation committee or other policy-level stakeholders.

While a foundational role of the information broker centered on the delivery of empirical information to the evaluation committee, this activity was not seen as exclusively important. Rather, the information broker helped develop deeper understanding and clarity of interpretation by communicating complex information in a way that was accessible to decision makers. This resulted in heightened trust and confidence in the decision making process and associated outcomes. In summary, the

information broker was viewed as an essential component of the committee's work to design and implement the evaluation system.

Summary: Influence of information. Based on participant narratives, providing ongoing empirical information impacted the decision environment in the following areas:

- *Changes to the Training Focus for Evaluators*: A shift of emphasis away from familiarity with evaluation structure and process onto improving evaluator skills, knowledge and authentic application of the (Danielson) scoring rubrics.
- *Increased Attention on Ensuring Evaluator Interrater Reliability*: An increased focus on the application of rubric criteria to real-world instructional settings for the purpose of improving and verifying scoring precision.
- *Conduct of a Formal Interrater Reliability Verification Study*: Design and implementation of an interrater reliability study to assess the degree of rater consistency and precision in the evaluation process.
- *Facilitating Changes in Perspective*: Empirical information facilitated changes in a priori perspectives held by decision makers by providing the basis for critical reflection, communication, and discussion.
- *Facilitating an Increase in Clarity and Confidence*: Empirical information provided authority for committee members to have enhanced confidence in decisions impacting evaluation structure and process.
- *Articulating the Need for Future Emphasis on Post-Conference Teacher Mentoring*: Empirical information revealed a need to incorporate post-conference mentoring and support activities into future evaluator training.

- *Implementation of the SPA (Support Plan) Initiative:* Review of empirical information assisted in articulating the need to add a Support Plan initiative to the evaluation model that focuses on struggling teachers.
- *Guiding the Standard Setting Decision Process*: Review of empirical information was critical to determining performance standards and related instructional classifications.
- *Devaluation of Test Scores in the Evaluation Formula*: Critical reflection of disconfirming empirical information raised questions, dialog, and critical reflection on the suitability of test scores to serve as a heavily weighted component in the evaluation framework. As a result, decision makers devalued the importance of achievement indicators to the minimum allowed under state legislation. The empirical information provided foundational support for this decision.

Stakeholder narratives outline numerous areas of impact including changes in training focus, changes in personal perspectives, the addition of new components to the evaluation environment, and the importance of information to making critical decisions regarding standard setting and component influence. In this way, provision of empirical information was viewed by participants as an important and influential activity to the evaluation system's design and implementation.

RQ2 closing assertions.

1. The provision of empirical information has a profound and lasting effect on the design and implementation of teacher evaluation systems by challenging a

priori perspectives, facilitating critical reflection and dialog, and providing foundational authority to decisions defining the evaluation environment.

2. The provision of accurate, complete, unbiased information into the decision making process is a critical condition for establishing authority and confidence in the final design and application of the evaluation activity. The focus and type of empirical information must align with the purpose of the evaluation process and offer account on a priori perspectives held by decision makers.
3. The position of information brokers serves an important role in the decision making process, bringing specialized skills, knowledge, and expertise into the policy environment. The information broker facilitates understanding of complex information, acts as an important arbiter of a priori perspectives, serves as a catalyst for critical reflection, and helps enable trust, confidence, and authority in the decision process.
4. Open communication, dialog, and critical reflection are necessary activities within the decision making process. They help ensure participation and engagement across stakeholder groups. Access and reflection on empirical information is the foundation of these activities. Transparency of information builds organizational trust and acceptance.

Research Question 3 (RQ3)

Third, to what degree have the perspectives and concerns of stakeholders been incorporated into, and influence, the system's implementation?

Implementation of long-term and sustainable organizational innovation is partially dependent on the trust and acceptance displayed by effected stakeholders **(**Hargreaves & Fink, 2004, 2006; Kotter, 1996; Rogers, 2003). Theorists posit that stakeholder empowerment in the design, decision, and implementation phases of change are critical to realizing successful transitions and long term viability (Kotter, 1996, 2010, 2011; Hord et al., 1987; Hall & Hord, 2011; Marzano et al., 2005).

In this regard, Research Question 3 examined stakeholder *voice* in the evaluation process. As used here, voice is conceptualized to mean *influence, impact,* and/or *effect*. In this way, decisions are affected such that different outcomes would have occurred in the absence of stakeholder voice. As such, the focus was on understanding the degree to which educators (teachers and administrators at the local and state policy level) were afforded input and influence in the design and implementation of the evaluation system. It is posited that greater educator voice reflects positively on the policy-imposed framework by evidencing greater alignment to instructional attributes held important by educational stakeholders. In contrast, lack of voice raises questions on the system's fidelity to its intended purpose: assessing the instructional competence of classroom teachers.

Educator voice. Stakeholder narratives reflected generally negative perspectives regarding teacher/educator inclusion in evaluation design and decision making. The collective narrative exposed a general lack of educator empowerment, viewing the system as externally imposed and implemented from the top-down: from policy maker (state and district leadership) to implementer (principal evaluators) to recipient (teachers). There is a consensual view that the fundamental legislative framework was derived without

substantive input from professional educators. In this regard, state policy participants made some distinction between the legislative and rule making processes, noting that key educators were involved in the interpretive discussions held by the State Board of Education and the Governor's Office, but not in forming the underlying legislative mandate. Nevertheless, the narrative reflected a general inability of the educator community to substantively influence the core design and implementation intentions of the framework.

Throughout the stakeholder narrative, distinctions were made concerning *opportunity to provide feedback*, *information sharing*, and *inclusion and/or influence* in the decision making process. When developing the analysis, the reflection emerged that these were not equivalent representations of voice. That is, providing input/feedback is not evidence of influence (i.e., being a member of the decision making body; having impact or effect). Similarly, communicating information from one group to another relates more to values of transparency and openness than inclusion in the decision making process. These distinctions were operationalized in the detailed analysis presented in Chapter 4.

District administrators (principal and central office members) acknowledge the numerous opportunities teachers were afforded to provide feedback during the local implementation phase of the activity. However, they believe that teachers generally failed to take advantage of these opportunities. For these individuals, having the opportunity to provide feedback is seen as evidence of teacher voice. Some central office participants equated teacher membership on the evaluation committee as evidence of teacher influence and voice in the decision process. However, there remained a collective

sentiment that teachers at large had little voice in the process, mostly because they failed to provide feedback and input.

As a group, principals viewed the evaluation system as inherently well-constructed and implemented. Thus, for this group, the system did not require substantive modification or change. Thus, the concept of educator voice, especially teacher input, was discounted. They also participated in many information sessions held during regularly scheduled administrator meetings. During these meetings, they received evaluation updates, were invited to dialog, and encouraged to provide feedback to the evaluation design team. Again, this type of information sharing was generally equated to a form of voice.

Teacher narratives reflected a general lack of empowerment and exclusion from the decision making process. For this group, the evaluation activity was externally developed and imposed. Despite some acknowledgment of opportunities to provide feedback/input, they viewed this as information sharing, meant to inform but not involve. As such, teacher participants viewed themselves as recipients of the evaluation process; it was being imposed on them rather than in partnership with them. In this way, teachers more accurately discriminated the concept of *voice*-as-*influence* from the *opportunity to provide feedback* and/or *information sharing*.

RQ3 summary. Overall, the collection of narratives indicate a lack of voice realized by stakeholders at each organizational level. Professional educators had little input to the legislative activity and only peripheral input in developing state-wide implementation guidelines and rules (i.e., as part of the larger collection of policy leaders external to the education community). District policy leaders, acting in response to state

legislative requirements, attempted to provide substantive opportunity for local input from site administrators and classroom teachers. In addition, the district engaged in substantive information sharing, communication with site administrators and staff. However, information sharing, communication, and process transparency do not equate to authentic inclusion and influence in the decision making process. Despite best efforts, teachers primarily viewed themselves as recipients of an externally imposed evaluation activity.

Lack of inclusion and empowerment of educators throughout the successive levels of policy leadership and decision making power suggests the evaluation framework may be disconnected from the professional practice perspectives of those it is designed to assess. In addition, lack of inclusion suggests stakeholder acceptance of the initiative will continue to be tenuous, threatening the long term sustainability of the system outside of the influence exerted by external power centers (i.e., state legislative authority, government mandates and guidelines, and associated requirements for district compliance).

RQ3 closing assertions.

1. The system is generally perceived at each successive stakeholder level as being externally imposed and administered in a top-down fashion with little opportunity for substantive modification.
2. Stakeholder narratives reflected a generally negative perspective on the degree of influence/inclusion educators held at all policy levels in the design, implementation, and/or utilization of the evaluation system.

3. Teachers expressed a fundamental lack of empowerment and exclusion from the decision making process. In this way, they view themselves as recipients of an externally imposed evaluation activity.
4. Local organizational efforts to provide opportunity for input, information sharing, communication, and transparency is differentiated from the authentic empowerment of stakeholders in the decision making process.
5. Long term sustainability of the evaluation initiative will be dependent on power exerted by external power centers including state legislative authority, government mandates and guidelines, and associated requirements for district compliance.

Chapter 6: Discussion

Introduction

This research study examined the construct validity of a state policy-imposed framework for evaluating Teacher Instructional Quality (TIQ). Construct validation attempts to assess "…the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests…" (AERA et al., 2014, p. 11). In this way, the study sought to evaluate the warrant by which inferences of instructional quality are made. To do so, it adopted a contemporary conceptualization of validity, bringing together many independent sources of evidence (AERA et al., 1999, 2014; Messick, 1989a; Kane, 2001; Gorin, 2007; Brualdi, 1999). Collectively, these evidences inform on the authority of inferences derived from the evaluation process as well as consequences realized from such inferences (Messick, 1989a, 1998; Kane, 2001; Kimball & Milanowski, 2009; AERA et al., 1999, 2014; Amrein-Beardsley, 2008, 2009)

Importantly, the study's purpose was not simply to inform local school administrators on the efficacy of the evaluation activity. Rather, it was embedded in a larger state and national policy context of teacher evaluation and education accountability (Ariz. Rev. Stat. §15-203A.38, 2010; USDOE, 2009; ADE, 2012b; Darling-Hammond, 1997; Erpenbach, 2011). That is, local implementation is directed by an overarching policy mandate (Ariz. Rev. Stat. §15-203A.38, 2010; ADE, 2012b). Thus, compliance to this mandate suggests that local experiences become important evidences of the larger policy construct.

In this context, Chapter 6 will discuss the following aspects of the study: Strengths and Significance of the Study, Limitations of the Study, Study Generalizability,

Recommendations for Future Research, and Recommendations for Implementing Teacher Evaluation Systems. In addition, Chapter 6 will present a discussion of primary Research Question 4 (RQ4): How did the process of engaging in this action-research study impact the investigator as a scholarly researcher and organizational leader? This section discusses challenges experienced during the research process and the impact these challenges had in terms of growth as an applied researcher, a scholar, and an organizational leader.

Strength and Significance of the Study

This study presented a compressive approach for evaluating construct validity within a high stakes, consequential, public policy context. It applied a contemporary view of validity as a unified construct, differing from historical conceptualizations (Westen & Rosenthal, 2005; Kane, 2001).

Modern validation study requires assembly of independent forms of evidence to inform on the nature and relationship of measures to the latent construct they purport to represent (Messick, 1989a, AERA et al. 1999, 2014; College Board, n.d.). The research presented herein attempted to adhere to this approach. In doing so it assembled multi-faceted information derived using multiple analytic traditions, data sources, and methods. In this regard, the study brought together reliability, criterion, content, and consequential evidences. In addition, importance was afforded to uncovering representations of the theoretical (teacher quality) construct upon which validation studies depend. It is believed that this type of comprehensive analysis more adequately informs on the policy framework and lends substantive credential to the study's findings, reflections, and recommendations

The specific analytic structure of the study also contributes to the authority of its findings and its relevance to the larger policy context. Here, the analysis focused on a well-defined population of teachers, self-contained general education elementary classrooms with class size of 10 or more students instructing in subject areas directly aligned to academic content measured by state standardized assessments. This population reduces confounding influences of alternative instructional settings and provides a tighter alignment with the conceptual intent of evaluation policy. Finally, this subset of teachers uniformly exercise the pedagogical competencies articulated by the Danielson Framework for Teaching (FFT). This makes inferences about the validity framework more precise and less subject to questions of situational variability.

From this population, a random sample of representative teachers was selected for use in the qualitative portions of the study. Here, sample selection was drawn from teachers located in high, middle, and low achievement (value-added) categories. In part, this was intended to ensure an unbiased representation of qualitative perspectives. Principal selection was purposeful, based on the same instructional location as the random sample of teachers. This structure afforded stronger comparative assertions of group perspectives since principals served as primary evaluators of classroom teachers.

Qualitative construct evidences were obtained from four independent stakeholder groups (n = 22, audio = 882:42 min. /sec., transcripts = 83,162 words), representing a hierarchy of influence and perspectives across the policy domain: teachers (recipients, n = 7), principals (implementers, n = 8), district policy (decision makers, n = 4), and state policy (architects/designers, n = 3). Each represents legitimate participation in the evaluation activity, offering unique insights and perspectives. Their collective reflections

enhance the authority of findings and permit nuanced reflections across the policy construct. Inclusion of state policy participants extends the examination beyond the local.

Finally, the situational context in which the research was conducted provides an important foundation to the significance of findings. The district spent the better part of three years developing the localized approach for operationalizing the state policy framework. Methodologically, the district utilized contemporary multi-level modeling techniques to account for the nested characteristics of achievement measures. In addition, it required all evaluators to annually attain national certification on the Danielson framework via the Teachscape program. Finally, the LEA provided all evaluators with ongoing (monthly) rubric training using authentic instructional situations. Arguably, substantive efforts were taken to ensure the soundness of the system's evaluation approach, structure, and implementation.

The district also developed custom information management systems that afforded every evaluator and teacher access to the complete body of evaluation information evolving in real-time throughout the course of the school year(s). Purposefully built into the evaluation activity were requirements for dialog, communication, and reflection between teachers and evaluators. Many meetings, workshops, and feedback sessions with stakeholders were also held to present and discuss technical and philosophical issues rising from the evaluation design process. Finally, the majority of members on the teacher evaluation design team were classroom teachers, instructional support representatives, and school administrators, and a separate teacher advisory group met repeatedly throughout the design and implementation phases of the process.

During the design and implementation process, committee members and LEA policy leaders attended numerous professional workshops and conferences concerning teacher evaluation, hearing from national experts in the fields of accountability, evaluation, measurement, and policy. In addition, the committee was informed about important insights and issues reported by other researchers through published academic literature. Finally, throughout the three plus years of development, decision makers spent most of their time reviewing and discussing a constant stream of technical data concerning data construction, integrity, reliability, and the computational aspects of the evaluation metrics.

As reported in this study, these efforts did not magically result in a perfect system for evaluating classroom teachers. Yet it cannot be said that the local implementation was haphazard, ill conceived, or without purposeful and considered intent. The time and effort the district invested in developing the best, most reliable system possible within externally imposed policy constraints is not in question. Indeed, the suggestion might be that it represents an investment in cost, time, and staff into design and quality attributes that most other public school districts either are unable or unwilling to undertake. Arguably, most public school districts in the United States simply do not have the resources or expertise to conduct this level of effort or focus. Collectively, these attributes strengthen the foundation on the findings presented herein.

In conclusion, this study addressed an important public policy topic. Under heightened calls for accountability, inferences based on measures of teacher competence are being used to justify consequential decisions impacting the personal and professional lives of teachers (Berliner, 2005, 2014; Amrein-Beardsley & Collins, 2012; Song &

Felch, 2011; Darling-Hammond, 1997; Darling-Hammond, Cook et al., 2012; Butrymowicz & Garland, 2012; Baum, 2010). Questions of validity including evaluation of measure quality and construct representation require empirical investigation using appropriate and comprehensive approaches. As reflected on by Danielson (2010) "… credibility in an evaluation system is essential. A principal or a superintendent must be able to say to the school board and the public, '… everyone who teaches here is good and here's how I know.'" (p. 36). Similarly, the principles outlined within the *Standards for Educational and Psychological Testing* (AERA et al., 1999, 2014), highlight the critical importance of validity evidence in legitimizing and authorizing claims of instructional competence. And Fast and Hebbler (2004) argue that in any type of accountability system "… validation evidence is necessary to support the accountability claims made about individuals and agencies and the accompanying imposition of stakes" (p. 2). It is believed that this study makes an important contribution to this discussion and knowledge base.

Limitations of the Study

It is recognized that the study harbors a number of limitations that may harm generalizability and reliability of findings. These limitations include the following categories: positionality of the researcher; limited stakeholder subgroup representation; restricted teacher representation; single district representation; nested FFT rating data; and unknown interrater reliability. Discussion of each is provided below.

Positionality

As discussed in more detail later in this chapter under the section *Research Question 4: Researcher Positionality*, I served as Director of Research in the study district. In addition, I was a member of the teacher evaluation committee and was

responsible for designing and implementing the computational approach used in the local evaluation process. In that role, I developed the computational perspectives that the committee approved for use in the evaluation process.

This role positioned me as an insider-researcher (England, 1994; Moser, 2008; Dyer, 2009; Moore, 2012; Bonner & Tolhurst, 2002), simultaneously negotiating membership in two communities: the decision making (policy) team and the role of an outside objective investigator (Wenger, 1998; Lave & Wenger, 1991). Throughout the activity, I tried to assume a neutral position: that of an information broker, providing objective data and analysis to policy leaders and decision makers. However, as I reflect later in this chapter, that was a naïve perspective to adopt.

While trying to serve as an information broker, I made decisions on what information to share with decision makers. This included the type of data, detail, and analysis as well as my explanation, interpretation, and "expert" opinion concerning its importance to the decision making process. In this way, my particular perspectives may have shape/influenced specific decisions due to the filters I imposed on the raw data, computations, and choice of presentation. It is difficult to objectively evaluate the impact this may or may not have had on the research presented in this report. As such, it stands as a possible limitation to the study's findings and reflections.

Qualitative: Limited Stakeholder Subgroup Representation

For this study, a total of 23 individuals were interviewed as part of the qualitative investigations. While this might be considered as a "large n" in terms of the collective data, the group was stratified by four levels of representation: teachers (n = 7), principals

(n = 8), district policy makers (n = 4), and state policy makers (n = 3). Thus, the counts for any single group were fairly small, especially for the district and state group.

In addition to the small n-counts, the sampling process differed across participant groups. Importantly, only the teacher group reflected a true random sample of eligible participants. However, even here the actual teacher sample population was reduced from 235 to 94 individuals based on their placement within one of three achievement categories: high (90^{th} percentile), middle (45^{th} to 55^{th} percentile), and low (10^{th} percentile). Thus, 141 teachers situated outside of these three designations were not included in the sampling process.

Importantly, principals were purposefully identified based on the location of the selected teachers. The intent was to provide an alignment between the perspectives expressed by teachers and their respective evaluators and/or administrative supervisor. The rationale was that observing consistency between the two groups would provide stronger evidence of cohesion or division (i.e., by eliminating variance due to school/community environments). However, it may be easily argued that this type of non-random selection actually introduced bias into the analysis by restricting such variation. Thus, generalization from the teacher-principal qualitative information to the rest of the district community may not be justified. Selection of the four district-level participants was also purposeful. However, in this case there were only five administrative positions identified in the population.

Finally, selection of the three state-level participants may be characterized as both purposeful and convenient. At the start of the qualitative portion of the data collection, access to individuals directly involved in formative policy development presented a

challenge. The legislative process occurred approximately two years prior to the start of this study. During that time, membership on legislative education committees had changed as had the status of many legislators.

Knowing that the number of state-level individuals would be limited by time, availability, and willingness to participate, it was decided to target individuals based on two selection criteria: first, persons known to be directly involved in the formative design and implementation stages of the state framework; second, individuals I personally knew through prior professional activities who I felt comfortable in requesting their contribution.

The three state participants possessed substantive knowledge about the framework's legislative and/or rule making activities. However, it could be argued that additional state representation was desirable. Perspectives from legislators representing each political party who were also members of the originating education committees seem warranted. In addition, obtaining perspectives from additional members of the state agency task forces responsible for developing the system's interpretive rules and guidelines also seems reasonable. Regardless, given the number of state policy persons involved in the formulation and implementation of the state's evaluation framework, inclusion of only three voices is arguably restrictive for making claims of representation. Thus, study findings for this participant group may lack representation to the larger complement of state policy leaders.

Restricted Teacher Generalization

This study focused on a restricted population of "Group A" teachers. They were composed of 235 self-contained general education elementary teachers who instructed

content areas assessed by the state's standardized achievement tests (thus, their Group A designation). However, the district employs approximately 1,400 certified pre-school through grade twelve teachers most of which instruct in content areas not assessed by state assessments. Examples include non-self-contained departmentalized middle school classes, special area subjects such as band, physical education, chorus, and Career and Technical Education (CTE). Under the state's evaluation framework, the achievement portion of their evaluation classification defaults to grade, school, or district level measures, depending upon the particular instructional setting.

For this study, non-Group A teachers were purposefully excluded in order to reduce construct-irrelevant variance due to incompatible instructional settings and a lack of correspondence between tested and instructional content. However, it could be argued that this unduly restricts the generalizability of findings since the majority (83%) of teachers are not represented, either in the professional practice (Danielson) data or by exclusion from the teacher (qualitative) sampling activity. Thus, the generalizability of the findings are restricted.

Single District Representation

This study was conducted using data from a single Arizona school district. As discussed in previous sections, it was situated within a middle income residential suburb of metropolitan Phoenix with an enrollment of approximately 25,500 PS-12 students of which approximately 44% receive services under the National Free and Reduced Price Lunch Program. This profile raises questions as to its representativeness of other Arizona school districts, many of which report substantively smaller enrollments, higher proportions of poverty, and are situated in more rural environs. Thus, the generalizability

of findings obtained from this study may be criticized as being unrepresentative at the larger state level. This observation suggests replicating the study in districts with dissimilar demographic characteristics as an area of future research.

Nested FFT Rating Data

One of the characteristics of the professional practice (Danielson) rating data is its nested structure. Discussed to some extent in Chapter 4, teachers are evaluated by their local school administrators. That is, principals and assistant principals serve as the evaluators for all certified staff on the campus. As such, the collection of evaluation scores are "nested" within both school and evaluator: the ratings of a group of teachers on a campus are aligned to either the school's principal or assistant principal, and since no external evaluators are utilized, to the school-community itself.

This creates a substantive methodological problem: all of the within-school and within-evaluator data may be correlated, thus violating distributional assumptions necessary for producing unbiased parameter estimates used in statistical procedures such as analysis of variance and confirmatory/exploratory factor analysis (Muthen, 1994). This study employed these types of analytic techniques without making specific adjustments for the nested nature of the rating data. This suggests the need for additional research efforts that utilize more sophisticated estimation approaches.

Unknown Interrater Reliability

Evaluators assign one of four possible performance ratings to each of the 22 behavioral components identified under the Danielson Framework for Teaching (FFT): Unsatisfactory (zero points), Basic (one point), Proficient (two points), or Distinguished (three points). Early review of the FFT distributional data suggested a substantive lack of

variation in the assignment of these ratings. That is, evaluators were assigning very few Unsatisfactory or Basic ratings. This lack of variation caused concern over evaluators' proper application of the Danielson rubric. As documented earlier in this report, this concern contributed to discussions regarding the need for additional training in the authentic application of the scoring criteria beyond that received during the Teachscape certification process.

Importantly, other than receipt of Teachscape Certification and the extensive amount of on-going scoring training conducted by the district, there was no formal analytic process in place to empirically assess interrater reliability or scoring drift over time. Thus, there was no basis to judge whether the clustering of performance scores into the Proficient and Distinguished categories reflected true instructional competency or were an artifact of evaluator's tendency to score favorably in a real-world consequential setting.

The possible impact of poor adherence to the rating criteria may be substantial. Lack of variation in the data directly impacts the location and distribution of scores used to establish classification cut points while scoring bias impacts the inferential qualities of the empirical data. In this way, questions of reliability or scoring bias undermine the authority of consequential decisions. Regardless, during the course of the study, there was no method to affirm ongoing rater reliability or the impact of external influences on the scoring process.

Discussions related to score distribution, potential bias, and the general lack of rater drift information eventually led the evaluation team to implement a formal study of interrater reliability. This study is currently underway. In its first phase, evaluators

(principals and assistant principals) at each school location are randomly assigned teachers to mutually assess. These teachers are internal to the school, thus rather than being evaluated by one of the two administrators, they are assessed by both. Each internal evaluator completes his/her observation independently. From the data, staff of the Research Office will compute various measures of interrater consistency. In the second phase, an external evaluator design will be implemented, where teachers are observed by persons external to their local campus: receiving one rating from their local administrator and another from an external administrator. Hopefully, these measures will inform on the degree to which evaluators are consistently applying the rubric criteria.

Comment on Generalizability

Most of the comments made regarding study limitations reflect a focus on experimental designs as the de facto standard of excellence. Indeed, the Institute for Education Sciences (IES, 2003) holds that best practices in conducting education research utilize randomized (control-treatment) experimental designs. Their concern is that education research is often characterized by "… poorly-designed and/or advocacy-driven studies…" that harm generalizability of results to broader educational contexts and prevent successful replication (p. 3). Indeed, the No Child Left Behind Act of 2001 advocated for such experimental methods as a needed change to evidence practices and programs that make real and substantive impacts on learning (NCLB, 2002). However, this so called "Gold Standard" aligns directly with quantitative research traditions that emphasize sample representation, measures of precision/error, and replicability (IES, 2003; Morse, 1999; Sparks, 2010). The

implication is that deviations from the *standard* harm the transferability and generalizability of findings.

Arguable, this perspective devalues more qualitative traditions that hold a less restrictive view of knowledge creation (Morse, 2003; Stake, 1978, 2010; Delmar, 2010). Rather, some qualitative theorists argue that evidence as a source of knowledge is inherently contextual, and the experimental criteria for authorizing generalizability are insufficient. Of this, Morse (1999) states emphatically "Of course, qualitative research is generalizable…" but that the "…criterion for determining generalizability … differs from quantitative inquiry…" (p. 5). Similarly, Stake (1978) argues for "naturalistic generalizability," where qualitative research is fully transferable to similar contexts and settings. For each, it is the contextual setting that determines the standard for transference. Further, Delmar (2010) contends that the issue of generalizability is really a question of what counts as knowledge, saying "…we need to accept that context-dependent knowledge, with its different mode of expression, can offer us true understanding…" (p. 121).

In this way, the limitations of this study are not intended to devalue the importance or contribution of its findings, or to suggest they are unduly limited because of the singular location from which the data was collected. Indeed, care was taken to select stakeholder participants from policy groups directly involved in the formulation and implementation of the state-imposed evaluation framework. In addition, districts throughout the state face similar compliance requirements and grapple with the same validity criteria. In this way, it may be argued that the contextual setting for the study makes transference of findings appropriate.

In this regard, Delmar (2010) argues that "The first condition for producing valid research is that the methodology is adapted to the specific object of examination…" (p. 116). Here, the object is the policy-imposed evaluation framework, common to all districts in the state. As long as the methods used are rigorous and properly reflect the contextual setting in which they were conducted, transferability to similar context may be warranted (Stake, 1978; Morse, 1999). Arguable, the methodology employed by this study were rigorous in that they examined construct validity from a multi-faceted perspective employing multiple methods, stakeholders, and data sources. At the same time, the local context equates to a case study (a single district among many across the state) and this has "… an epistemological advantage over other inquiry methods as a basis for naturalistic generalization…" (Stake, 1978, p. 7). Again, the authority for transference becomes the alignment of the contextual setting.

Yeager, Bryke, Muhich, Hausman, and Morales (2013) posit that "…accelerating the field's capacity to learn *in and through* practice is one key to transforming promising ideas in education into tools, interventions, and professional development initiatives that achieve effectiveness reliably at scale…" (p. 1). They argue that traditional forms of measurement and research are not exclusive in their ability to contribute to the generalized body of knowledge and/or understanding. Rather, localized research for the purpose of organizational improvement is a legitimate and important method for knowledge contribution because it situates research in the context of problem solving that may be shared on a larger scale. They state "… we know from numerous sectors, such as industry and health care, that

such inquiries can transform promising change ideas into initiatives that achieve efficacy reliably at scale…" (p. 6). The power of the effort originates from the need to solve localized problems through practical experience that may then be transported externally to similar contexts. The lesson learned through the applied research process increases with each implementation where "…"learning by doing" in individual clinical practice can culminate in robust, practical field knowledge…" (Yeager et. al., 2013, p. 6).

The perspective advanced by Yeager et al. (2013) seems applicable to the validity study presented herein. The localized problem involved the implementation of an evaluation system that met legislative standards and displayed sufficient measure quality to warrant high stakes consequential decision making. In this sense, the LEA was learning by doing, grappling with and examining the best approaches for completing the task. This study attempts to share much of that experience with the larger education community, what Yeager et al. (2013) term as *networks of research and practitioners engaged in improvement research* (p. 39).

Similarly, Bryk, Gomez, and Grunow (2011) argue for educational research that is problem-based and localized, an approach they term as Design-Educational Engineering and Development (DEED). They suggest that localized research should be conducted in the context of *Network Improvement Communities* that serve both a localized context but that transfer to a larger context and application (p. 10). While this study was not designed and conducted by a network of researchers, it was situated internal to an organization engaged in an applied problem-solving context. It is argued that this context is shared by districts

throughout the state, making its findings and experiences relevant to a wider audience of practitioners and policy leaders.

Recommendations for Extending the Organization's Current Evaluation System

The research presented herein suggests the following extensions and modification to the current system:

1. *Increase professional development for administrator-evaluators in the area of teacher mentoring and post-conference support.*

   Comment: District policy leaders acknowledge that the primary purpose of conducting the evaluation activity is to improve the instructional competency of classroom teachers. However, the main focus of evaluator training to-date has been on the structural application of the Danielson Framework for Teaching (FFT), the scoring process. While evaluation metrics provide empirical identification of weak instructional competences, it is the consequential action taken on these measures that align the process to the system's stated purpose. Raised as an issue by some study participants, district policy leaders agree that additional emphasis should be placed on pre-post conference mentoring and support. The recommendation offered herein is to implement this change now that other facets of the evaluation process have been completed.

2. *Increase focus on building evaluator-teacher relationship through communication, dialog, and critical reflection conversations*.

   Comment: This recommendation is based on a primary finding of the study and is tied to Recommendation 1. That is, the importance of the teacher-evaluator relationship is affirmed by stakeholder reflections of the evaluation environment.

The desire is to strengthen the understanding of instructional practice through collaborative reflection and communication. Data suggest that the opportunities afforded by formal/informal observations are insufficient to develop this intimacy. Time (lack thereof) is also a contributing factor. Thus the recommendation is to explore opportunities to expand this important affective component of the evaluation process.

3. *Design and conduct internal research on the efficacy of the evaluation system's ability to actually improve the instructional practice of classroom teachers*.

    Comment: Since the main purpose of the evaluation system is to improve instructional competency, evidence of such impact is critically important for assessing system efficacy. Currently, little empirical evidence exists to support the premise that the evaluation process is substantively changing professional practice. Indeed, findings reported herein suggest some (but not all) stakeholders are pessimistic of the system's ability to effect meaningful improvement. Therefore, the recommendation is to design and implement a formal research design to assess this program goal.

4. *Expand the amount of time evaluators are able to spend with individual teachers.*

    Comment: Issues related to time are a major finding of the study. Specifically, some stakeholders believe evaluators afford insufficient time to gain a full and accurate understanding of instructional practice. This perspective is nuanced in that the time required to evaluate all teachers on a campus is substantial from the evaluator's perspective, but insufficient from the perspective of the individual

teachers. This item is also intertwined with Recommendations 1 and 2 (i.e. mentoring, support, relationship).

Solutions to this issue are difficult because they require changing the current structure of the evaluation process. In its current form, the evaluation process is time consuming for evaluators which reduces the time allocated to individuals. However, findings from the study suggest that not all of the FFT components are viewed equally for identifying and distinguishing good/effective teachers. Indeed, results of factor analytic and content validation portions of the research suggest the same level of information regarding instructional quality might be obtained from a smaller subset of empirical measures, requiring less time to collect/document.

Reducing the number of components formally assessed by the FFT or re-structuring the evaluation approach may provide some flexibility. Regardless, attention and discussion should be afforded solutions that enhance the evaluator-teacher connection, of which time is a critical component.

5. *Develop and utilize additional measures of student affective influences and affective dimensions of professional practice.*

Comment: This is a major finding of the study. Issues of attribute omission and underrepresentation negatively impact the alignment between the reified evaluation framework and conceptualizations of good/effective teaching. In its current form, the framework over emphasizes pedagogical competencies of instructional behavior and fails to afford sufficient attention to personal, emotional, and non-academic influences on students. There is a lack of

representation to indirect curriculums not defined in terms of content attainment. And affective, non-pedagogical attributes of teacher professional identity are lacking.

The recommendation suggests that a broader array of affective influences be identified, measured, and utilized to afford a stronger alignment between the empirical and theorized instructional quality constructs. This includes additional aspects of teacher professional practice that more closely associate with professional identity, commitment, and involvement in local school communities.

6. *Extend the type and number of achievement measures used by the evaluation system.*

   Comment: This is a major finding of the study. Stakeholder reflections suggest that single-event, standardized test scores, are insufficient indicators of instructional efficacy. In addition, the influence of non-instructional factors may harm score reliability and bias inferential judgments of instructional efficacy. State assessments measure a narrow array of content not associated with the majority of teachers including some subjects taught in self-contained, general education classrooms. The recommendation is not to eliminate this metric, but to extend the types of content assessed and utilize multiple forms of measure in addition to mastery-level standardized multiple choice assessments.

7. *Complete the full two-phase Interrater Reliability Study.*

   Comment: The current evaluation context lacks sufficient evidence of interrater reliability. This brings into question data quality issues associated with the professional practice measures. The issue is that these questions remain unexplored. Leadership has started implementation of a same-school reliability

study design. This effort should be extended to include the proposed external evaluator design. Finally, ongoing rater-drift information should become a formal part of the ongoing evaluation process. These data will help ensure the suitability of the professional practice scores for informing on instructional quality.

8. *In its current form, continue to de-value the emphasis placed on single-event standardized test scores as a component of the policy-imposed evaluation system.*

    Comment: The evidence provided herein questions the suitability of single-event standardized test scores as adequate indicators of instructional quality. In this regard, the recommendation is to continue to assign minimum weight to this metric as per state legislative policy.

9. *Continue the current extended timelines associated with identifying* Ineffective *teachers and associated pathways for determining consequences*

    Comment: This is an important outcome of the study. In aggregate, the collection of evidences suggest that the current framework and associated metrics do not display the level of measure quality, content representation, or construct alignment sufficient to warrant use in making high stakes, consequential decisions impacting teacher's personal and profession identity, especially if those impacts include termination of employment.

    Currently, the pathway between initial identification of poor instructional performance and potential termination is two school years. During the first year, teachers are evaluated using ongoing, repeated observational assessments. Teachers receive feedback at each point in this process and become aware of instructional weakness prior to assignment of the first year performance

classification. At the end of the school year, if performance has not improved, an *Ineffective* classification may be assigned and the individual placed onto a formal improvement plan.

From the start of the second year, individuals have opportunity to work on and improve their practice: they may receive additional mentoring and support from colleagues, attend targeted professional development and training. As before, ongoing, progressive observational assessments continue to occur throughout the second year. It is only at the close of the second year that a new formal evaluation classification is assigned. If performance has not improved, further consequential action may occur.

This study brings into question the appropriateness of making short-term consequential decisions based on any single-year evaluation outcome. Measurement and construct issues suggest that longer term and more robust measures of performance are required to arrive at accurate performance inferences. Given the current characteristics of the measures, single-year evaluation outcomes should not be used to make high stakes consequential decisions. Thus, the recommendation is to continue with the current performance timelines utilized by the district. Finally efforts should be resisted to compress these timelines pending empirical evidence of improved measurement quality.

Recommendations for Designing and Implementing Teacher Evaluation Systems

The findings presented in this study lead to the following generalized observations for designing and implementing localized teacher evaluation systems. All of

these are personal reflections from the research process and my role as information broker:

1. *Data Accessibility and Transparency*: Access to empirical information is a critical attribute for successful implementation of new innovations such as a teacher evaluation system. Specific observations in this regard include:
    - Having access to a comprehensive set of supporting data acts to drive discussion, critical thinking, and reflective questioning on the part of designers and decision makers.
    - The type of empirical information made available to decision leaders should range from the most elemental to the aggregated and analytical. Access to elemental data promotes alternative thinking because stakeholders have confidence that their detailed questions may be easily evaluated.
    - Data accessibility acts to challenge tendencies toward political expediency, adherence to policy traditions, and deference to practicality because closely held propositions may be easily evaluated.
    - Data accessibility promotes decision making within a context of supporting evidence. Decisions made contrary to evidence are more open to scrutiny and require verifiable justification.
    - Transparency equates to allowing all designers, decision makers, and stakeholders equal access to the same set of information. Data transparency promotes representational equity, promotes discussion, and trust in the decision process.

- Transparency, data accessibility, and communication are equivalent concepts necessary for the successful implementation of new innovations, especially with regard to high stakes consequential outcomes.

2. *Data Quality*: Empirical information utilized in the decision making process must be of high quality in terms of reliability and validity. Trust in the innovation depends on knowing that policy decisions are informed by unbiased and accurate empirical data. When stakeholders question the quality of the underlying information, trust in the decision making process becomes suspect. Data quality should not be assumed, rather is should be empirically evaluated and presented.
3. *Information Broker:* Information brokers are defined by the specialized skills and knowledge they bring to the decision environment. They act as conduits between decision activities and its empirical foundations. In so doing, information brokers add authority to the decision making process.
4. *The Importance of Evidence*: Inherent in the role of information broker is the concept of evidence as being distinct from data. Evidence is a form of argument requiring standards of data quality, context, meaning, and inference. The information broker navigates transitions from data to information to evidence in service of decision makers.
5. *Stakeholder Voice:* During the design, implementation, and application phases of evaluation design, the voice of recipient stakeholders (i.e., teachers) need to be influential in the decision-making process. When recipients lack voice, it harms trust and acceptance of the innovation, making it less likely that desired outcomes will be realized.

6. *Construct Articulation*: Decision making lacks authority in the absence of a well-defined context that all stakeholders can understand. Implementation of a teacher evaluation system that lacks clear articulation of what, how, and why the activity is being undertaken will create disconnect between stakeholders. As such, construct articulation should be considered a foundational requirement for successful implementation of any evaluation system. It facilitates connection between what is being measured and how it is being measured.

Recommendations for Future Evaluation Research

As a result of conducting this study, the following recommendations for future research on teacher evaluation systems and related policy environs are offered:

1. Replicate this study in districts with different student-teacher demographics, location, and size.

   Comment: The study data is situational and embedded in the context of local demographics and community characteristics. As such, generalizability to broader contexts may be questioned. This type of study utilizing a unified conceptualization of construct validity, examined through application of single-phase mixed-method designs should be replicated in other situational settings.

2. Re-compute selected statistical procedures using robust estimators that account for the multi-level, nested characteristics of the professional practice rating data.

   Comment: Methodological issues arise from the nested characteristics of the professional practice rating data. Confirmatory and exploratory factor analytic techniques utilizing estimators that account for the nested structure need to be re-

estimated. Findings should be compared to the more traditional techniques used in this study and any adjustments to conclusions noted.

3. Extend the construct research by developing highly articulated, independent definitions of good/effective teaching from multiple stakeholder perspectives based on larger participation counts

    Comment: This study attempted to explore and differentiate conceptualizations of good/effective teaching across multiple stakeholder groups including teachers, principals, district policy and state policy representatives. However, the participant counts in each of these groups were generally small. In addition, the need to examine multiple dimensions of the evaluation framework limited the focus afforded specifically to construct articulation.

    The recommendation is to more comprehensively examine the theoretical conceptualizations of good/effective teaching held by a wider variety of stakeholder groups. Added to the list would be business, community, parent, and student representatives as well as an expansion of the district and state policy/legislative groups. This research would provide a contemporary and robust specification of the teacher instructional quality construct that could be compared to reified policy-imposed evaluation frameworks.

4. Empirically assess changes in professional practice attributable to teacher evaluation activities

    Comment: The policy-imposed teacher evaluation system posits that the act of evaluation improves the instructional practices of classroom teachers. In a larger framework, possibly utilizing national education databases, design and implement

research designs specifically to evaluate this premise. Results would either support or challenge the efficacy of the policy assertions.

5. Explore the issue of policy-imposed evaluation systems from the perspective of organizational change theory

    Comment: Organizational change theory is usually explored in terms of new initiatives imposed by localized leadership (i.e., companies, organizations, agencies, etc.). Contemporary teacher evaluation systems have their origin in external power centers located at state and national policy levels. Localized implementation of externally-imposed innovations presents a unique situational context from which to explore the efficacy and sustainability of evaluation systems.

6. Triangulate findings from this study with results reported by the district's internal program evaluation study

    Comment: As a separate activity, the district conducted a system-wide survey of stakeholder perspectives regarding the evaluation process. To do so, a questionnaire was distributed to all certified staff and school administrators. It was administered after the data collection used in this research study was completed. As such, it serves as a potential source of additional triangulation and validation.

7. Formulate an experimental design to examine the impact that access to empirical information has in the decision making process

    Comment: One of the primary questions evaluated in the study was the role that access to empirical information had on the decision making process. The

analytic approach was qualitative and non-experimental, relying on stakeholder perspectives to inform on the question.

It is possible to re-specify a more experimental methodological approach to this question. Here, a treatment-control group design could be conceptualized. That is, a sample of stakeholders (teachers, administrators, and policy leaders) might be randomly assigned to one of two groups. Each group would be asked to devise a functional framework for evaluating teacher instructional quality including the types of metrics utilized. The treatment group would be provided access to published literature on designing teacher evaluation systems, problems associated with various metrics and methods. In addition the treatment group would be provided simulation data and related analysis of evaluation data as well as the services of a specialized information broker. In this context, comparison of purported evaluation systems including associated justifications might be compared and contrasted.

8. Value-added (VAM) Model Research: Substitute alternative measures of student achievement and evaluate effect on performance classifications.

    Comment: Using the same professional practice scores, apply alternative academic growth measures and evaluate the stability of teacher performance classifications. At the time of writing, the state's Student Growth Percentile Model provided data for the same two years of VAM estimates used in this study. By substituting the state growth measures for the VAM-based information, the stability of the performance classifications might be assessed.

9. Deconstruction of evaluation outcomes by teacher, student, and school characteristics.

Comment: Further research should be conducted exploring the correlation of teacher evaluation outcomes by secondary community, school, and teacher characteristics. Distributions of performance classifications by community poverty, Title 1 status, teacher experience, administrative leadership, etc. might be examined to see if factors external to instructional efficacy are influencing the outcomes. Similar independent explorations might be conducted for both the VAM and PP measures.

Research Question 4: Personal Reflections

Introduction

Research Question 4 represents a reflective commentary regarding the research process and its effect on personal identity and perspective. The primary research question was formulated as:

RQ4: How did the process of engaging in this action-research study impact the investigator as a scholarly researcher and organizational leader?

At the outset of the study, two supporting research questions were posited in order to help organize the reflections into manageable sections. These supporting questions were formulated as:

RQ4 (a): What barriers or impediments were encountered during the course of the study? How were they overcome and/or handled?

RQ4 (b): What were the salient lessons learned from this study? How did the researcher grow professionally and personally? How will these learnings be incorporated into future research & leadership activities?

The data sources employed to address this question originate from research journals (166 handwritten pages), online blogging (thirty posts, 6,949 words), and additional foundational reflections (3,324 words) recorded into a supporting Word document. Journal and blog posts were sequentially recorded primarily during the implementation and the data analysis phases of the study (September 2012 through summer 2014). Reflective (Word document) annotations were assembled mostly during the writing phase of the study (Spring 2014 to January 2015). Throughout the discussion, the researcher's current sentiments are integrated along with the journal entries and notations.

Construct Overview

An outline of component constructs relevant to RQ4 is provided below, organized by the two supporting research questions: RQ4 (a) and RQ4 (b).

RQ4 (a) - Barriers/Impediments: What barriers or impediments were encountered during the course of the study? How were they overcome and/or handled?

Discussions relevant to RQ4 (a) are organized into five topic areas: *TEval: Technical Issues/Problems, TEval: Complexity and Communication, TEval: Power & Influence/Ineffective Classifications, Researcher: Positionality,* and *Researcher: Complexity of Study*. For each, reflections on the barriers, impediments, and solutions are addressed. The five discussion topics under RQ4 (a) are outlined in Table 66.

Table 66

Outline of Discussion Topics Addressed Under RQ4 (a)

Component	Description
1. TEval: Technical Issues/Problems	• Data Integrity/Problematic Data Sets (2011 Course Schedule) • Data Design (Student-Teacher assignment) • Problematic Teacher Assignments (Extended Leave, Team Teaching, Mixed Assignment)
2. TEval: Communication and Complexity	• VAM Complexity • Stakeholder Communication & Understanding
3. TEval: Power & Influence/Ineffective Classifications	• Concern of Ineffective Classifications o Distributions, Definition o Trust in Process (Policy Level) o Impact: Public Perspectives o Impact: Organizational (Morale, Retention, Recruitment)
4. Researcher: Positionality	• Role of Researcher in Decision Making Process (Information Broker)
5. Researcher: Complexity of Study	• Data Organization, Global Constructs, Data Management, Technology Tools

RQ4 (b) – Salient Lessons Learned: What were the salient lessons learned from this study? How did the researcher grow professionally and personally? How will these learnings be incorporated into future research & leadership activities?

Discussions relevant to RQ4 (b) are organized into six topic areas: *Reflections on Analytic Skill Sets, Reflections on Qualitative Analysis, Reflections on the Writing Process, Transition to Scholarly Writing, Personal Significance,* and *Next Steps*. For

each, reflections on personal importance and meaning are discussed. The six discussion topics under RQ4 (b) are outlined in Table 67.

Table 67

Outline of Discussion Topics Addressed Under RQ4 (b)

Component	Description
1. Reflections on Analytic Skill Sets	Technical Knowledge, Skills, and Abilities Necessary to Complete the Study
2. Reflections on Qualitative Analysis	Transition Between Quantitative and Qualitative Analytic Perspectives
3. Reflections on the Writing Process	Expansion of Writing Ability; Link Between Writing and Thinking
4. Transition to Scholarly Writing	Growth as a Researcher
5. Personal Significance	Personal Reflections; Personal Meaning
6. Next Steps	Future Research; New Leadership, Personal Goals

Researcher Narrative

RQ4 (a) - Barriers/Impediments: What barriers or impediments were encountered during the course of the study? How were they overcome and/or handled?

TEval: Technical issues/problems – 2011 course schedule. During the main design phase of the evaluation system (spring through fall, 2013), the evaluation committee decided to utilize three years of value-added (VAM) growth measures. This decision was consistent with the committee's review of published literature, discussions

regarding problematic issues on the use of VAMS in high stakes decision making, and with the information district policy leaders received while attending numerous professional and state-level meetings and conferences.

One of the issues discussed concerned incorporating multiple years of achievement data into the evaluation calculations in order to improve the stability of the aggregated indicators. Both the committee and the district Superintendent believed this was an important design feature. Indeed, the Superintendent expressed an expectation that this would occur. My journal notes regarding this decision were:

> [May 2013] Multi-year VAM scores: This was a good decision. Had numerous discussions about VAM issues. Literature shows stability issues. Supt feels strongly that we need to use three years data, after coming back (I think?) from ASA or ADE meetings. The process will be much more complex. Need to do this year-by-year. Need archived IC [Infinite Campus – student information system] files back to 2011. Using a single year is already complex. This makes it more so.

The activity required assembling three years of achievement data and linking each year's student information to the appropriate course and teacher. This was accomplished by matching the test scores to the district's internal course schedule for each of three years using teacher and student identifiers. Construction of the final statistical models took place during fall 2013 (i.e., SY2013-14), utilizing testing data from 2011, 2012, and 2013. This necessitated accessing course scheduling files for school years 2010-11, 2011-12, and 2012-13.

During the process of verifying student-teacher-course alignments, I suspected that the achieved SY2010-11 course schedule data was inaccurate. Discussions with staff in the Information Technology (IT) department revealed that the historical information could no longer be reconstructed from original files. Thus, there was no way to validate

that the existing file data was accurate. Therefore, a decision needed to be made whether or not to utilize the information from SY2010-11 in construction of the pilot evaluation ratings, reducing the annual data points from three to two.

A blog entry made in September 2013 reflected on this decision point:

> [September 2013] After significant efforts to ensure data integrity across years of teacher-course data, the 2011 database became problematic. I cannot verify the information archived from the on-line student information system for that year with the raw data from the underlying data tables. For this reason, I approached the Director of Human Resources [HR] and the Superintendent and indicated that I cannot use three years of data for the pilot year development (i.e. 2011, 2012, & 2013). I indicated that I needed to use only data that has integrity in the underlying data structures/values and that I could ensure was properly linked to teaching staff.
>
> They agreed with my recommendation to use the last two years for the pilot and then add the third year in Summer 2014 - when the TEval system becomes 'Live'. Therefore, I am now completing the initial analysis using two years (2012 & 2013) data.

The decision went against the superintendent, evaluation committee, and my belief that three years of data was needed to improve the accuracy of teacher ratings. Even though the pilot results were intended to be non-consequential, releasing it to teachers would certainly impact professional identity. On the down side, using less than complete data might harm teacher trust in the system since the VAM portion might be less stable. It would also make it more difficult to restore that trust in subsequent iterations. Regardless, internal data issues compelled me to recommend removal of the 2011 data point.

Fortunately, policy leaders agreed that suspect data should not be used for the pilot project. It then became a communication issue with stakeholders to clarify that two years of VAM data would be used in the pilot run of the system, but that three years

information would be utilized in the "live" analysis scheduled to be released in fall, 2014. The change to the pilot VAM components were communicated in follow up meetings with administrators, subsequent site-based meetings with certified staff, and in related documentation. Interestingly, we did not receive any feedback suggesting that this change was inappropriate. Perhaps this was because everyone understood that the "live" 2014 results would incorporate the full three-year composite VAM scores, or that the non-consequential nature of the 2013 pilot results lowered stakeholder concern/interest.

TEval: Technical issues/problems – student-teacher assignments. The process of connecting students with courses and teachers is complex. This complexity is made more difficult if the underlying historical data files do not display high levels of field integrity. That is, in order to connect the achievement (VAM) scores of students to the proper teacher, there needs to be some mechanism to array and connect the complete mix of courses, instructional location(s), and enrolled students. Errors in these alignments result in improper attribution of achievement (VAM) measures, biasing the final evaluation ratings.

During the implementation phase of the pilot evaluation, selected key data fields in the district's course and employment files were utilized to identify and differentiate teacher's instructional context (i.e., job title, location, subjects, course names, etc.). From a practical perspective, it becomes difficult to properly identify the instructional characteristics of teachers if the values entered into these fields are incorrect and/or incomplete.

From the start of the process, integrity issues were revealed for a substantive number of teacher records. The initial result was to misclassify teachers in terms of the

subjects taught, location assignment, and student attribution. In some cases, job descriptors (i.e. Teacher Grade 7 Science) failed to reflect multi-grade assignments (i.e., instructing Grade 7 and Grade 8 students) or multi-subject assignments (i.e., teaching sections of Science and Social Studies). In addition, some location identifiers (i.e., school name) were incorrect or incomplete (i.e., persons assigned to multiple locations or the current location was incorrect).

To correct these types of inaccuracies, records for approximately 1,500 certified teachers were manually cross-referenced against aggregated listings of course names, school names, grade levels, and student counts in order to reveal problematic entries. Incorrect/incomplete field representations were modified resulting in a new standardized teacher-attribute database. In some cases, incompatible field representations exiting between course schedule and employee databases also needed to be investigated and resolved.

These integrity checks were completed independently for each of the three data years required by the evaluation analysis. The activity took months of staff time and brought to light data integrity issues previously unknown to district technical staff. However, the effort improved the integrity of the data and highlighted integrity issues to staff in the research, IT, and human resources departments. Maintaining an audit trail has enabled staff to check, verify, and where necessary correct evaluation values on a case-by-case basis, enhancing trust and credibility in the evaluation procedures.

TEval: Technical issues/problems – problematic teacher assignments. For a variety of reasons, each year a substantive number of teachers change their instructional assignments and/or take extended leaves of absence. For example, some teachers teach

for only a portion of the school year as a consequence of medical or personal considerations (i.e., maternity leave). In addition, subsets of classrooms are characterized by mixed/multiple instructors. These contexts are not represented as an attribute in the teacher-course databases and require investigation on a case-by-case basis. Importantly, these conditions also raise specific policy issues regarding how and when to attribute achievement (VAM) scores to specific teachers.

More than a few teachers take leave of absences that require the assignment of guest or long-term substitute teachers to the effected classrooms. Since the historical student-course database is archived at the end of each calendar year, the teacher-course data files record the teacher-of-record present at the close of the school year. That is, these databases are not transactional and do not reflect changes occurring throughout the school year. This raises a number of questions such as (1) how long does an individual need to be the teacher-of-record to qualify for achievement score attribution; (2) what criteria are applied when multiple teachers need to be assigned concurrent achievement measures; and (3) do co-teaching classrooms mean that both teachers-of-record receive the same end-of-year achievement scores? Permutations to these issues reoccurred in each analysis year. Each raised a number of policy and technical issues that needed review and resolution by policy-level decision makers (human resources department, evaluation committee members, and district administration). Uncovering and openly discussing these issues ensured that solutions were arrived at transparently with the consensus of multiple policy leaders and decision makers.

TEval: Communication and complexity. This construct refers to the difficulty with communicating the complexity of the evaluation system's computational structure

and methods to stakeholders, primarily classroom teachers and principals. From the outset of the planning process, members of the evaluation committee and I felt it important to clearly communicate how the evaluation results were constructed as well as the rationale for making computational decisions. Indeed, everyone involved agreed that transparency and communication were essential for building trust in the system.

Early in the development process, I brought up some general findings from selected published research regarding the roles transparency and understanding have on building trust in high stakes teacher evaluation systems (i.e., Amrein-Beardsley & Collins, 2012). I suggested that the published literature indicates teachers lack a basic understanding of the structure and purpose of VAM models in determining final performance ratings. This suggests the need to provide substantive professional development and a clear understanding of how performance ratings are derived.

To address this and their commitment to transparency in general, committee members took extensive steps to ensure ongoing communication with teachers and administrators. Some of these activities included:

- Held 24 site-specific staff meetings at the start of the development process to communicate representation on evaluation committee, review aspects of the legislatively mandated framework, highlight local philosophical and computational challenges, and present the committee's current thinking regarding computational design and methods.
- Implemented an online system for gathering stakeholder feedback, questions, and comments. As a follow-up, the committee responded to all questions and

comments in the form of an FAQ posted to the district's evaluation web site and in subsequent school/regional meetings.

- Held four regional feeder-school meetings detailing the computational structure of the proposed VAM models and how achievement components are aggregated with professional practice ratings.
- Convened a teacher advisory committee composed of two representatives from each campus. The purpose was to discuss aspects of VAM models, computational methods, determination of performance criteria (i.e., cut scores), and to review contents of the proposed evaluation reports. The hope was that these representatives would serve as ambassadors to each campus to ensure all teachers had a clear understanding of the system's technical aspects and related decision frameworks. In addition, these individuals provided ongoing input to the evolving design of the evaluation system.
- On multiple occasions presented technical design details to school administrators so they would have a deep understanding of analytic methods being considered. In addition, administrators were presented with the many philosophical issues being confronted by evaluation committee. The hope was that these site administrators would also serve as a central source for teacher questions and feedback.
- Recorded and posted two videos describing the evaluation process, methods, and computations. These videos were required viewing by all staff as part of the district's online professional development environment.

- Conducted two web-streamed presentations on the computational details used in the evaluation framework.

As mentioned, the purpose of these activities was to provide transparent and complete information to teachers and administrators in an effort to ensure understanding and trust in the system. A journal entry from one early information session is provided below:

> [Oct 2012] TEval Information Workshop: Presented computational details to group of teachers and IGTs (Instructional Coaches) at [location name]. Most had no understanding of VAM models or methods. Confused VAM with OYG [One Year of Growth] used in previous RIFF [Reduction in Force] and Prop 301 formulas. I went through the student covariates and they seemed to understand and appreciate the effort to control for student background factors. Note: The view was clearly that this [evaluation] is being imposed on them. They questioned aspects but didn't seem to be empowered to do anything about it.

This session was held after completion of the initial 24 site-based staff meetings and near the start of follow-up regional feeder school meetings. It became clear that attendees had little understanding of the VAM modeling approach. My journal entry below was made five months later after additional regional meetings had been conducted:

> [3.1.2013] Teacher Evaluation Committee Meeting: Discussed that attendance at feeder school TEval communication session was not well attended but there were some good questions. Seems like not many teachers or administrators care very much. Administrators are in ongoing rubric training sessions and have heard my presentation on value-added and computations a number of times, but teachers have chosen not to attend these sessions. Is this because we did [last year] a presentation at each school site? But this didn't have any detail. Or is it that everyone fully understands? I think not.

The sentiment is that teachers either lacked interest in the evolving designs of the system, felt powerless to effect any changes, or were satisfied with the information already received. I reflect that the latter conclusion is unlikely. Attendance at regional

meetings was voluntary and poorly attended by both teachers and principals despite repeated emphasis at previous administrator meetings of the evaluation's importance.

It was not until the following spring (April, 2013) that the pilot, non-consequential evaluation results using two years of achievement data were released. At that point, many more meetings and communications concerning technical details and methods had transpired. As a follow-up, the teacher advisory committee was reconvened in August 2014 to discuss feedback they received from colleagues regarding the pilot information (i.e., reactions, concerns, comments, etc.). It was hoped that this feedback would help finalize preparations for the release of the first 'live' performance evaluations scheduled for September 2014. The journal entry made for that meeting was:

> [8.21.2014] August 2014 Teacher Site Representative Meeting: All 24 schools; 2 teachers per school. Asked for feedback from the spring 2014 release of the pilot data and the room went silent. It was as if no one had any feedback one way or another. It was not that they were shy. It was more that they didn't have any feedback to give.
>
> It took a lot to get them to discuss. One of the general attitudes was that it was no big deal. [Name] and I were taken off-guard by this response. Teachers didn't seem to care - it wasn't real, or they didn't feel it impacted them very much (because it was just a pilot and had no impact on them professionally?). They had already completed their 1:1 discussions regarding the FFT with their principals. The pilot report didn't really change anything.

The reaction by the teacher representatives was perplexing. The realization began to sink in that, despite all of the communications and attempts at transparency, teachers may not value the information because they did not see any immediate personal impact from the pilot evaluation activity. The reactions lead to the following personal reflection (journal entry):

> [8.21.2014] Was the entire activity more important to us than to them? Did we have a lens of 'importance' while those it was about didn't? Will this change with the release of the 2014 information - which is 'live' and consequential?
>
> Maybe we all overvalued the project because it was part of us. We created it. We invested ourselves in it. So therefore, it should have been important to everyone? That may be the fallacy - if it's important to us, it's important to everyone else?
>
> Some teachers seem to indicate that many staff didn't even LOOK at the report! Didn't discuss it with colleagues, didn't react to it.
>
> Maybe it's because we did SUCH a good job of information sharing, teachers knew exactly what the pilot was about and were COMPLETELY satisfied with the district's efforts and the classifications? Yeah, right. NOT!

The personal reflection questioned the impact of the efforts made to communicate the complexity, detail, and purpose of the evaluation framework. However, the response seemed to suggest that, because the pilot data was not personally consequential, many did not see it as a priority concern. My reflection was really about the assumption placed on the importance of transparency in building stakeholder trust. While it brought comfort to me in my role as researcher and policy advisor, the reaction seemed to suggest it was not as important to those affected, at least not at this point in system development.

Another attempt to examine whether stakeholders understood the complexity and structure of the system took the form of a stakeholder survey administered to all administrators and certified staff in April through May 2014. The survey was conducted just after release of the pilot evaluation results, so teachers had opportunity to review their performance classifications and discuss the results with their principals. The survey was conducted as part of a formal program evaluation conducted by the research staff. In the survey, teachers were asked to indicate their level of comfort in explaining the evaluation components to lay-persons. The wording of the survey items was:

Please indicate your level of agreement with the following statements:

I feel very comfortable explaining to a non-educator how my...

- Professional practice (Danielson) score is calculated.
- Value-added (student growth) score is calculated.
- Overall Effectiveness Classification (i.e., Highly Effective, Effective, Developing, Ineffective) is determined.

A four-point Likert response scale was utilized: *Strongly Agree, Agree, Disagree, and Strongly Disagree*. A comparison between teacher and administrator responses is provided in Table 68.

Table 68

Stakeholder Comfort in Explaining Evaluation Components

Comfort Category	Teacher Positive (Agree + Strongly Agree)	Principal Positive (Agree + Strongly Agree)
[How my] Professional practice (Danielson) score is calculated	62% (*n*=799)	88% (*n*=40)
[How my] Value-added (student growth) score is calculated	41% (*n*=793)	71% (*n*=40)
[How my] Overall Effectiveness Classification is determined	57% (*n*=797)	93% (*n*=40)

Survey results suggest that approximately 60% of teachers did not fully understand how the value-added component was computed. It might be reasonable to assume they did not fully understand how the calculation might/might not address questions of fairness, bias, or equity across different classrooms or instructional settings.

In addition, 38% of teachers did not even feel comfortable explaining how their professional practice (Danielson) score was computed and 43% could not explain how their overall performance classification was derived. These responses contrast substantively with those of principals, who uniformly report higher proportions of computational comfort.

The policy challenge for the organization was, in part, to convey a level of understanding in the computational details that lead to a teacher's consequential performance classification. The evidence suggests that despite all the efforts to communicate this understanding, large proportions of teachers did not fully understand how their evaluations are constructed. The larger question is how these findings impact trust in the system overall and how a lack of understanding impacts staff morale and professional identity?

TEval: Power and influence/ineffective classifications. This construct concerns the influence of power centers on the development of the evaluation system. From the outset of the study, my role was to provide policy makers empirical information from which to make informed decisions. The presumption was that such information facilitates shifts in perspectives as a response to objective evidence. However, this presumption was challenged with regard to the criteria used to classify teachers as *Ineffective*.

By early December 2013, the first pilot evaluation results were being prepared for review and release to teachers across the district. These results were the culmination of approximately two years of focused development by the evaluation committee. This included extensive review of the empirical information, collegial discussion and reflection, leading to the establishment of the proposed performance classification criteria

(i.e., setting of cut scores to segment teachers into four performance classification categories: *Ineffective, Developing, Effective,* and *Highly Effective*).

Approximately mid-December, 2013, the initial distribution of teacher classifications was presented to the evaluation committee and selected central office policy leaders. The purpose was to review and affirm pilot year results prior to public release. The information included tabulations by classification category, teacher listings (with names), and site locations. The data revealed that approximately 11% of teachers would fall into the *Ineffective* classification due to low placement along the composite (VAM plus PP) performance scale. In addition, there seemed to be larger concentrations of lower performing teachers at certain school locations. Finally, the classifications for selected teachers raised concern for some policy leaders.

The journal entry (below) records this researcher's reaction to the initial reaction received from policy leaders:

> [Reader's Note: This portion of the report contained personally identifiable information and has been removed prior to publication]

It is important to note that at no time did any policy leader direct me or the committee to change, alter, or otherwise adjust, the final outcomes. All of the discussions focused on ensuring valid representations of instructional quality. Regardless, the concerns raised were multifaceted. There was a concern that the substantive number of teachers classified as *Ineffective* might harm organizational morale and/or negatively affect the district's public persona. In turn, this would make it more difficult to retain/recruit qualified teachers.

[Reader's Note: This portion of the report contained personally identifiable information and has been removed prior to publication]

Here, reactions by some policy leaders exposed concern over the larger public perceptions of the district. This perspective is being driven, in part, by the lack of commonality between district systems. That is, policy leaders recognize the possibility that adjacent districts may rate teachers more leniently in order to minimize the number of teachers designated as *Ineffective*. As a consequence, districts that impose higher standards of integrity are at a disadvantage. Indeed, the consequence of reporting large numbers of *Ineffective* teachers might be considerable with impacts on staffing, employee turnover, student membership, funding, passage of voter-approved overrides and bond initiatives, etc. These concerns make it paramount that teacher evaluation ratings are accurate and defendable in a larger context.

[Reader's Note: This portion of the report contained personally identifiable information and has been removed prior to publication]

As noted, the eventual policy decision, supported by the evaluation committee, was to delay release of the pilot information until the issues had been thoroughly discussed. This decision was important because it reflected the sentiment that the data itself was not faulty. Rather, the problem was the lack of an operational definition for the term *Ineffective*. I believed this was the correct action since I held the perspective that data lacks meaning without some type of associated construct definition. Essentially, do

not simply adjust data to fit the perception, but articulate the latent construct and then do one's best to measure it.

Indeed, the context of this entire discussion was driven by the absence of a state-provided definition of Ineffective. Up to this point, the Ineffective classification was simply based on a teacher's relative placement along the composite performance scale (i.e., the weighted composite of both VAM and PP ratings). The direct effect of the policy concern was to seek clarification and operational definition of what it meant to be an Ineffective teacher: that is, what traits are consistent with the meaning of the term Ineffective? In turn, this led to a review of the behavioral components contained within the Danielson framework.

Resolution: At the conclusion of the discussion series, evaluation committee members adopted a definition of Ineffective that closely matched the criteria already used by the district to place teachers onto a formal Improvement Plan (i.e., an Improvement Plan is an official notice from the district for the need to improve instructional ability in specific Danielson components or risk further consequence inclusive of termination). The criteria for being placed on a formal improvement plan were as follows: Probationary teachers (certified staff employed in the district for three or fewer years) who receive Unsatisfactory ratings on four or more of the 22 Danielson components; Continuing teachers (employed four or more years) who receive one or more Unsatisfactory ratings. The Committee felt this brought a clear definition to the meaning of *Ineffective*, one that had been previously used and also met the intent of the evaluation's policy framework. In doing so, the district defined the term Ineffective solely on the basis of instructional behavior exclusive of achievement measures.

A secondary effect of the Ineffective discussion was to provide additional time to search for anomalies in the data and verify that the empirical information for selected teachers. As discussed earlier, the job assignments of some teachers were unique making it difficult to align the proper achievement metrics. The delay in releasing the pilot results allowed more time to uncover and resolve some of the case-by-case issues.

Researcher: Positionality. Another challenge I faced concerned my positionality as both researcher and member of the administrator-level decision team. In this context, I served as both system designer and decision maker, raising the possibility of bias leading to an inaccurate representation of findings and study conclusions (Wenger, 1998; Lave & Wenger, 1991; England, 1994; Moser, 2008; Dyer, 2009; Moore, 2012; Bonner & Tolhurst, 2002).

In reflecting on this, England (1994) cites two concerns related to researcher positionality. First, personal experience impacts the research activity because "… different personal characteristics … allow for certain insights, and as a consequence some researchers grasp some phenomenon more easily and better than others…" (p. 85). Thus, personal context forms a boundary through which data is filtered and interpreted. Second, special problems "… derive from the nature of the power relationships in the research encounter" (p. 85) Here, England comments that "… field work is inherently confrontational in that it is the purposeful disruption of other people's lives…" (p. 85). He argues that nature of this power relationship is an important arbiter of the analysis process, influencing both inferential perspectives and the information offered by study participants. Collectively, the responsibility of the researcher is to understand and

acknowledge these factors and then account for their potential impact on claims of objectivity in the research activity.

For me, I was aware that I was occupying dual roles. My initial reaction was to view myself as an external researcher in an effort to objectively conduct the research activity. However, as will be discussed below, this was a naive perspective. In this regard, Moser (2008) discusses the issue of positionality, saying "… it was long believed that the ideal scholar needed to strive for absolute neutrality so as not to 'taint' the research with his or her individuality…" (p. 384). However, "… social scientists have grown increasingly suspicious of the possibility of such claims to objectivity and neutrality…" (p. 384). Moser (2008) argues that researcher personality is a threat to these claims and that "… we never shed our identities or biographies to become neutral observers…" (p. 384). Similarly, Dyer (2009) concludes that "… the personhood of the researcher, including her or his membership status in relation to those participating in the research, is an essential and ever-present aspect of the investigation…" (p. 55). Thus, my attempt to see myself as an outsider-researcher was most likely not consistent with reality.

During the research process, it was difficult for me to see how this dual positionality of external researcher and internal administrator impacted my study. I could not step outside myself and critically evaluate whether my actions were affected by the power centers around me, or how I might influence the information I was receiving from study participants. Regardless of my desire to maintain objectivity, it was impossible to remove myself from an insider position because I was the Director of Research in the study district.

Bonner and Tolhurst (2002) discuss the problem of being an insider-researcher in her own qualitative research in the nursing field saying:

> …there was the risk that over-familiarization with the setting might lead me to make assumptions about what I was observing without necessarily seeking clarification for the rational underpinning particular actions… (p. 10)

In my case, as an administrator in the organization, I certainly risked over-familiarity with the evaluation context, the policy environment, and how the evaluation process was being implemented. Conceivably, this could have impacted my analysis of participant interviews, the discussions I had with colleagues, the questions and probes I explored, and whether or not I imposed my own perceptions over the perceptions shared by others.

In contrast to Bonner and Tolhurst (2002), Dyer (2009) sees insider-outsider dichotomy as more of a distinction of perspective. In his view, neither "… makes me a better or worse researcher; it just makes me a different type of researcher…" The author argues that each has its own benefits and drawbacks regarding threats to study validity. Essentially, it is the responsibility of any researcher to be cognizant his/her position and protect from letting issues of positionality distort or compromise the validity of claims made from the experience.

I tried to address the issue of positionality in two ways. First, I adopted the persona of information broker, perceiving that my primary role was to provide empirical information to decision makers and to do so without filtering the information, albeit, as best as I could. This perspective was partially influenced by my readings of Wenger's conceptualization of boundary brokers and the exchange/negotiation of knowledge between social groups (Wenger, 1998). For me, the intent was to work in a service

capacity to the evaluation committee and decision makers, providing empirical information, facilitating discussion, and enabling understanding.

The second strategy which I believe helped mitigate issues of positionality involved using various forms of data triangulation within my research design. Here, participant review of interview transcripts, sharing/discussing analysis with colleagues, use of secondary researchers as a check on my analysis, and the presentation of findings to stakeholders hopefully lessened the possibility of injecting positionality error into the study.

Researcher: Information broker. During the period in which policy leaders were reviewing the first set of evaluation results, I made the following journal entry (below) reflecting on the evaluation committee's reaction to the initial pilot evaluation results:

> [12.16.2013] [Name] reviewed first set of TEval reports and immediately reacted to the [uneven] distribution of results across schools. [Different Name] made similar observations and reflected on needed changes to evaluator training in terms of using the rubrics – expressed concern that some principals are not scoring properly because of high concentrations of high/low ratings on particular campus.
>
> The data displays sparked thinking on rater training, accuracy, and consistency. This seemed to shape leadership thinking on the evaluation measures. Objective data influences thinking, influences policy, raises questions, challenges beliefs. Information broker: role as researchers and policy influence.

Up to this point, the evaluation committee did not have a complete representation of the rating and performance classifications across the district. Here, I am recording what I believe to be the influence data is having on policy thinking. The uneven distribution of evaluation scores led some policy leaders to question scoring reliability and the consistent application of rubric criteria across the population of evaluators.

A few days later, I made the following related journal entry (below). The context pertains to the reaction additional policy leaders had on the information, believing it reflected substantive problems with rater reliability and accuracy. Reflecting on my personal perspective:

> [12.20.2013] Organizational Leadership/TEval Classifications: Can't help reflecting on my position in the organization. I have done much to design and implement the TEval system, but I can't control, nor is it my position to control, the policy environment.
>
> [Reader's Note: This portion of the report contained personally identifiable information and has been removed prior to publication]
>
> The empiricist's approach would be to validate these concerns. The approach is to retrain these individuals, and collect interrater reliability data to verify that they are assigning rubric scores based on objective evidence gathered from classroom observations. In addition, I will recommend implementing a more rigorous verification system that includes random external evaluators. In this way, the integrated reliability can be quantified and measured.

My response to the policy discussion was to think about the issue from a validation point of view (i.e., attempting to assume an objective position) and not jump to conclusions. After all, given the comprehensive development work done by the committee and the metrics, plus the fact that all evaluators were certified through Danielson's Teachscape training regiment, the evaluation scores might be presenting an accurate differentiation of instructional capacity across schools and teachers. As previously discussed, I felt that the necessary approach would be to first gather reliability evidence on the scoring abilities of the problematic evaluators, verify that training issues did/did not exist, and then implement an appropriate solution. Clearly, my reaction was an attempt to challenge an as yet unsubstantiated perception of cause and effect and

maintain an empiricist's point of view. Indeed, I felt this was my role – to maintain objectivity and caution against making inferences without having an evidentiary basis.

Again pertaining to my role as information broker, I made the following journal entry. The context concerned the issue of post-conference training for administrators:

> [12.10.2012] My second research question is 'How does providing an evolving dialog on validity change organizational policy?' My 12.5.2012 discussion with [name] about threats to observational validity may lead to changes in training on how principals conduct observations [collect evidences] and the need to increase training and skills in post-conferencing and mentoring. [Name] was very interested and seemed receptive. [Name] reacted to my threats-to-validity discussion as input to improving system efficacy.
>
> This leads into my research question 4: My positionality? Am I able to be a researcher and also influence policy thinking by providing evidences? In this way I am an information broker as well as a policy agent.

The entry reflects one of the basic premises of my research activity: that access and understanding of empirical information impacts the decision environment of policy leaders. In the entry above, I recorded my sense of that influence, believing that the perspectives of these policy leaders were substantively affected by the information. A few months later, I made a follow-up entry concerning discussions with two other district policy leaders, stating:

> [2.7.2013] I discussed some of the principal interview reflections with [name] and [name]. The comments concerned not being comfortable with post conference mentoring capabilities of principals. [Name] made note and commented that increased professional development on counseling should be considered. I also discussed CFA [Confirmatory Factor Analysis] and EFA [Exploratory Factor Analysis] analysis (of the FFT rating data) and they seemed to understand the implications. Although, not sure what they took away as an actionable item.

Here, I continued to raise the issue of post-conferencing training. At the time, I believed that I was staying true to my role as information broker. However, upon reflection, I realize that as more and more stakeholder narratives were being collected and

analyzed, I was also acting as an arbitrator of that information: deciding what was important, how it was analyzed, and to whom I would share the information (i.e., the evaluation committee, key policy leaders and/or decision makers, colleagues associated with the evaluation process, etc.). My sense is that this was probably true for both the qualitative and the quantitative information.

Reflection back on these and other journal entries as well as on the writings of qualitative researchers such as England (1994), Moser (2008), Bonner and Tolhurst (2002), and Dyer (2009), I more clearly see how concepts of positionality intersect with my role of information broker and the issues impacting the research process.

I viewed myself as an objective outsider-researcher insisting on evidence for all claims made on the data. Upon reflection, I now acknowledge that this position imposes its own paradigm. Indeed, in retrospect, the idea that I could somehow remain objectively detached from the decision environment by serving as an outsider-research was naïve. In fact, the decisions I was making as an information broker were effecting the decision environment in a non-neutral manner. Add to this my personal relationships as a member of the organization's administrative team, and the façade of objectivity becomes even more unclear. Dyer (2009) reflects on this context, saying:

> As qualitative researchers we are not separate from the study … We are firmly in all aspects of the research process and essential to it … Just as our personhood affects the analysis, so, too, the analysis affects our personhood … The intimacy of qualitative research no longer allows us to remain true outsiders to the experience under study and, because of our role as researchers, it does not qualify us as complete insiders. We now occupy the space between, with the costs and benefits this status affords. (p. 61)

His insightful observation of the "…the space between..." seems to properly describe my positionality: that of a researcher-practitioner who negotiates between two

simultaneous roles within the policy context for the purpose of facilitating information exchange and informed decision making.

Researcher: Complexity of study. A major challenge I faced in this study concerned the organization, analysis, and reporting of the many forms and dimensions of the data collected. Quantitative approaches included correlational, reliability, factor analytic, statistical modeling, and ANOVA methods with numerous analytic techniques applied within each procedural context. A total of eight primary and secondary research questions required quantitative evidences.

The qualitative information was segmented by four stakeholder groups: teachers, principals, district policy, and state policy. In addition, qualitative methods were applied to 16 primary and secondary research questions. As a result, a total of 64 qualitative components were independently evaluated.

In order to manage the complexity and scope of the information, I chose to sequentially address each primary and secondary research question independently. This approach is consistent with the conceptual framework of construct validity adopted for this study, bringing many forms of independent evidences together to inform on a unified concept of construct validity (Messick, 1989a; Cronbach & Meehl, 1955; Kane, 2001).

Data handling of the quantitative evidences was straightforward, treating each data set and each research question individually. My previous background in quantitative methods, data processing, and computational software made this activity approachable and deliberate. I began the process with a clear understanding of how to structure, manipulate, and reformulate the raw data so as to align with each statistical procedure.

In contrast, my approach to managing the qualitative information was less certain. Being less experienced in handling large-scale qualitative data sets made this portion of the study difficult and technically complex. From the beginning, it was clear the narrative information was not properly structured and would prevent deep exploration of the numerous constructs posed by the research questions. I realized that I would need to extensively reformulate the raw data to make it compatible with the study's research design, facilitate comprehensive analysis, and provide an efficient approach for reporting. A detailed description of this process was previously provided in Chapter 4 under *Part 2: Descriptive Summary: Qualitative Data Collections.*

Briefly, the narrative data was originally organized by stakeholder participant (n = 22). For each participant, the same set of interview prompts were utilized, each aligned to a specific secondary research question. Thus, the raw data relevant to any particular research question was distributed across 23 different file locations. To restructure the data, all of the narrative information was brought into the HyperResearch software program. In these initial rounds, the individual narratives were reviewed independently, subjecting each to multiple rounds of coding, categorization, and annotation. This brought a board context to each individual's "voice," connecting across interview prompts and discussion topics.

In the next step, coding groups were formed within HyperResearch that compartmentalized the participant narratives into Global Constructs. These Global Constructs were more closely associated to one or more research questions. Using features present in the software program, new data structures were constructed that reorganized the narrative information by Global Construct. Once this was complete, it

became possible to engage in additional rounds of coding and categorization that were directly aligned to the questions posed by the study. The final data processing step brought together all codes, categories, annotations, and supporting exemplars into a single file. It was this organizational structure that facilitated the discussions presented herein.

In order to report on the qualitative analysis, a structured approach to presenting the information needed to be adopted. To do so, discussions of each qualitative question were organized into five general sections: Introduction, Construct Overview, Stakeholder Narrative, Summary, and Petite Assertions. This was an important organizational decision because it provided a framework for efficiently summarizing the codification activity, the operational constructs revealed by the data, and the systematic presentation of narrative evidence. The structure permited the reader to fully review the narrative evidence as well as judge inferential insights made from the data.

Finally, each qualitative research question was treated as a stand-alone analysis. Again, this was consistent with the conceptual framework adopted by the study, a unified conceptualization of construct validity informed by many independent sources of evidence. My perspective is that this was the most efficient method of dealing with the substantive amount of qualitative data informing on wide array of topics.

Summary RQ4 (a). Research Question 4 (a) concerns the challenges and barriers encountered during the research process. As discussed, concerns over data integrity, stakeholder communication, policy influence, researcher positionality, and the general complexity of the study itself conspired to make the research activity both challenging and interesting. Simultaneously navigating through these factors tested my

abilities to design and conduct a large-scale mixed-method research activity, forced reflection on my dual role as researcher and administrator, and allowed me to exercise and enhance my leadership skills. Fortunately, I believe all of the challenges encountered were surmountable. I also found that negotiating transparent access and communication of empirical information helped clarify, shape, and resolve many policy-related issues. Attention to technical details regarding information integrity and management helped facilitate comprehensive analysis and formulate the inferential findings presented in the study.

RQ4 (b) – Salient learnings: What were the salient learnings from this study? How did the researcher grow professionally and personally? How will these learnings be incorporated into future research & leadership activities?

Research Question 4 (b) explores the salient learnings I obtained during the course of the study. The discussion is organized into six areas: Reflections on Analytic Skills Sets, Reflections on Qualitative Analysis, Reflections on the Writing Process, My Transition to Scholarly Writing, Personal Significance, and Next Steps.

Reflections on analytic skill sets. To conduct this research study, I was able to utilize and expand on my analytic skill set in terms of data processing, data analysis, statistical knowledge and methods, database design, application programming, research methods, and the simultaneous application of multiple analytic traditions. The size, complexity, and scope of the research topic necessitated use of this diverse set of analytic tools not usually required by more constrained research activities. An outline of the skills and toolsets used to conduct the study is presented in Appendix Q.

The analytic skill sets used in this study extended and challenged my technical capabilities. It was a chance to bring together past training in quantitative and qualitative methods onto a single large-scale study. The experience expanded my understanding of contemporary statistical practices, application programming, and research methods. It challenged nearly every facet of my analytic skill set and provided an opportunity to acquire new knowledge and technical abilities.

Reflections on qualitative analysis. As mentioned, the qualitative analysis activities represented the most technically difficult and time consuming part of the study. This was partly due to my weaker foundation in qualitative methods. At the same time, this weakness provided the opportunity to experience the most growth, both as a researcher and personally.

The discussion below was constructed mostly from journal entries made (approximately) between March-December 2014, the period I was concentrating on the narrative information. I qualified these dates because as I began writing my analysis, I purposefully created a Word document titled *Chapter 5 & 6* to begin entering a scattering of personal notes on relevant topics. In retrospect, I should have done this as a formal journaling activity, noting appropriate dates, but at the time I was simply 'getting a start' on what I thought would be portions of my narratives and topic ideas. Over time, the collection of personal notes and reflections in the document began to take the form of a formal journaling activity.

Printed documents/touching the data. The following reflection concerns my frustration at working completely within an electronic environment while assimilating the narrative information. I simply found it too constraining; it was preventing me from

understanding the evolutionary development of the constructs present in the data. My entry was as follows:

> I need to 'touch' the data: a heathy dose of printed transcripts, narrative sections, code structures, annotations, and exemplars. I need to take the narrative with me so that electronic restrictions/limitations would not prevent me from thinking when the time suited – coffee shops, quiet moments, and instances when inspirational thoughts/question would pop up based on my topic, my focus at the time. I need to 'turn the page,' jump back and forth between pages of narrative accented by an ever growing set of margin notations, color highlighting, and otherwise insightful (at the time) notes. The physical incorporated into the electronic.

The entry above is all about the thought process, about maximizing critical reflection, of getting to the essence of the participant's "voice." I was trying to be as close to the data as possible and feeling I needed alternative approaches to dig into the information. Saldana (2014) speaks about the importance digging into qualitative data and taking the time and effort it requires, saying:

> …Sweatin' analysis is gittin' your hands dirty in the data. It's muscle work, and you git stronger and smarter with each project. Nothin' wrong with the craftspersonship of it. An honest day's qualitative work for an honest day's quantitative pay… (p. 978)

I think having access to the printed and the electronic, and the analytic benefits each provides, was me attempting to do an "honest day's work" and honor the participant's voices.

I remember thinking at the time that my age probably is preventing me from fully acclimating to the electronic environment. After all, I would never think about using a hand calculator and scratch paper to do complex statistical analysis. But for some reason, analysis of the qualitative information was different. The voices in the narrative were personal, complex, nuanced, and less apparent. This requires more attention to context, connection, and intention. At one point, I felt the electronic format was getting in the way

of sorting through everything, rearranging and comparing context. I needed to see codes, annotations, multiple stakeholder exemplars, etc., spread out in front of me.

After spending more time writing, printing out sections, and refining my activity, I made an additional (subsequent) entry in my Word document notes:

> To write the final analysis for each of the supporting qualitative research questions, I found it both necessary and comforting to have printouts of all the Global Constructs beside me. Attempting to navigate through the electronic forms of the main documents and associated subsections proved too confusing, time consuming, and difficult. Because of this, I felt restricted in my ability to quickly connect my thoughts or explore new questions in the data. In essence, printing the sections, with their color annotations, margin notes, and insightful scribbling's, served as proxy [i.e. low tech] video screens that I could switch back and forth through in an instant – far more efficient and effective because I would have had to have a dozen video monitors laid out in front of me to duplicate the effect.
>
> The ability to physically switch back and forth between documents, sections, pages impacted my thinking right up to the end. During final writing of the manuscript, I found myself verifying, questioning, and checking my assertions. Having the printed documents made it easier by being able to grab hold and open the different documents lying around me.
>
> Perhaps I am too old and set in my ways to change exclusively to a technology-based qualitative analysis, but using a combination of advanced technology-based analytic tools (HyperResearch), mainstream document applications (MS Word), and physical media (printed sections with color highlights along with margin notes using different colored pens) provided a satisfying and hopefully productive environment for understanding the qualitative information collected for this study. At the very least, it permitted me to 'touch' the data many, many times, each time refining and clarifying my thinking.

The entry references the benefits of using highlight and memo features in Word to annotate, format, and emphasize sections of the text. When the pages are output to a color printer, all the editorial aspects are clearly evident and very useful. When these multi-color documents are printed, additional rounds of editing and formatting becomes much easier, manually applying low-fi tools such as colored highlighters and colored pens. I reflected on the utility of this round of analysis in another short entry saying:

> Extensive use of color highlighters, physical color pen edits/notations, was critical to further develop & organize my thinking.

Low-tech. I also made reflections on my use of Word in this analytic process. As mentioned, HyperResearch was the primary qualitative coding application. The program is very capable, easy to use, and powerful in certain features. However, at one point I restructured the data and exported exemplar-linked codes into Microsoft Word so that I could review the information on devices not set up with HyperResearch. This was a profound step, not entirely apparent at the time. The following is a reflection on this jump out of the high-tech and into a low-tech environment:

> Microsoft Word's Track Changes allowed assignment of different color coding to different types of edits/reflections. Rather than 'accepting' track changes, I left them visible in order to maintain the color coding linked to various types of edits and annotations. (Annotations = green, edits = red, personal memos/comments = dark orange, use comment boxes as an additional source of reflection)

This alternative environment provided a different set analytic tools not present in HyperResearch. I later concluded that any quantitative researcher with sufficient experience knows that you use the application that provides the best tool for the specific task (i.e., Excel, SPSS, Mplus, SQL, HLM, AMOS, etc.). Why would this be different in qualitative research? It becomes the responsibility of the researcher to use the tools that best examine the data in question. In my case, I reached the point where I felt using Word represented a useful extension to the qualitative tool kit, just as printing the physical pages were to further rounds of analysis.

Cycles of analysis. Another "journal" entry regarding learnings I attained while conducting the qualitative analysis concerned C*ycles of Analysis*. The entry stated:

> Cycles of Analysis: There is NO substitute for reading, re-reading, and re-reading the narratives. Reading, making notation, thinking, noting questions or aspects of

> what you 'believe' may be embedded in the narrative, and then walking away. Then come back and re-read and re-formulate your thinking – does it hold after re-reading it fresh?
>
> In addition, NOT being afraid to restructure the constructs and components, to re-define, re-order, re-think about what YOU believe is present. Constant memoing, annotation, notation, documenting questions and thoughts, are all critical to developing the underlying constructs present in the narratives.
>
> This requires command of many different forms of data processing techniques and data organizational techniques – use of software, coding structures, re-grouping codes into competing categories, etc. Also, exporting to text, using color, section headers, paragraph formatting (indentation, bullets, etc.).

This was an important learning for me. Having worked on many smaller qualitative data sets usually connected with a survey activity, I had a working understanding of analyzing and coding narrative data. In addition, my university coursework provided an intellectual understanding of qualitative methods. Reading the work of researchers such as Corbin and Strauss (2008), Creswell (2009), Stake (2010), and Wolcott (2009) provided me an appreciation of the difficulty inherent in qualitative research. However, it was not until I became immersed in this study that I fully realize the expenditure of effort necessary to really connect with, and understand, the voices of participants and how those voices connect with the underlying constructs being investigated. Indeed, in one entry I state "This is hard work, incredibly time consuming, and incredibly important" I think this realization might have led to my next entry regarding time:

> Time: Time is important. It [the analysis] cannot be rushed. Initial thinking must be allowed to settle, rest, and then subjected to new cycles of re-reading, re-thinking, and critical reflections. Is the evidence really there? Are there competing interpretations? Is there bias in the perspective being imposed by the analyst?

The researcher hopes that time, effort, and cycles of analysis all work together to produce a viable and valid understanding of participants voice. In my experience, there does not seem to be shortcut for this process, especially if the analysis must carry the weight of authority, the importance of claim, or of consequence. In this regard, I subsequently reflected on the validity of my qualitative work saying:

> Once more of the proposed constructs are developed, do they hold tightly together, are they distinct, or are there still ambiguities, overlap, and uncertainties? If someone else were to read my operational definitions of the construct, with the exemplar evidences laid out, would they come to the same conclusions? Would they say "yes, that construct interpretation is clearly visible in the narrative"?

I was (and still am) concerned that my work to date is incomplete, not fully formed, and missing context and understanding. This concern starting to develop as I was writing the qualitative sections of the dissertation, realizing that I was still reformulating constructs as I was writing. I took this to mean that I had not articulated the essence of the participant perspectives and there was still more work to be done. I take solace that Corbin and Strauss (2008) comment "… even after publication, they [qualitative researchers] view their work as modifiable and open to negation as new knowledge is accrued…" (p. 13). It seems that every time I either reviewed the data or re-read my initial reflections, I accrued new knowledge.

Reflections on the writing process. During the writing process, I made a number of reflections in my *Chapter 5 & 6* Word document about the qualitative writing process. The data was complex and voluminous and I did not have a clear idea of how long or how much time the process would take. If I spent too much energy and time making things perfect on one section, I would not have sufficient time to similarly invest in

subsequent sections. I needed a strategy, particularly for the qualitative writing. In this regard, I made the following entry:

> I really need to develop a process & organizational method for completing the writing. The process that seems promising is to develop a writing cycle, working on multiple sections at once: Write a section, then leave the section; Begin writing the next section; After a few days, go back to previous section and read, edit, revise, then leave; Cycle through these sequentially until arriving at the last section.

This seemed promising at first and was used to some degree throughout. I wrote a section (still spending too much energy on trying to perfect every word and sentence), put it aside and started a new section. Then I revisited the first section, made more edits and sent it out for review (to colleagues, my spouse, or members of my committee). Next, I went back to the new section and continued writing. When I received edits from my reviewers, I made changes, and so on and so forth. Whether that was the best way, I guess additional experience will tell.

During the activity of writing up the qualitative results, I quickly learned that my initial thinking, analysis, and conclusions were not as complete as I first believed. I began to question everything. The following extended entry exposes this personal reflection:

> *OMG. I thought the final writing would be merely to document my conclusions.* NOT! The act of writing everything up is an integral part of qualitative data analysis. It forces yet another round of reflection and questioning. Does this perspective hold up under the data? Do my conclusions follow from the analysis and exemplars? Did I think this through enough?
>
> From the very beginning, I found I had to re-think, re-write, and dig back into the data. I found myself questioning everything – every sentence, use of exemplars supporting the ideas, every summarization of the collective voices.
>
> The time it is taking is directly related to my desire to be honest, transparent, and accurate. I am realizing that my analysis and conclusion are very imperfect, restrictive, and in many was misleading – that I still am not capturing the essence of the narratives. There are simply too many interconnected ideas.

> It is a painstaking process, requiring attention to words, sentence structures, and context. There is a constant fear of correctly understanding voice, of making proper attributions to the person speaking, of not injecting personal bias into the interpretation. There is constant danger of going too far in contextualizing voice, misrepresenting, or failing to see the larger context of the limited comment.
>
> To do this, one MUST agonize over words, phrases, sentences. One must see the larger context as well as the expression of narrow sentiments. Asking "is this phrase indicative of the emotional core of the person, or is this just one article in an emotional drawer of a much larger dresser.

Transition to scholarly writing. From my perspective, both the dissertation activity and the ability to write for scholarly publication differs substantively from my professional experience as a researcher-practitioner in public K-12 education. For most of my career I have been embedded in educational organizations conducting internal research to inform local policy decisions. In this context, emphasis is placed on communicating empirical findings; much less concern is afforded to documenting or publishing comprehensive explanations of the research process or contributing to a larger body of scientific understanding.

The dissertation activity represents a fundamentally different perspective. The expectation is that the dissertation demonstrates one's capacity to critically conduct and present all aspects of the research activity. Here, much more emphasis is placed on documentation, verification, explanation, and justification. All claims of authority are to be explicitly substantiated, critically assessed, and cogently presented. The process is comprehensive and exhaustive by its nature.

From my perspective, writing for journal publication is still another level of authorship. Space restrictions and the emphasis on scholarly contribution impose its own

type of criteria on the writer. For me, this represents the most difficult form of writing and the one that I have the least experience.

In this context, I see the process dissertation writing as a transformational experience, moving from policy advisor to scholar capable of contributing to a larger body of collective knowledge. The difficulty I have had with this transformation has been the focus of some reflection, as represented by the following:

Difficulty of transitioning to scholarly writing.

1st: Applied Analysis in Organizational Settings: I have always focused on providing policy makers with scientific, research-based analysis that may be useful for decision-making. The constraints are that (1) time is very restricted; eliminating comprehensive documentation or time for peer-review, multiple validation studies; (2) focus is on delivery of localized findings; and (3) findings are rarely shared outside the immediate organization and, even then, restricted to a small group of decision-makers.

2nd: Writing for Academic Learning: My graduate level classes require writing to demonstrate content mastery & competency. The focus is on proving to the instructor that you have learned the narrowly defined set of subject matter. Here, expanding on detail and justification becomes a norm. The task is to be narrowly exhaustive. It is a pathway toward the dissertation.

3rd: Dissertation: A larger representation of the capacity to conduct research. Broader in scope, exhaustive in detail. In my case the dissertation is not a journal publication but more like a book or extended treatise of my topic. My hope is that I may publish several articles from the work I have completed.

4th: Writing for Scholarly Publication: Here, there is no room to publish all the detail. The task is to be succinct while at the same time complete (as opposed to extensively documenting every aspect of the research process). Here one presents the analytic approach, not every detail, in a scholarly context, followed by the findings and conclusions. It is specifically focused on adding to the collective body of knowledge.

Personal Note: I have done the first two for so long, I find it very difficult to write my dissertation. I have this need to insert everything: process, data, details on analysis, and findings. This makes the task very complex, large, and very time consuming. I know that many of my cohort colleagues (and most doctoral

> students) outlined much more restricted research questions. But mine concerns construct validation, which is a big topic requiring many sources of evidence.

In re-reading this entry, it appears disjointed but reflects my struggle to develop my ability to become more of a scholarly writer and move beyond the role "research technician." The dissertation provides an opportunity to represent my capabilities but at the same time reveals the distance I still have to travel, putting my research into a framework that may one day be published. While this may seem like a discussion of challenge, I see it more of how much I have learned from the process. I can recognize the next level of scholarship and writing ability required to warrant legitimacy to my work.

Personal significance. The process of obtaining a doctorate has been a deeply personal experience. It represents a life goal; one that I have worked hard to attain. I believe the dissertation experience has been transformational, enhancing my skills and knowledge as a researcher, and expanding my perspectives as an academic. As a result of the dissertation activity, I believe I am fundamentally better in my role as a scientific policy advisor and organizational leader. This is mostly due to developing a wider intellectual frame of reference and a refined capacity to fully integrate quantitative and qualitative analytic methods. Used together, I am better positioned to inform policy leaders and meaningfully contribute to the decision making process.

Importantly, university coursework has made me more learned in my field. The process of completing my dissertation has arguably improved my writing skills, sharpened my thinking, and helped me become a better communicator. The guidance, support, and encouragement received from my instructors and mentors have profoundly influenced my professional and personal identity. My hope is that my experience has

empowered me to one day contribute to the knowledge base in my field. Finally, by conducting my research, I believe I have positively impacted my organization and helped shape an improved teacher evaluation environment that otherwise would not have occurred.

In this context, I reflected (below) on my own personal context by making the following (Word) entries during the fall 2014:

> Personal Reflections: This work was born from a personal desire to obtain a doctorate. This was personal, a long desired life goal.
>
> - It was not born from a desire to further my career: I am nearing the end of it.
> - Or to obtain higher salary: this degree has cost a considerable amount of money including tuition, fees, and a considerable amount of lost wages during the year it took to complete the writing of the dissertation – all of which could have been used to make my family's life more comfortable.
> - Or to achieve higher professional status - I have enjoyed an active, successful career in applied education research and assessment and I believe I am well-respected by my peers.
>
> However, the reality of it is that it has caused considerable disruption to family (inconvenience) and friends (no time to nurture relationships, give back to others, or develop new connections). In addition, there has been considerable impact on my staff. When I went to part time, I was not in the office and unable to fully mentor and support, causing disengagement from those that worked for me.

The economist in me recognizes the sacrifices made as an opportunity cost of obtaining the degree. These are real costs in terms of money, time, effort, health, and impact on family and friends. Thus, conduct of the dissertation has meant much more to me than completion of a university graduation requirement. It is an affirmation of the sacrifices endured by those around me. As a result, the degree will have that much more significance.

Regarding the dissertation itself, I made the following entry at approximately the same time as my personal reflections of meaning, effort, and sacrifice. The intent was to remind myself of what I wanted to write in Chapter 6 as a retrospective on the dissertation document and the personalized context in which I approached the task. However, now I believe the bulleted items can stand without much additional explanation.

> What does this [dissertation] effort represent (to me)?
> - It is not for publication in a journal.
> - It is not condensed, reduced, or otherwise a summary of some other body of analysis.
> - I have not tried to simplify the analysis, since the context requires that this level of analysis to be done.
> - It is transparent, complete, rigorous, and representative of the research questions being asked.
> - This is the reference from which other dimensions of summary can be examined: i.e., particular perspectives of the [teacher evaluation] question.
> - It is me. It is the best I can do within the context of my life.
> - I have thought deeply, extensively, comprehensively about the questions and evidences. This document reflects that thinking in its entirety. It is not open to question how I interpreted each aspect of the activity. It is laid out for all to see. Feel free to disagree with my analysis, but not that I have intentionally excluded anything.
> - I believe it to be scholarly, recognizing that I am a naïve scholar.
> - I am new at this.

Next steps. What is next (for me)? I made the following reflective entry in my *Chapter 5 & 6* notes regarding future personal goals and activities.

> What is next for me? How will this experience be incorporated into my future research and leadership?
>
> - Publish; Share [my research] at national conferences, share at local conferences
> - Continue to engage in policy discussions and write to impact/influence policy decisions
> - Get back to reading the published literature in this area [evaluation], accountability systems, validity studies, value-added models, and measurement in public policy

- Focus [future work] on evaluation systems, validity studies, accountability, value-added, research methods – both quantitative and qualitative
- [Expand my capacity on] Accounting for Nested data within empirical methods (i.e. HLM, CFA/EFA, etc.)

The collection of "next steps" reflects my desire to participate more fully in the public policy dialog, improve my skills regarding contemporary research methods (i.e., problems related to nested, hierarchical data), and to generally continue to make a difference in public education. I add to this list, my hope that I can once again teach classes in research methods, statistics, testing and measurement, and/or data literacy training to persons interested in education and public policy. I believe my experience and training have something important to offer beyond the technical.

Summary RQ4 (b). Research Question 4 (b) concerned personal reflections on salient learnings, growth, and future activities. The narrative focused mostly on the impact of the dissertation activity (i.e., the research study and the process of writing the dissertation). The perspective highlighted improvements in technical skills/knowledge as well as aspects of personal/professional identity. In this regard, I described the activity as transformational. The overall impact cited improvements in analytic skillsets, expansion of intellectual framework, enhance efficacy as a policy scientist, heightened leadership capabilities, and scholarly preparation. The overall experience sets the stage for future endeavors as a policy scientist, organizational leader, future scholar, instructor, and professional mentor.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Belmont, CA: Wadsworth, Inc.

American Education Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1974). *Standards for educational and psychological tests and manuals.* Washington, D.C.: American Psychological Association.

American Education Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

American Education Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

American Psychological Association (APA). (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, D.C.: American Psychological Association, Inc.

American Psychological Association (APA). (2001). *Appropriate use of high stakes testing in our nation's schools.* Retrieved from http://www.apa.org/pubs/info/brochures/testing.aspx

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher*, *37*(2), 65-75.

Amrein-Beardsley, A. (2009). Value-added tests: Buyer, be aware. *Educational Leadership*, *67*(3), 38-42.

Amrein-Beardsley, A, & Barnett, J. H. (2012). Working with error and uncertainty to increase measurement validity. *Educational Assessment, Evaluation and Accountability, 24*(4), 369-379. http://dx.doi.org/10.1007/s11092-012-9146-6

Amrein-Beardsley, A., & Collins, C. (2012). The SAS education value-added assessment system (SAS® EVAAS®) in the Houston independent school district (HISD): Intended and unintended consequences. *Education Policy Analysis Archives, 20*(12), 1-31.

Anderson, G. L., Herr, K., Nihlen, A. S. (2007). *Studying your own school: An editor's guide to practitioner action research* (2nd ed.). Thousand Oaks, CA; Corwin Press.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H.I. Braun, *Test validity* (pp. 19-32). Hillsdale, NJ: Lawrence Erlbaum Associates.

Apple, M. (2007). Education, markets, and an audit culture. *International Journal of Educational Policies, 1*(1), 4-19.

Arbuckle. J. L. (2005). *AMOS 6.0 users guide*. Spring House, PA: Amos Development Corporation

Arizona Department of Education (ADE). (2012a). *Arizona framework for evaluating educator effectiveness fact sheet*. Phoenix, AZ: Arizona Department of Education

Arizona Department of Education (ADE). (2012b). *Race to the Top scope of work*: *June 2012*. Retrieved from http://www2.ed.gov/programs/racetothetop/state-scope-of-work/arizona.pdf

Ariz. Rev. Stat. §15-203A.38 (2010).

Armor, D. J. (1974). Theta reliability and factor scaling. In H. Costner (Ed.), *Sociological methodology* (pp. 17-50). San Francisco, CA: Jossey-Bass.

ATS Statistical Consulting Group. (2012). *Annotated SPSS output: Factor analysis.* Los Angeles, CA: Institute for Digital Research and Education. Retrieved from http://www.ats.ucla.edu/stat/hlm/seminars/hlm6/outline_hlm_seminar.pdf

ATS Statistical Consulting Group. (2013). *Introduction to multilevel modeling using HLM 6* [Presentation]. Los Angeles, CA: Institute for Digital Research and Education. Retrieved from http://www.ats.ucla.edu/stat/hlm/seminars/hlm6/outline_hlm_seminar.pdf

Au, W. (2010). Neither fair nor accurate: Research based reasons why high stakes tests should not be used to evaluate teachers. *Rethinking Schools, 25*(2), 34 – 38.

Austin Independent School District. (2012). *Teacher evaluation system: ASID Reach, 2012-2013*. Austin, TX: Austin Independent School District.

Baird, J., & Black, P. (2013). The reliability of public examinations. *Research Papers in Education, 28(*1), 1-4. http://dx.doi.org/10.1080/02671522.2012.754232

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Sheppard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers. EPI briefing paper # 278*. Washington, D.C.: Economic Policy Institute.

Baum, G. (2010). N.Y. union seeks to curb teacher data. *L.A. Times*. Retrieved from http://www.latimes.com/news/local/teachers-investigation/la-me-teachers-new-york-union,0,7294443.story

Berliner, D. C. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education*, *56*(3), 205-213.

Berliner, D. C. (2009). *Poverty and potential: Out-of-school factors and school success*. Boulder, CO and Tempe, AZ: Education and the Public Interest Center & Education Policy Research Unit. Retrieved from http://epicpolicy.org/publication/poverty-and-potential

Berliner, D. C. (2014). *The teacher evaluation paradox: Confusing teachers and the public for years*. Unpublished paper presented at the Annual Conference of the Arizona Education Research Organization, Phoenix, Arizona.

Berliner D. C., & Biddle, B. J. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. Reading, MA.: Addison-Wesley Publishing Company.

Betebenner, D. W. (2008). *Norm and criterion-referenced student growth*. Dover, NH: National Center for the Improvement of Educational Assessment.

Bill and Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the met project's three-year study*. Retrieved from www.edweek.org/media/17teach-met1.pdf

Blank, R. F. (2010). *State growth models for school accountability: Progress on development and reporting measures of student growth.* Washington, D.C.: Council of Chief State School Officers.

Bmuthen. (2002, September 10). Categorical data modeling [Discussion post]. Retrieved from http://www.statmodel.com/discussion/messages/23/208.html?1031672785

Bonanomi, A., Ruscone, M. N., & Osmetti, S. A. (2013, June). *The polychoric ordinal alpha, measuring the reliability of a set of polytomous ordinal items*. Paper presented at the Italian Statistical Society (SIS) 2013 Statistical Conference, University of Brescia, Milan, Italy. Retrieved from http://meetings.sis-statistica.org/index.php/sis2013/ALV/paper/viewFile/2651/424

Bond, T. G., Fox, & C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human science* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates

Bonner, A., & Tolhurst, G. (2002). Insider-outsider perspectives of participant observation. *Nurse Researcher*, *9*(4), 7-19.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Policy Information Center, Educational Testing Service

Brown, J. D. (2009). Choosing the right type of rotation in PCA and EFA. *JALT Testing and Evaluation SIG Newsletter*, *13*(3), 20-25.

Brown, S. (2011, April). Measures of shape: Skewness and kurtosis. Retrieved March 15, 2015 from http://www.tc3.edu/instruct/sbrown/stat/shape.htm.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NJ: Guildford Press.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230-258.

Brualdi, A. (1999). *Traditional and modern concepts of validity.* Washington, D.C.: ERIC Clearinghouse on Assessment and Evaluation.

Bryk A. S., Gomez L. M., & Grunow A. (2011). Getting ideas into action: Building networked improvement communities in education. Stanford, CA: Carnegie Foundation for the Advancement of Teaching. Retrieved from http://www.carnegiefoundation.org/spotlight/webinar-bryk-gomez-building-networkedimprovement-communities-in-education

Bulmer, M. G. (1979). *Principles of statistics*. Cambridge, MA: M.I.T. Press.

Butrymowicz, S., & Garland. S. (2012). New York City teacher ratings: How its value-added model compares to other districts. *Huffington Post.* Retrieved from http://www.huffingtonpost.com/2012/03/02/new-york-city-teacher-rat_n_1316755.html

Butterfield, K., & Amator, J. (2012). *Arizona framework for measuring educator effectiveness: Statewide awareness presentation*. Phoenix, AZ: Arizona Department of Education. Retrieved from http://www.azed.gov/teacherprincipal-evaluation/files/2012/04/statewide-awareness-presentation.pdf

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Newberry Park, CA: Sage Publications.

Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment, 20*(3), 333–346. http://dx.doi.org/10.1111/j.1468-2389.2012.00604.x

Cavanaugh, J. E. (2009). *Model selection: Introductory principles, concepts, and procedures* [Lecture notes]. Iowa City, IA: Author. Retrieved from http://myweb.uiowa.edu/cavaaugh/ms_lec_1_ho.pdf

Chou, C. P., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37-55). Thousand Oaks, CA: Sage Publications, Inc.

Citroen, C. L. (2011). The role of information in strategic decision-making. *International Journal of Information Management, 31*(6), 493–501. http://dx.doi.org/0.1016/j.ijinfomgt.2011.02.005

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (2006). Social, emotional, ethical, and academic education: Creating a climate for learning, participation in democracy, and well-being. *Harvard Educational Review*, 76(2), 201-237.

Cohen, J, & Cohen, P (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

College Board. (n.d.). Validity handbook. Retrieved from http://research.collegeboard.org/services/aces/validity/handbook

Collins, C. (2012). *Houston, we have a problem: Studying the SAS Education Value-Added Assessment System (SAS EVAAS) from teacher's perspectives in the Houston Independent School District* (Doctoral dissertation). Retrieved from Proquest Dissertations and Theses. (UMI No. 3547764)

Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods*, *6*(2), 147-168. http://dx.doi.org/10.1177/1094428103251541

Corbell, K. A., Osborne, J. W., & Grable, L. L. (2008). Examining the performance standards for in-service teachers: A confirmatory factor analysis of the assessment of teachers' NETS-T expertise. *Computers in Schools*, *25*(1-2), 10-24. http://dx.doi.org/10.1080/07380560802157683

Corbin, J., & Strauss, A. (2008). *Basic of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Corcoran, S. P. (2010). *Can teachers be evaluated by their student's test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice*. Providence, RI: Annenberg Institute for School Reform at Brown University.

Council of Chief State Schools Officers (CCSSO). (2012). *Chief's research primer A summary of measures of effective teaching's (MET) recent findings, gathering feedback for teaching: Combining high-quality observations with student surveys*

and achievement Gains. Washington, D.C.: CCSSO Research and Development Service.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Harcourt College Publishers.

Cronbach, L. (1971). Test validation. In R. L. Thorndick (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, D.C.: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer, & H. I. Braun (Ed.), *Test Validity* (pp. 3-18). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(5), 281–302.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16-29.

Daley, G., & Kim, L. (2010). *A teacher evaluation system that works*. Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved from http://www.tapsystem.org/publications/wp_eval.pdf

Danielson, C. (2007). *Enhancing professional practice, A framework for teaching*. Alexandria, VA: ASCD.

Danielson, C. (2010). Evaluations that help teachers learn. *Educational Leadership*, *68*(4), 35-39.

Danielson, C. (2011). *The framework for teaching evaluation instrument: 2011 edition*. Princeton, NJ: The Danielson Group.

Danielson, C. (2012). Teacher evaluation: What's fair? What's effective? *Educational Leadership*, *70*(3), 32-37.

Danielson Group. (2013). *The framework.* Retrieved from http://danielsongroup.org/framework/

Darling-Hammond, L. (1997). School reform at the crossroads: confronting the central issues of teaching. *Educational Policy*, *11*(2), 151–156.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Kappen*, *93*(6), 8–15.

DeCoster, J. (1998). *Overview of factor analysis.* Retrieved from http://www.stat-help.com/factor.pdf

Delmar, C. (2010). "Generalizability" as recognition: Reflections on a foundational problem in qualitative research. *Qualitative Studies*, *1*(2), 115-128.

Dewey, J. (1900). *The school and society*. Chicago, IL: The University of Chicago Press.

Dewey, J. (2009). *Democracy and education: An introduction to the philosophy of education*. Radford, VA: Wilder Publications. (Original work published 1916)

Dyer, S. C. (2009). The space between: On being an insider-outsider in qualitative research. *International Journal of Qualitative Methods*, *8*(1), 54-63.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179-197.

Embretson, S. E. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Embretson, S. (2007). A universal validity system or just another test evaluation procedure? *Educational Researcher*, *36*(8), 449 - 455. http://dx.doi.org/0.3102/0013189X07311600

England, K. V. L. (1994). Getting personal: Reflexivity, positionality, and feminist research. *The Professional Geographer*, *46*(1), 80-89.

Erpenbach, W. J. (2009). *Determining adequate yearly progress in a state performance or proficiency index model*. Washington, D.C.: Council of Chief State Schools Officers.

Erpenbach, W. J. (2011). *Statewide educational accountability systems under the NCLB act: A report on 2009 and 2010 amendments to state plans*. Washington, D.C.: Council of Chief State Schools Officers.

Fast, E. F., & Hebbler, S. (2004). *A framework for examining validity in state accountability systems.* Washington, D.C.: Council of Chief State Schools Officers.

Ferguson, G. T., & Takane, Y. (1989). *Statistical analysis in psychology and education*. New York, NY: McGraw-Hill Publishing.

Freire, P. (1970). *Pedagogy of the oppressed*. New York, NY: Herder and Herder.

Freire, P. (1994). *Pedagogy of the hope*. New York, NY: Continuum.

Fullan, M. (2009). Have theory, will travel: A theory of action from system change. In A. Hargreaves & M. Fullan (Eds.), *Change wars*. Bloomington, IN: Solution Tree.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, *17*(3), 1-13.

Gelo, O., Braakmann, D., & Benetka, G. (2008). Quantitative and qualitative research: Beyond the debate. *Integrative Psychological & Behavioral Science*, *42*(3), 266-290.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference*. 11.0 update (4th ed.). Boston, MA: Allyn & Bacon

Gergen, K. J. (2009). *An invitation to social construction* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Gibbons, J. D. (1993). *Nonparametric measures of associations*. Newbury Park, CA: Sage Publications.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine Publishing Company.

Goldhaber, D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *American Economic Review*, *100*(2), 250–255. http://dx.doi.org/10.1257/aer.100.2.250

Good, T. L. (1999). The purpose of schooling in America. *The Elementary School Journal, 99*(5), 383–389.

Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Research*, *36*(8), 456–462. http://dx.doi.org/10.3102/0013189X07311607

Green, S. B., & Salkind, N. J. (2011). *Using SPSS for Windows and Macintosh: Analyzing and understanding data*. Boston, MA: Prentice Hall.

Greene, J. C. (2007). *Mixed methods in social inquiry*. San Francisco, CA: Wiley and Sons.

Greene, V. L. (1977). A note on theta reliability and metric invariance. *Sociological Methods & Research, 6*(1), 123-128.

Haertel, E. (2008). *Instability of teacher effects estimates from value-added models* [Presentation materials]. Retrieved from http://www.ctc.ca.gov/seminars/VAM/Haertel.pdf

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27. http://dx.doi.org/10.1111/j.1745-3992.2004.tb00149.x

Hall, G., & Hord. S. (2011). *Implementing change: Patterns, principals and potholes*. New York, NY: Ablongman Pearson Education.

Hall, G. E., Loucks, S. F., Rutherford, W. L., & Newlove, B. W. (1975). Levels of use of the innovation: A framework for analyzing innovation adoption. *Journal of Teacher Education*, *26*(1), 52–56.

Hargreaves, A., & Fink, D. (2004). The seven principles of sustainable leadership. *Educational Leadership*, 61(7), 8-13.

Hargreaves, A., & Fink, D. (2006). *Sustainable leadership*. San Francisco, CA: Wiley and Sons

Hargreaves, A., & Fullan (2009). *Change wars*. Bloomington, IN: Solution Tree.

Hargreaves, A., & Shirley, D. L. (2009). *The fourth way: The inspiring future for educational change*. Thousand Oaks, CA: Corwin Press

Harris, D. N. (2013). *How do value-added indicators compare to other measures of teacher effectiveness? Knowledge brief 5*, Stanford, CA: Carnegie Knowledge Network, Carnegie Foundation for the Advancement of Teaching. Retrieved from http://carnegieknowledgenetwork.org/briefs/value-added/value-added-other-measures

Harris, D N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal*, *51*(1), 73–112. http://dx.doi.org/10.3102/0002831213517130

Heneman III, H. G., & Milanowski, A. (2003). Continuing assessment of teacher reactions to a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education*, *17*(3), 171-195.

Henry, G. T., Kershaw, D. C., Zulli, R. A., & Smith, A. A. (2012). Incorporating teacher effectiveness into teacher preparation program evaluation. Journal of Teacher Education, 63(5), 335-355. http://dx.doi.org/10.1177/0022487112454437

Herr, K., Anderson, G. L. (2005). *The action research dissertation: A guide for students and faculty*. Thousand Oaks, CA: Sage Publications.

Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management*, *28*(4), 692–712. http://dx.doi.org/10.1002/pam.20463

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1994). *Applied statistics for the behavioral sciences*. Boston, MA: Houghton Mifflin Company

Ho, A. D., & Yu, C. C. (2014). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement.* http://dx.doi.org/10.1177/0013164414548576

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education*, 1(3), 207–219.

Hord, S. M., Rutherford, W. L., Huling-Austin, L., & Hall, G. (1987). *Taking charge of change*. Alexandria, VA: ASCD.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 69*(1), 1–55.

Institute of Education Sciences (IES). (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide,* Washington, D.C.: U. S. Department of Education, National Center for Education Evaluation.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*(1).

Kane, M. T. (1992a). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535.

Kane, M. T. (1992b). The assessment of professional competence. *Evaluation and Health Professions*, *15*(2), 163–182.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319-342. http://dx.doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2010). Validity and fairness. *Language Testing*, *27*(2), 177–182.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. http://dx.doi.org/10.1111/jedm.12000

Kane, M. T., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education*, *17*(3), 221-240.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2010). *Identifying effective classroom practices using student achievement data*. Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w15803.pdf

Kim, J. O., & Mueller, C. W. (1978). *Introduction to factor analysis*. Newberry Park, CA: Sage Publications.

Kimball, S. M., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly*, *45*(1), 34-70. http://dx.doi.org/10.1177/0013161X08327549

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Klugh, H. E. (1986). *Statistics: The essentials for research* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kotter, J. (1996). *Leading change*. Boston, MA: Harvard Business School Press.

Kotter, J. (2010, November 3). Why should anyone trust your vision*? Harvard Business Review*. Retrieved from https://hbr.org/2010/11/why-should-anyone-trust-your-v.html

Kotter, J. (2011, February 16). Before you can get buy-in, people need to feel the problem. *Harvard Business Review*. Retrieved from https://hbr.org/2011/02/before-you-can-get-buy-in-peop/

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value-added assessment system. *Educational Evaluation and Policy Analysis*, *25*(3), 287-298. http://dx.doi.org/10.3102/01623737025003287

Labaree, D. F. (1997). Public goods, private goods: The American struggle over educational goals. *American Educational Research Journal*, Vol. 34, No. 1, 39-81.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563-575.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. New York, NY: Cambridge University Press.

Learning Sciences Marzano Center. (2012a). *Guidance in reading selected research documents regarding the Marzano teacher evaluation model*. Retrieved from http://www.marzanoevaluation.com/files/Guidance%20in%20Reading%20Selected%20Research%20Documents%20Marzano%20Teacher%20Evaluation%20Model.pdf

Learning Sciences Marzano Center. (2012b). *The Marzano causal teacher evaluation model* white paper prepared for the Oklahoma State Department of Education.

Retrieved from http://www.marzanoevaluation.com/files/Oklahoma-Marzano-Teacher-Evaluation-White-Paper.pdf

LeCroy, C. W., & Krysik, J. (2007). Understanding and interpreting effect size measures. *Social Work Research*, *31*(4), 243-248.

Leo, S. F., & Lachlan-Hache, L. (2012). *Creating summative educator effectiveness scores: Approaches to combining measures*. Washington, D.C.: American Institute for Research.

Li, H., & Wainer, H. (1997). Teacher's corner: Toward a coherent view of reliability in test theory. *Journal of Educational and Behavioral Statistics*, *22*(4), 478-484. http://dx.doi.org/10.3102/10769986022004478

Linn, R. (2008). *Validation of uses and interpretations of state assessments*. Washington, D.C.: Council of Chief State Schools Officers.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36(8), 437-448.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing.

Loucks, S. F, & Hall, G. E. (1979, April 12). Implementing innovations in schools: A concerns-based approach. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California.

Loucks-Horsley, S. (1996). Professional development for science education: A critical and immediate challenge. In R. Bybee (Ed.), *National standards and the science curriculum*. Dubuque, IA: Kendall/Hunt Publishing Company.

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage Publications, Inc.

Maguire, T., Hattie, J., & Haig, B. (1994). Construct validity and achievement assessment. *The Alberta Journal of Educational Research*, *40*(2), 109-126.

Manna, P, & Ryan, L.L. (2011). Competitive grants and educational federalism: President Obama's Race to the Top program in theory and practice. *The Journal of Federalism*, *41*(3), 522-546.

Marsh, H. W., Hau, K. T., & Wen, Z (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320-341. http://dx.doi.org/10.1207/s15328007sem1103_2

Marzano, R. (2011). *Research base and validation studies on the Marzano Evaluation Model*. Retrieved from http://www.marzanoevaluation.com/files/Research_Base_and_Validation_Studies_Marzano_Evaluation_Model.pdf

Marzano, R., Waters, T., & McNulty, B. (2005). *School leadership that works*. Alexandria, VA: ASCD.

Mason, E. J. (2007). Measurement issues in high stakes testing, *Journal of Applied School Psychology, 23*(2), 27-46. http://dx.doi.org/10.1300/J370v23n02_03

Mazerolle, M. J. (2004). *Making sense out of Akaike's Information Criterion (AIC): Its use and interpretation in model selection and inference from ecological data*. Quebec City, Quebec, Canada: Université Laval. Retrieved from http://archimede.bibl.ulaval.ca/archimede/fichiers/21842/apa.html

McCaffrey, D. F., & Hamilton, L. (2007). *Value-added assessment in practice: Lessons from the Pennsylvania value-added assessment system pilot project*. Santa Monica, CA: RAND Corporation.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models of teacher accountability*. Santa Monica, CA: RAND Corporation.

McClellan, C. (2012). Teacher evaluator training & certification: Lessons learned from the measures of effective teaching project. San Francisco, CA: Teachscape. Retrieved from http://www.teachscape.com/resources/teacher-effectiveness-research/2012/02/teacher-evaluator-training-and-certification.html

McCoach, D. B., & O'Connell, A. A. (2012, April). *An introduction to hierarchical linear modeling for educational researchers*. Workshop presented at the American Education Research Association Annual Conference, Vancouver, British Columbia, Canada.

McGee-Banks, C. A., & Banks, J. A. (1997). Reforming schools in a democratic pluralistic society. *Educational Policy*, *11*(2), 183–193.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, *16*(2), 16-18.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*(11), 1012-1027.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer, &H. I. Braun, *Test validity* (pp. 33-46). Mahwah, NJ: Lawrence Erlbaum Associates.

Messick, S. (1989a). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Phoenix, AZ: American Council on Education and Oryx Press.

Messick, S. (1989b*).* Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5-8.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*(1-3), 35-44.

MET Project. (2010). *A composite measure of teacher effectiveness*. Seattle, WA: Bill and Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Value-Add_100710.pdf

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and students achievement evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33-35.

Milanowski, A. (2011, April 10). *Validity research on teacher evaluation systems based on the framework for teaching*. Paper presented at the 2011 annual meeting of the American Education Research Association, New Orleans, Louisiana. Retrieved from http://files.eric.ed.gov/fulltext/ED520519.pdf

Milanowski, A., & Kimball, S. M. (2005, April 13). *The relationship between teacher experience and student achievement: A synthesis of three years of data*. Paper presented at the American Educational Research Association Annual Meeting, Montreal, Canada.

Morse, J. M. (1999). Qualitative generalizability. *Qualitative Health Research*, *9*(1), 5-6.

Moser, S. (2008). Personality: A new positionality? *Area*, *40*(3), 383-392. http://dx.doi.org/10.1111/j.1475-4762.2008.00815.x

Moss, P. A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice*, *14*(2), 5-13.

Muthen, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*(4), 338-354.

Muthen, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, *22*(2), 376-398.

Muthen, B.O. (2011, September 8). Exploratory factor analysis [Discussion post]. Retrieved from http://www.statmodel.com/discussion/messages/8/205.html?1348513453

Muthen, L. K. (2014, May 24). Re: Structural equation modeling [Discussion post]. Retrieved from http://www.statmodel.com/discussion/messages/11/411.html

Muthen, L. K., & Muthen, B. O. (2012). *Mplus, statistical analysis with latent variables, users guide* (7th ed.). Los Angeles, CA: Muthen & Muthen.

National Commission on Excellence in Education. (1983). *A nation at risk*. Washington, D. C.: U.S. Department of Education.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18*(23), 1–27.

Nicholson-Crotty, S., & Staley, T. (2012). Competitive federalism and race to the top application decisions in the American states. *Educational Policy*, *26*(1), 160-184.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, *32*(3), 396-402.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163 –193. http://dx.doi.org/10.3102/0002831210362589

Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123-141.

Patton, M. L. (2012). *Understanding research methods*. Glendale, CA: Pyrczak Publishing.

Pearson, M., & Yan, H. (2012). Official: No deal yet between Chicago teachers and school system. *CNN*. Retrieved from http://www.cnn.com/2012/09/10/us/illinois-chicago-teachers-strike/index.html

Peifer, A. (2014). The purpose of public education and the role of the school board. *National Connection, National School Boards Association*. Retrieved from http://www.nsba.org/sites/default/files/The%20Purpose%20of%20Public%20Education%20and%20the%20Role%20of%20the%20School%20Board_National%20Connection.pdf.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Thousand Oaks, CA: Sage Publications.

Plano Clark, V. L., & Creswell, J. W. (2010). *Understanding research: A consumer's guide*. Upper Saddle River, NJ: Pearson Education, Inc.

Popham, W. J. (1997). Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publishing, Inc.

Rigdon, E. E., & Ferguson, C. E. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research*, *28(*4), 491-497.

Rogers, E. (2003). *Diffusion of innovations* (5th ed.). New York, NY: Free Press

Saldana, J. (2009). *The coding manual for qualitative researchers*. Thousand Oaks, CA: Sage Publications.

Saldana, J. (2014). Blue-collar qualitative research: A rant. *Qualitative Inquiry*, *20*(8), 976-980. http://dx.doi.org/10.1177/1077800413513739

Sanders, W. L. (1998). Value-added assessment. *The School Administrator*, *55*(11), 24-27.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system: Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*(3), 299-311.

Schlacter, L., & Thun, Y. M. (2004). Paying for high- and low-quality teaching. *Economics of Education Review*, *23*, 411–430.

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains*. Washington, D.C.: Nation Center for Education Evaluation and Regional Assistance, U.S. Department of Education.

Shepard, L. A., (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education*, (Vol. 19, pp. 405-450). Washington, D.C.: AERA.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-24

Sloan, W. M. (2012). What is the purpose of education? *ASCD Education Update*, *Quality Feedback*, 54(7), 1-3.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modelling.* London, England: Sage Publications.

Sparks, S. D. (2010). Federal criteria for studies grow. *Education Week*, *30*(8), 1-12.

Stake, R. E. (1978). The case study method in social inquiry. *Educational Researcher*, *7*(2), 5-8.

Stake, R. E. (2010). *Qualitative research: Studying how things work*. New York, NY: Guilford Press.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Strauss, A., & Corbin, J. M. (1990). *Basics of qualitative research: Grounded theory procedures and* techniques. Thousand Oaks, CA: Sage Publications, Inc.

Stringer, E. T. (2007). Action research (3rd ed.). Thousand Oaks, CA: Sage Publications.

Stronge, J. H. (2012). Stronge evaluation system report. Retrieved from http://www.strongeandassociates.com/files/Stronge%20Evaluation%20System%20Report.pdf

Stronge and Associates. (2012). *Teacher and leader effectiveness performance evaluation system* [Presentation]. Retrieved from http://www.strongeandassociates.com/files/STRONGE%20PRESENTATION%203.6.12%20web.pdf

Stumbo, C., & McWalters, P. (2010). Measuring effectiveness: What will it take? *Educational Leadership*, *68*(4), 10.

Subedi, B. R., Swan, B., & Hynes, M. C. (2011). Are school factors important for measuring teacher effectiveness? A multilevel technique to predict student gains through a value-added approach. *Education Research International*, *2011*(2011), 1-10. http://dx.doi.org/10.1155/2011/532737

Sullivan, J. L. (1979). *Reliability and validity assessment*. Newberry Park, CA: Sage Publications.

Tan, X., & Michel, R. (2011, September). Why do standardized testing programs report scaled scores? Why not just report the raw or percent-correct scores? *R&D Connections*, *16*.

Task Force on Teacher and Principal Evaluations. (2011). *Arizona framework for measuring educator effectiveness*. Phoenix, AZ: Arizona Department of Education.

Templin, J. (2008). *Assessing the adequacy of hierarchical models* [Presentation]. Retrieved from http://jtemplin.coe.uga.edu/files/hlm/hlm08mi/hlm08mi_lecture05.pdf

Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues.* Thousand Oaks, CA: Sage Publications.

TojibTraub, R. E., & Rowley, G. L. (1991). Understanding reliability. *ITEMS: Instructional Topics in Educational Measurement*, 37-45. Retrieved from http://ncme.org/linkservid/65F3B451-1320-5CAE-6E5A1C4257CFDA23/showMeta/0/

Tuytens, M., & Devos, G. (2009). Teachers' perception of the new teacher evaluation policy: A validity study of the policy characteristics scale. *Teaching and Teacher Education*, *25*(6), 924-930. http://dx.doi.org/10.1016/j.tate.2009.02.014

Uebersax J. S. (2006). *The tetrachoric and polychoric correlation coefficients*. Retrieved from http://john-uebersax.com/stat/tetra.htm

U.S. Department of Education (USDOE). (2009). *Race to the top program: Executive summary*. Washington, D.C.: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

U.S. Department of Education (USDOE). (2015). Race to the top fund: Awards. Retrieved March 3, 2015 from http://www2.ed.gov/programs/racetothetop/awards.html

Viteritti, J. P. (2012). The federal role of school reform: Obama's Race to the Top. *Notre Dame Law Review, 87*(5), 2087-2120.

Wainer, H., & Braun, H. I. (1988). *Test validity*. Mahwah, NJ: Lawrence Erlbaum Associates.

Watanabe, T. (2011). Value-added teacher evaluations: L.A. unified tackles a tough formula. *Los Angeles Times*. Retrieved from http://www.latimes.com/news/local/la-me-adv-value-add-20110328,0,3903343.story

Wayne, A. (2010). Neither fair nor accurate: Research-based reasons why high-stakes tests should not be used to evaluate teachers. *Rethinking Schools,* (Winter). Retrieved from http://www.rethinkingschools.org/archive/25_02/25_02_au.shtml

Webber, K. C., Rizo, C. F., & Bowen, N. K. (2012). Confirmatory factor analysis of the elementary school success profile for teachers. *Research on Social Work Practice, 22*(1), 77-84. http://dx.doi.org/10.1177/1049731511415549

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. New York, NY: Cambridge University Press.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project.

Weiss, E. (2014). Mismatches in race to the top limit education improvement: Lack of time, resources, and tools put lofty state goals out of reach. *The Education Digest*, *79*(5), 60-65.

Westen, D., & Rosenthal, R. (2005). Improving construct validity: Cronbach, Meehl, and Neurath's ship. *Psychological Assessment*, *17*(4), 409–412.

White House. (2009). *Fact sheet: Race to the Top.* Retrieved from http://www.whitehouse.gov/the-press-office/fact-sheet-race-top

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow, & D. E. Wiley (Eds.), *Improving inquiry in the social sciences: A volume in honor of Lee J. Cronbach* (pp. 75-107). Hillsdale, NJ: Erlbaum.

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, *45*(3), 197-210.

Wolcott, H. F. (2009). *Writing up qualitative research* (3rd ed.). Los Angeles, CA: Sage.

Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256-293). Newbury Park, CA: Sage.

Yeager, D., Bryke, A. S., Muhich, J., Hausman, H., & Morales, L. (2013). *Practical measurement*. Austin, TX: Carnegie Foundation for the Advancement of Teaching.

Yeomans, K. A., & Golder, P. A. (1982). The Guttman-Kaiser criterion as a predictor of the number of common factors. *Journal of the Royal Statistical Society*, *31*(3), 221-229.

Zumbo, B. D., Gadermann, A. M., & Cornelia Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods*, *6*(1), 21-29.

APPENDIX A

DANIELSON FRAMEWORK FOR TEACHING

Example Domain Performance Descriptors for Domain 1a: Knowledge of Content and Pedagogy

1a Performance Level Descriptors

Unsatisfactory	Basic	Proficient	Distinguished
In planning and practice, teacher makes content errors or does not correct errors made by students. Teacher's plans and practice display little understanding of prerequisite relationships important to student's learning of the content. Teacher displays little or no understanding of the range of pedagogical approaches suitable to student's learning of the content.	Teacher is familiar with the important concepts in the discipline but displays lack of awareness of how these concepts relate to one another. Teacher's plans and practice indicate some awareness of prerequisite relationships, although such knowledge may be inaccurate or incomplete. Teacher's plans and practice reflect a limited range of pedagogical approaches to the discipline or to the students.	Teacher displays solid knowledge of the important concepts in the discipline and the ways they relate to one another. Teacher's plans and practice reflect accurate understanding of prerequisite relationships among topics and concepts. Teacher's plans and practice reflect familiarity with a wide range of effective pedagogical approaches in the discipline.	Teacher displays extensive knowledge of the important concepts in the discipline and the ways they relate both to one another and to other disciplines. Teacher's plans and practice reflect understanding of prerequisite relationships among topics and concepts and provide a link to necessary cognitive structures needed by students to ensure understanding. Teacher's plans and practice reflect familiarity with a wide range of effective pedagogical approaches in the discipline, anticipating student misconceptions.

1a Behavioral Possible Examples

Unsatisfactory	Basic	Proficient	Distinguished
The teacher says, "The official language of Brazil is Spanish, just like other South American countries." The teacher says, "I don't understand why the math book has decimals in the same unit as fractions." The teacher has students copy dictionary definitions each week to help his students learn to spell difficult words.	The teacher plans lessons on area and perimeter independently of one another, without linking the concepts together. The teacher plans to forge ahead with a lesson on addition with regrouping, even though some students have not fully grasped place value. The teacher always plans the same routine to study spelling: pretest on Monday, copy the words 5 times each on Tuesday and Wednesday, test on Friday.	The teacher's plan for area and perimeter invites students to determine the shape that will yield the largest area for a given perimeter. The teacher realized her students are not sure how to use a compass, so she plans to practice that before introducing the activity on angle measurement. The teacher plans to expand a unit on civics by having students	In a unit on 19thcentury literature, the teacher incorporates information about the history of the same period. Before beginning a unit on the solar system, the teacher surveys the class on their beliefs about why it is hotter in the summer than in the winter.

1a Critical Attributes

Unsatisfactory	Basic	Proficient	Distinguished
Teacher makes content errors. Teacher does not consider prerequisite relationships when planning. Teacher's plans use inappropriate strategies for the discipline.	Teacher is familiar with the discipline but does not see conceptual relationships. Teacher's knowledge of prerequisite relationships is inaccurate or incomplete. Lesson and unit plans use limited instructional strategies, and some may not be suitable to the content.	The teacher can identify important concepts of the discipline and their relationships to one another. The teacher consistently provides clear explanations of the content. The teacher answers student questions accurately and provides feedback that furthers their learning. The teacher seeks out content-related professional development.	In addition to the characteristics of "proficient": Teacher cites intra- and interdisciplinary content relationships. Teacher is proactive in uncovering student misconceptions and addressing them before proceeding.

APPENDIX B

OUTLINE OF STUDY RESEARCH QUESTIONS

Research Question #1: To What degree does the validity evidence generated by the LEA's policy-directed teacher evaluation system support inferences of Teacher Instructional Quality (TIQ)? This Research Question is defined by five types of component evidence: (A) Criterion, (B) Content, (C) Consequential, (D) Scale Reliability, and (E) Theoretical Construct Definition.

RQ1A. Criterion Evidence

Methodological Approach	Methodological Type	Participants	Data Sources & Collections	Effect Indicators
Correlation, Means Testing	Quantitative (Q)	Data for Teacher/Classroom, Grades 3-8, n ≥ 10 Students, Reg. Ed., Self-Contained	PP Scores (Evaluator); PP Scores (Teacher); VAM Estimates (Teachers)	r, r^2, χ^2, ANOVA, t-Test, F-Tests, Post Hoc Tests of Group Mean Differences

Supporting Research Questions:

Item	Description	Approach	Measure
RQ1A(a):	To what degree do value-added measures of instructional effectiveness correlate with measures of professional practice (PP)?	Correlation	r, r^2
RQ1A(b):	To what degree do measures of PP assigned by qualified evaluators correlate with teacher's self-assessment of PP?	Correlation	r, r^2
RQ1A(c):	To what degree do VAM estimates of instructional effectiveness in reading and mathematics correlate?	Correlation	r, r^2
RQ1A(d):	Do PP sub-scale scores display similar degrees of correlation with VAM measures?	Correlation	r, r^2
RQ1A(e):	To what degree are high, middle, and low VAM estimates of instructional effectiveness able to differentiate PP	ANOVA, Means Testing	t, F tests, Post Hoc Tests of Group Mean Differences

RQ1B. Content Evidence

Methodological Approach	Methodological Type	Participants	Data Sources & Collections	Effect Measures
Lawshe CVI (Q); EFA/CFA (Q); Expert Review (Q/QL)	Quantitative (Q) Qualitative (QL)	Stakeholders (Teachers, Principals, Policy); Expert Reviewers: • Curriculum Developers • Instructional Coaches	Lawshe CVI FFT Survey; Stakeholder Interviews; Stakeholder Questionnaire	Lawshe CVI item/scale threshold criteria; EFA/CFA model fit statistics (factor scores, χ^2, AIC); PP Coded Interview Responses.

Supporting Research Questions:

Item	Description	Approach	Measure
RQ1B (a):	To what degree do empirical ratings of PP correspond with the theoretical FFT construct?	Exploratory & Confirmatory Factor analysis	Factor Extractions; Factor Loadings; χ^2; AIC; α
RQ1B (b):	To what degree does the factor analytic structure of empirically-based PP scores differ between less experienced and more experienced teachers?	Exploratory & Confirmatory Factor analysis	Factor Extractions; Factor Loadings; χ^2; AIC; α
RQ1B (c):	To what degree do the twenty-two elements contained within the theoretical FFT framework adequately represent the latent TIQ construct?	Lawshe CVI Questionnaire, Stakeholder Interviews	CVI; Coded Interview Responses
RQ1B (d):	Do perspectives differ among stakeholders regarding the capacity for VAM and PP measures to adequately represent and differentiate the instructional quality of classroom teacher?	Stakeholder Interviews; Stakeholder Questionnaire	Coded Interview Responses; Questionnaire Item Responses

RQ1C. Consequential Evidence

Methodological Approach	Methodological Type	Participants	Data Sources & Collections	Effect Measures
Stakeholder Interview	Qualitative	• Stakeholders (Teachers, Principals, Policy)	Interview	Coded Interview Responses

Supporting Research Questions:

Item	Description	Approach	Measure
RQ1C (a):	In what way has implementation of the teacher evaluation system affected the PP of classroom teachers (Instruction, student learning, professional capacity building, job satisfaction, etc...)	Semi-Structured Interview	Coded Interview Responses;
RQ1C (b):	Do the perspectives of efficacy and system affect differ across stakeholder groups (Teachers, Principals, and Policy Makers)? By VAM Group?	Semi-Structured Interviews	Coded Interview Responses;

RQ1D. Reliability Evidence

Methodological Approach	Methodological Type	Participants	Data Sources & Collections	Effect Measures
PP Scale and sub-scale Reliability (Internal Consistency); VAM Model fit; Correlation; Standard Errors	Quantitative	Data for Teacher/Classroom, Grades 3-6, n ≥ 10 Students, Reg. Ed., Self-Contained	PP Scores (Evaluator); PP Scores (Teacher); VAM Model Estimates	α, r, r^2, model residual & tests of normality, violations of regression assumptions; Prediction error & Confidence Intervals around estimates

Supporting Research Questions:

Item	Description	Approach	Measure
RQ1D (a):	What are the reliability indices for the PP and VAM scales used to form measures of TIQ?	Scale and sub scale reliability indices; Correlation; measurement error; Tests of Normality	PP Scale and sub-scale item correlations; tests of VAM regression assumptions; Prediction Error (SEM);

RQ1E. Theoretical Construct Articulation

Methodological Approach	Methodological Type	Participants	Data Sources & Collections	Effect Measures
Stakeholder Interview	Qualitative	Stakeholders (Teachers, Principals, Policy)	Interview	Coded Interview Responses;

Supporting Research Questions:

Item	Description	Approach	Measure
RQ1E (a):	What is the theoretical construct definition held by stakeholders regarding high quality teaching?	Semi-Structured Interview	Coded Interview Responses
RQ1E (b)	Do the theoretical construct definitions differ by stakeholder group? By VAM Group?	Semi-Structured Interview	Coded Interview Responses
RQ1E (d)	What are stakeholder perspectives regarding the purpose and intended outcome of teacher evaluation? Does this perspective differ across stakeholder groups?	Semi-Structured Interview	Coded Interview Responses

Research Question #2: How does providing an evolving dialog on validity evidence influence organizational policy decisions regarding implementation of the LEA's teacher evaluation system.

Evidence Category	Methodological Approach	Methodological Type	Participants	Data Sources & Collections	Effect Measures
A. Documentation of Affect	Stakeholder Interview	Qualitative	Stakeholder: • Policy Makers • Teacher Evaluation Committee Members • Professional Development (Decision Makers)	Interview; Research Journal	Documentation of Organizational Decisions

Supporting Research Questions:

Item	Description	Approach	Measure
RQ2 (a):	To what extent do policy-level stakeholders value the collection and review of validity evidences as an important input to the system's on-going development?	Semi-Structured Interviews Research Journal	Coded interview responses; Coded Journal Entries
RQ2 (b)	To what extent does validation evidence prompt changes in organizational decisions regarding system implementation?	Semi-Structured Interviews Research Journal	Coded interview responses; Coded Journal Entries

Research Question #3: To what degree have the perspectives and concerns of stakeholders been incorporated into, and influence, the system's implementation plan?

Evidence Category	Methodological Approach	Methodological Type	Participants	Data Sources & Collections	Effect Measures
A. Documentation of Affect	Stakeholder Interviews	Qualitative	Stakeholder (Teachers & Principals)	Interview	Documentation of Organizational Decisions

Supporting Research Questions:

Item	Description	Approach	Measure
RQ3 (a):	Do stakeholders perceive that policy decisions have been made inclusive of their perspectives?	Semi-Structured Interviews	Coded interview responses
RQ3 (c)	Do stakeholders believe that they have a role in shaping future directions of the evaluation system?	Semi-Structured Interviews	Coded interview responses

Research Question #4: How did the process of engaging in this Action-Research study impact the investigator as a scholarly researcher and organizational leader?

Evidence Category	Methodological Approach	Methodological Type	Participants	Data Sources & Collections	Effect Measures
A. Documentation of Affect	Research Journal	Qualitative	Researcher-Practitioner	Journal Reflections	Documentation of Research Reflections

Supporting Research Questions:

Item	Description	Approach	Measure
RQ4 (a):	What barriers or impediments were encountered during the course of the study? How were they overcome and/or handled?	Research Journal	Coded Journal entries
RQ4 (c)	What were the salient learnings from this study? How did the researcher grow professionally and personally? How will these learnings be incorporated into future research & leadership activities?	Research Journal	Coded Journal entries

APPENDIX C

CONTENT VALIDITY ASSESSMENT QUESTIONNAIRE

Introduction
Thank you for taking time to complete this survey. The purpose of this survey is to obtain your perspective of (1) what it means to be a "Good/Effective" teacher, and (2) which elements in the Danielson Framework best capture/measure these attributes. The survey is divided into three parts:

1. Part I asks you to indicate whether you are currently in an administrative or instructional position including the number of years you have been employed in that context.

2. Part II asks you to briefly reflect on what YOU feel it means to be a "good/effective" teacher.

3. Finally, Part III asks you to identify whether obtaining a measure on a particular FFT element is essential/not essential for identifying a "good/effective" teacher.

All responses are confidential. No personally identifiable information is being requested on this survey. If you have questions concerning this survey, please feel free to contact Ed Sloat at edward.sloat@yahoo.com.

Thank you for your time.

Part I: Background Information:
1. Please indicate which category best describes your current employment position
(Choose one option)

☐ Classroom Teacher
☐ Instructional Support (IGT, Interventionists, etc.)
☐ School Administrator
☐ District Administrator
☐ Other (please specify):

2. Approximately how long have you been in your CURRENT position? Count the 2013-14 school year as 1 full year. (Choose one option)

☐ 1 Year ☐ 2 Years ☐ 3 Years ☐ 4 Years ☐ 5 or More Years

Part II: Define Good/Effective Teacher

3. Please use the space below to briefly outline/describe what YOU believe it means to be a "Good/Effective" teacher. Feel free to insert sentences/paragraphs or simply enter key phrases, concepts, or words that reflect your thinking.

[Space provided for comments]

Part III: Rating the Relevance of the Danielson FFT Elements

Please read the following before continuing:

To complete Part III, recall your answer to the previous question (i.e. attributes YOU believe characterize a good/effective teacher.) Based on your understanding and definition of good/effective teaching, you will be asked to evaluate the importance of each component in the Danielson Framework. For example, are all elements equally important for identifying good/effective teaching or are some more essential than others?

Use the following rubric to guide your responses:

Not Necessary/Not Important: Obtaining a high rating on this item does not necessarily inform on whether the person is a good/effective classroom teacher. This item does not necessarily represent an important attribute of high quality or effective teaching. Omitting this item would NOT significantly detract from efforts to identify a good/effective teacher.

Useful/Important: Obtaining a high rating on this item provides some indication as to whether or not the person is a good/effective teacher. However, it is not an essential measure. Omitting this item may detract from attempts to identify a good/effective teacher.

Essential/Very Important: This item represents an essential attribute of high quality and effective instruction. Omitting this item would SIGNIFICANTLY detract from one's ability to identify a good/effective teacher.

Danielson Domains and Elements: On the table below, please indicate whether obtaining a rating on the specific element is Not Necessary, Useful, or Essential for identifying a good/effective teacher.

Domain 1: Planning and Preparation	*Not Necessary/Not Important*	*Useful/Important*	*Essential/Very Important*
1a. Demonstrating Knowledge of Content and Pedagogy	❑	❑	❑
1b. Demonstrating Knowledge of Students	❑	❑	❑
1c. Setting Instructional Outcomes	❑	❑	❑
1d. Demonstrating Knowledge of Resources	❑	❑	❑
1e. Designing Coherent Instruction	❑	❑	❑
1f. Designing Student Assessments	❑	❑	❑

Domain 2: Classroom Environment	*Not Necessary/Not Important*	*Useful/Important*	*Essential/Very Important*
2a. Creating an Environment of Respect and Rapport	❑	❑	❑
2b. Establishing a Culture for Learning	❑	❑	❑
2c. Managing Classroom Procedures	❑	❑	❑
2d. Managing Student Behavior	❑	❑	❑
2e. Organizing Physical Space	❑	❑	❑

Domain 3: Instruction	*Not Necessary/Not Important*	*Useful/Important*	*Essential/Very Important*
3a. Communicating With Students	❑	❑	❑
3b. Using Questioning and Discussion Techniques	❑	❑	❑
3c. Engaging Students in Learning	❑	❑	❑
3d. Using Assessment in Instruction	❑	❑	❑
3e. Demonstrating Flexibility and Responsiveness	❑	❑	❑

Domain 4: Professional Responsibilities	*Not Necessary/Not Important*	*Useful/Important*	*Essential/Very Important*
4a. Reflecting on Teaching	❑	❑	❑
4b. Maintaining Accurate Records	❑	❑	❑
4c. Communicating with Families	❑	❑	❑
4d. Participating in a Professional Community	❑	❑	❑
4e. Growing and Developing Professionally	❑	❑	❑
4f. Showing Professionalism	❑	❑	❑

THANK YOU FOR PARTICIPATING IN THIS EVALUATION ACTIVITY!

APPENDIX D

SUMMARY OF DATA COLLECTION PROTOCOLS

A. Stakeholder Interview Protocols

Date & Location of Interview: _____
Interviewee Identification: _____
Time: Approximately ½ hour to 45 minutes per session
Location: Face-to-Face; Location - School, office, etc.
Method of Data Collection: Audio recording, interview notes, artifacts (documents, etc…)

Interview Session Activities:

- Introduction to Study (Background, purpose & context, Expected time frame)
- Review of Study Participation Agreement (confidentiality, risk, contact information, signatures)
- Interview Activity
- Discussion of Subsequent Data Processing Activities: Transcription, inclusion in interview data bank, review and coding
- Request for transcript review (Interviewee verification, modifications, clarifications, additions)

Interview Prompts: In your opinion…	Question Alignment
1. What is the purpose of k-12 education in American society?	RQ5
2. What role do teachers serve in this purpose?	RQ5
3. What does it mean to be a 'good teacher'?	RQ1E(a)(b)
4. What is the purpose of Teacher Evaluation?	RQ1E(d)
5. How well does the district's teacher evaluation identify and distinguish between 'good teachers'?	RQ1B(d)(d); RQ1E(c)
o Do the FFT elements cover all aspects of what it means to be a 'good teacher'?	
o Are characteristics of 'good teaching' missing in the FFT Framework?	
o Are test scores a suitable indicator of 'good teaching'?	
o Which would you place more confidence in for identifying a 'good teacher': evaluator observations or test scores? Why?	
6. How has participation in the teacher evaluation process impacted your instructional practice? You personally?	RQ1C(a)(b)
7. How much input do you feel teachers (yourself) have had in the design, implementation, and use of the evaluation system?	RQ3

Guiding Conceptual Structure of Stakeholder Discussions

RQ#(x)	**Interview Prompt**
Content Evidence	
RQ1B (c)	In your opinion, how well does the Danielson FFT components (22) and domains represent what it means to be a "good teacher" o *Does the Danielson FFT elements cover all aspects of teaching and its desired impact on students?* o *Are their characteristics of 'good teaching' that are missing or not well represented in the Danielson FFT?*
Consequential Evidence	
RQ1C (a), (b)	How has the teacher evaluation process impacted your instructional practice (1) as a classroom teacher; (b) as an educator; (c) personally?
Construct Evidence	
RQ1E (a), (b)	In your opinion, what does it mean to be a 'Good Teacher'? o *What impact does a 'Good Teacher' have on students?*
RQ1E (d)	In your opinion, what is the purpose and intended outcome(s) of the teacher evaluation system
RQ1E(c)	The teacher evaluation system combined student achievement and observations of professional practice to arrive at an overall teacher evaluation rating. In your opinion, does this combination of measures adequately encapsulate what it means to be a 'good teacher'? o *Is this combination of measures able to adequately distinguish between levels of instructional quality and how instruction impacts students?*
Construct Definition (Extended)	
RQ5	1. What is the purpose of K-12 education/schools in American society? 2. What role do teachers serve in this purpose? o *Impact teachers have on students?* 3. What is the purpose of teacher evaluation in schools? o *Purpose, outcomes of evaluation process, impact on teacher quality* 4. How well does the AZ/DUSD (legislative criteria from SB1040) approach to evaluating teachers serve this purpose?
Organizational Value of Evidence	
RQ2	In what way have the evidence gathered regarding the teacher evaluation system impacted your perspective of how the system might be improved/changed in the future? o *Consequential Decisions on teacher tenure and retention?* o *Professional development, training, resource allocation?*
Participant Inclusion (Voice)	
RQ3	Regarding the design and implementation of the district's teacher evaluation system, how well have policy decisions incorporated the perspectives, concerns, and suggestions of classroom teachers? o *Have teacher's voices, concerns, and perspectives been adequately represented in the teacher evaluation system design* o *... regarding consequences based on evaluation ratings*

B. Observation Protocol

Date & Location of Interview: _____
Participant Identification: _____
Time: Length of observation session
Location: Location of event
Method of Data Collection: Scripted notes, event artifacts (documents, handouts, presentation materials …)

Reflective Prompts: (Included during and immediately after event)

- Power relationships, gatekeepers, influence brokers
- Issues & challenges in the discussion
- Impact on TEval validity constructs (Pro/con, if any)
- Participant voices, perspectives
- Resolution, outcomes, decisions
- Personal (researcher) reflections & memos

C. Journaling Protocol

Entry Date: _____
Topic Identification: _____
Location: Location that entry is recorded
Participants: If applicable: who is involved in the reflection

Reflective Prompts: (referenced during/after journal entry)

- Positionality, power relationships, gatekeepers, Influence Brokers
- Change Agency (factor, issues, barriers, actions, solutions & resolutions)
- Reflection & personal voice, concerns, …
- Personal growth, learning's, future …

APPENDIX E

PARTICIPANT INFORMED CONSENT FORM

Examining the Validity of a State Policy-Directed Framework for Evaluating Teacher Instructional Quality: Informing Policy, Impacting Practice

Dear _____

I am a doctoral candidate in the Mary Lou Fulton Teachers College Leadership and Innovation Program at Arizona State University. In addition, I serve as the Director of Research and Accountability within the [Name]. For my doctoral dissertation, I am examining the technical attributes of a new state-policy directed teacher evaluation system as implemented within the [Name]. I would like to personally invite to participate in my study.

Purpose: The purpose of my study is to assess the validity evidence associated with implementing Arizona's teacher evaluation framework. Included are perspectives held by various stakeholder groups such as classroom teachers, school administrators, and district/state policy makers. Findings will directly inform on local policy decisions and help improve future implementations of the evaluation framework. It is also hoped that results will be provide useful understandings of implementation issues faced at the state and national level.

Participation: If you agree to participate, I would like to conduct an interview lasting approximately 45 minutes regarding your perspectives of teacher evaluation. With your permission, I would like to audio record the interview. However, I will not do so without your explicit prior approval and you may change your mind even after the interview is in progress. Participants in the study will also be asked to complete a survey requiring approximately 15-20 minutes of your time. Finally, if you are a classroom teacher, I may ask to observe one of your classroom instructional activities. I would be glad to provide you with advanced copies of the interview questions and survey questionnaire.

Voluntary Participation: You are under no obligation to participate in this study. If you choose to participate, you may withdraw at any time, for any reason, without providing prior notification to this researcher. Similarly, you may elect to not to answer any specific interview or questionnaire item. At any time before, you may direct me not to utilize any/all of the information you provided and to delete/destroy any/all of your data. There are no negative consequences related to your choice to participate and/or withdraw from any aspect of this study.

Confidentiality: All of the information you provide will be strictly confidential. I will not disclose your participation or any information you may provide. All data/information will be combined with the responses from other participants and analyzed/reported in aggregate form. No names, position titles, locations, or references that may be attributed to you will be released and/or published in any form. All electronic data will be maintained in secure password protected electronic files. Written notes, questionnaire responses, or other artifacts and documents will be maintained by this researcher in a secure location.

Dissemination of Information: This study will be published as a doctoral dissertation and reside in the public realm. In addition, copies of the dissertation will be provided to district policy makers and all stakeholders participating in the research. Finally, data/information collected as part of the study may be utilized/published in subsequent articles, papers, and/or reports.

Risks to Participant: The only foreseeable risk associated with this study might be the inadvertent disclosure of personally identifiable information through theft or oversight. However, as discussed above every effort is being made to eliminate this possibility.

I would be glad to answer any questions you may have regarding any aspect of the research and/or your contribution. You may contact me directly at esloat@asu.edu (phone number). In addition, you may also direct questions about your rights as a subject/participant in this research study, or if you feel you have been placed at risk, by contacting the following individuals:

- Chair of the Human Subjects Institutional Review Board, through the ASU Office of Research Integrity and Assurance, at [Phone Number]
- The study's Principal Investigator, Dr. Keith Wetzel, Professor, Arizona State University, at [Phone Number] (Keith.Wetzel@asu.edu)
- Ms. [Name], Director Human Resources, [Name], at [Phone Number]

I thank you for your consideration.

Sincerely,

Edward F. Sloat
esloat@asu.edu
[Phone Number]

****

Your signature below indicates that you consent to participate in the above study.

_____	_____	_____
Subject's Signature	Printed Name	Date

INVESTIGATOR'S STATEMENT

"I certify that I have explained to the above individual the nature, purpose, potential benefits, and possible risks associated with participation in this research study. In addition, I have answered all questions and/or concerns raised, and have witnessed the above signature. These elements of Informed Consent conform to the Assurance given by Arizona State University to the Office for Human Research Protections to protect the rights of human subjects. Finally, I have provided the subject/participant a copy of this signed consent document."

Signature of Investigator_____ Date_____

APPENDIX F

INFORMATION LETTER-INTERVIEWS

Examining the Validity of a State Policy-Directed Framework for Evaluating Teacher Instructional Quality: Informing Policy, Impacting Practice

Date: July 19, 2013

Dear _____:

I am a doctoral candidate under the direction of Professor Keith Wetzel in the Mary Lou Fulton Teachers College Leadership and Innovation Program at Arizona State University. I am conducting a research study to examine the technical attributes of a new state-policy directed teacher evaluation system as implemented within the [Name].

I am inviting your participation, which will involve an interview lasting approximately 45 minutes regarding your perspectives of teacher evaluation. Participants in the study may also be asked to complete a related survey requiring approximately 15-20 minutes of your time. Finally, if you are a classroom teacher, I may ask to observe one of your classroom instructional activities. You have the right not to answer any question, and to stop the interview/survey/observation at any time.

Your participation in this study is voluntary. If you choose not to participate or to withdraw from the study at any time, there will be no penalty. You must be 18 or older to participate in the study.

Your participation will help build an understanding of the perspectives held by various stakeholder groups such as classroom teachers, school administrators, and district/state policy makers. Findings will directly inform local policy decisions and help improve future implementations of the evaluation framework. There are no foreseeable risks or discomforts to your participation.

All of the information you provide will be strictly confidential. I will not disclose your participation or any information you may provide. All data/information will be combined with the responses from other participants and analyzed/reported in aggregate form. No names, position titles, locations, or references that may be attributed to you will be released and/or published in any form. All electronic data will be maintained in secure password protected electronic files. Written notes, questionnaire responses, or other artifacts and documents will be maintained by this researcher in a secure location. Results from this study may be used in reports, presentations, and/or publications. Again, your name will not be disclosed.

I would like to make digital/audio recordings of the interview. However, the interview will not be recorded without your permission. Please let me know if you do not want the interview to be taped; you may also change your mind after the interview starts, just let me know. All audio recordings will be kept in a secure location accessible only to me. All audio recordings will be deleted/erased at the completion of the study (approximately June 2014).

If you have any questions concerning the research study, please contact the research team at:

Co-Researcher:	Ed Sloat @ [Phone Number] or esloat@asu.edu
Principal Investigator:	Dr. Keith Wetzel, Professor, Arizona State University @ [Phone Number] or Keith.Wetzel@asu.edu

If you have any questions about your rights as a subject/participant in this research, or if you feel you have been placed at risk, you can contact the Chair of the Human Subjects Institutional Review Board, through the ASU Office of Research Integrity and Assurance, at [Phone Number]. Please let me know if you wish to be part of the study

APPENDIX G

ALIGNMENT BETWEEN SELECTED RESEARCH QUESTIONS AND QUALITATIVE DATA SOURCES:

INTERVIEW PROMPTS, RESEARCH JOURNAL, OBSERVATIONS

Research Question 1: To What degree does the validity evidence generated by the LEA's policy-directed teacher evaluation system support inferences of Teacher Instructional Quality (TIQ)?

RQ: ID	RQ: Description	Data Collection Method	Interview Prompts (Generalized)
RQ1B (c):	To what degree do the twenty-two elements contained within the theoretical FFT framework adequately represent the latent TIQ construct?	Lawshe CVI Questionnaire, Stakeholder Interviews	To what extent do the Danielson FFT elements adequately represent what it means to be a 'good teacher'? Are missing characteristics in the Danielson FFT Framework?
RQ1B (d):	Do perspectives differ among stakeholders regarding the capacity for VAM and PP measures to adequately represent and differentiate the instructional quality of a classroom teacher?	Stakeholder Interviews	How well does the district's (Arizona's) teacher evaluation framework identify and distinguish 'good teachers'? Please Explain? Taken on their own, are test scores a suitable indicator of 'good teaching'? Which would you place more confidence in for identifying a 'good teacher': evaluator observations or test scores? Why?
RQ1C (a):	In what way has implementation of the teacher evaluation system affected the PP of classroom teachers?	Semi-Structured Interview	How has participation in the teacher evaluation process impacted your Professional/Instructional practice? You personally? In what way is Arizona's teacher evaluation process impacting the Professional/Instructional practices of classroom teachers?
RQ1C (b):	Do the perspectives of efficacy and system affect differ across stakeholder groups (Teachers, Principals, and Policy Makers)?	Semi-Structured Interviews	How has participation in the teacher evaluation process impacted your Professional/Instructional practice?
RQ1E (a):	What is the theoretical construct (definition) held by stakeholders regarding high quality teaching?	Semi-Structured Interview	What is the purpose of k-12 education in American society? What role do teachers serve in this purpose? What does it mean to be a 'good teacher'? How well does the Arizona/district's teacher evaluation identify and distinguish between 'good teachers'?
RQ1E (b)	Does the theoretical construct (definition) regarding high quality teaching differ by stakeholder group?	Semi-Structured Interview	What is the purpose of k-12 education in American society? What role do teachers serve in this purpose? What does it mean to be a 'good teacher'?
RQ1E (c)	To what degree do stakeholders believe that the empirical measures derived from the teacher evaluation system align with the theoretical construct of quality teaching?	Semi-Structured Interview	What does it mean to be a 'good teacher'? How well does the Arizona/district's teacher evaluation identify and distinguish between 'good teachers'?
RQ1E (d)	What are stakeholder perspectives regarding the purpose and intended outcome of teacher evaluation? Does this perspective differ across stakeholder groups?	Semi-Structured Interview	What is the purpose of Teacher Evaluation?

Research Question 2: How does providing an evolving dialog on validity evidence influence organizational policy decisions regarding implementation of the LEA's teacher evaluation system.

RQ: ID	RQ: Description	Data Collection Method	Interview Prompts (Generalized)
RQ2 (a):	To what extent do policy-level stakeholders value the collection and review of validity evidences as an important input to the system's on-going development?	Semi-Structured Interviews Research Journal	Research Journal; Researcher observations; Interview Prompt: Do you feel that the implementation decisions were aided/influenced by the availability of on-going empirical analysis?
RQ2 (b)	To what extent does validation evidence prompt changes in organizational decisions regarding system implementation?	Semi-Structured Interviews Research Journal	Research Journal; Researcher observations; Interview Prompt: Do you feel that the implementation decisions were aided/influenced by the availability of empirical analysis?

Research Question 3: To what degree have the perspectives and concerns of stakeholders been incorporated into, and influence, the system's implementation plan?

RQ: ID	RQ: Description	Data Collection Method	Interview Prompts (Generalized)
RQ3 (a):	Do stakeholders perceive that policy decisions have been made inclusive of their perspectives?	Semi-Structured Interviews	How much input do you feel Arizona Educators (teachers, administrators) have had in the design, implementation, and application of the district's (Arizona's) evaluation system?
~~RQ3 (b)~~	~~Do stakeholders harbor concerns over the system's validity, integrity, and/or its ability to fairly and accurately assess their instructional competence?~~	~~Semi-Structured Interviews~~	~~How well does the Arizona/district's teacher evaluation identify and distinguish between 'good teachers'?~~
~~RQ3 (c)~~	~~Do stakeholders believe that they have a role in shaping future directions of the evaluation system?~~	~~Semi-Structured Interviews~~	~~How much input do you feel Arizona Educators (teachers, administrators) have had in the design, implementation, and application of the district's (Arizona's) evaluation system?~~

Research Question 4: How did the process of engaging in this Action-Research study impact the investigator as a scholarly researcher and organizational leader?

RQ: ID	RQ: Description	Data Collection Method	Interview Prompts (Generalized)
RQ4 (a):	What barriers or impediments were encountered during the course of the study? How were they overcome and/or handled?	Research Journal	Coded Research Journal entries, Researcher observations
RQ4 (b)	What aspects of the action-research process provided the most growth and/or learning for the Researcher-Practitioner?	Research Journal	Coded Research Journal entries , Researcher observations
RQ4 (c)	What were the salient learnings from this study? How did the researcher grow professionally and personally? How will these learnings be incorporated into future research & leadership activities?	Research Journal	Coded Research Journal entries, Researcher observation

APPENDIX H

DESCRIPTIVE SUMMARY OF QUALITATIVE COMPONENTS

Stakeholder Group	Individual	Audio Length	Transcription (Words)	Co-Research Transcript Review
District Policy	301	32:51:	5,099	Y
	302	32:17	3,804	
	303	38:09	5,897	Y
	304	27:50	3,999	
Subtotals:	**N=4**	**131:07**	**18,799**	
State Policy	401 (Supt PI)	56:48	9,183	
	402 (ASA/Gov)	56:08	7,324	Y
	403 (SBE)	47:01	6,783	Y
Subtotals:	**N=3**	**159:57**	**23,290**	
Principal/Evaluator	201	39:35	6,651	
	202	29:53	4,561	Y
	203	31:18	4,586	
	204	59:21	6,876	
	205	31:11	5,240	Y
	206	29:24	4,202	
	207	54:25	7,809	
	208	40:03	5,733	
Subtotals:	**N=8**	**315:10**	**45,658**	
Teacher (Group A)	101	39:37	5,538	
	102	26:26	3,714	Y
	103	31:07	4,639	
	104	44:53	6,922	
	105	64:56	9,850	
	106	27:19	4,281	
	107	Pending	Pending	
	108	41:17	6,645	Y
	109	Pending	Pending	
Subtotals:	**N=7**	**275:35**	**41,589**	
	Placeholder	**Length**	**Words**	
IGT (Survey)	1	na	119	Y
	2	na	89	Y
	3	na	28	Y
	4	na	186	Y
	5	na	58	Y
	6	na	57	Y
	7	na	287	Y
	8	na	354	Y
	9	na	55	Y
	10	na	25	Y
	11	na	91	Y
	12	na	108	Y
Subtotals:	**n=12 (N=24)**		**1457**	
TEval Committee (Survey)	1	na	103	Y
	2	na	173	Y
	3	na	86	Y
	4	na	124	Y
	5	na	97	Y
	6	na	158	Y
Subtotals:	**N=6 (N=12)**		**741**	
Grand Totals:		**882:42 Min./Sec 15.24 Hrs./Min.**	**83,162**	

APPENDIX I

POLYCHORIC CORRELATIONS: FFT ELEMENTS

	D1A	D1B	D1C	D1D	D1E	D1F	D2A	D2B	D2C	D2D	D2E	D3A	D3B	D3C	D3D	D3E	D4A	D4B	D4C	D4D	D4E	D4F
D1A	1																					
D1B	0.73	1																				
D1C	0.74	0.67	1																			
D1D	0.69	0.76	0.67	1																		
D1E	0.82	0.62	0.79	0.63	1																	
D1F	0.68	0.78	0.73	0.79	0.72	1																
D2A	0.69	0.72	0.55	0.59	0.65	0.56	1															
D2B	0.76	0.70	0.74	0.69	0.67	0.68	0.74	1														
D2C	0.70	0.58	0.74	0.62	0.69	0.67	0.70	0.75	1													
D2D	0.59	0.62	0.70	0.60	0.59	0.58	0.64	0.63	0.69	1												
D2E	0.45	0.61	0.60	0.73	0.56	0.66	0.63	0.65	0.56	0.57	1											
D3A	0.69	0.71	0.71	0.67	0.68	0.72	0.70	0.80	0.71	0.59	0.61	1										
D3B	0.75	0.58	0.81	0.68	0.81	0.66	0.58	0.74	0.70	0.69	0.50	0.67	1									
D3C	0.77	0.63	0.83	0.64	0.83	0.68	0.66	0.77	0.71	0.66	0.50	0.65	0.79	1								
D3D	0.72	0.72	0.79	0.68	0.70	0.85	0.61	0.71	0.67	0.68	0.63	0.55	0.72	0.67	1							
D3E	0.66	0.70	0.64	0.77	0.63	0.74	0.56	0.62	0.37	0.52	0.56	0.57	0.62	0.64	0.68	1						
D4A	0.75	0.69	0.61	0.65	0.61	0.62	0.62	0.67	0.48	0.53	0.58	0.59	0.53	0.58	0.59	0.63	1					
D4B	0.65	0.68	0.58	0.62	0.54	0.70	0.55	0.66	0.58	0.52	0.48	0.58	0.58	0.55	0.62	0.60	0.62	1				
D4C	0.59	0.65	0.57	0.64	0.55	0.70	0.47	0.60	0.55	0.46	0.58	0.56	0.55	0.47	0.63	0.57	0.60	0.76	1			
D4D	0.55	0.59	0.59	0.62	0.51	0.65	0.43	0.56	0.53	0.50	0.59	0.47	0.50	0.51	0.70	0.46	0.52	0.65	0.64	1		
D4E	0.77	0.75	0.73	0.79	0.66	0.74	0.67	0.69	0.54	0.50	0.62	0.56	0.65	0.66	0.68	0.72	0.78	0.68	0.66	0.68	1	
D4F	0.70	0.69	0.59	0.69	0.61	0.63	0.71	0.66	0.52	0.49	0.67	0.69	0.56	0.55	0.66	0.66	0.74	0.72	0.65	0.74	0.80	1

APPENDIX J

STANDARD ERRORS OF POLYCHORIC CORRELATIONS: FFT ELEMENTS

	D1A	D1B	D1C	D1D	D1E	D1F	D2A	D2B	D2C	D2D	D2E	D3A	D3B	D3C	D3D	D3E	D4A	D4B	D4C	D4D	D4E	D4F
D1A																						
D1B	0.056																					
D1C	0.052	0.063																				
D1D	0.062	0.053	0.067																			
D1E	0.042	0.07	0.047	0.072																		
D1F	0.061	0.048	0.059	0.053	0.058																	
D2A	0.067	0.057	0.074	0.076	0.062	0.08																
D2B	0.048	0.054	0.051	0.06	0.059	0.048	0.054															
D2C	0.062	0.069	0.057	0.069	0.063	0.064	0.06	0.051														
D2D	0.078	0.068	0.063	0.078	0.076	0.079	0.059	0.067	0.063													
D2E	0.077	0.075	0.061	0.068	0.067	0.077	0.074	0.07	0.076	0.083												
D3A	0.061	0.057	0.059	0.067	0.061	0.058	0.046	0.044	0.058	0.073	0.075											
D3B	0.054	0.065	0.049	0.066	0.047	0.054	0.066	0.048	0.057	0.067	0.073	0.06										
D3C	0.049	0.068	0.041	0.07	0.039	0.063	0.061	0.048	0.06	0.067	0.068	0.066	0.048									
D3D	0.056	0.056	0.049	0.068	0.063	0.041	0.073	0.051	0.064	0.066	0.08	0.075	0.06	0.065								
D3E	0.068	0.059	0.05	0.054	0.07	0.059	0.077	0.065	0.079	0.067	0.085	0.076	0.051	0.068	0.066							
D4A	0.053	0.059	0.073	0.069	0.071	0.072	0.073	0.06	0.084	0.068	0.062	0.056	0.064	0.074	0.077	0.07						
D4B	0.067	0.041	0.061	0.072	0.08	0.064	0.077	0.059	0.073	0.068	0.075	0.071	0.062	0.08	0.058	0.076	0.068					
D4C	0.062	0.048	0.057	0.066	0.067	0.061	0.082	0.059	0.074	0.076	0.082	0.073	0.061	0.062	0.059	0.077	0.07	0.051				
D4D	0.078	0.07	0.062	0.075	0.065	0.069	0.083	0.064	0.067	0.086	0.079	0.074	0.07	0.067	0.064	0.089	0.077	0.067	0.063			
D4E	0.051	0.052	0.058	0.05	0.068	0.057	0.067	0.057	0.076	0.068	0.058	0.058	0.055	0.066	0.063	0.059	0.049	0.063	0.064	0.064		
D4F	0.049	0.058	0.057	0.064	0.055	0.069	0.059	0.061	0.074	0.074	0.066	0.06	0.063	0.064	0.067	0.066	0.054	0.057	0.064	0.053	0.046	

APPENDIX K

Z-VALUE FOR POLYCHORIC CORRELATIONS: FFT ELEMENTS

	D1A	D1B	D1C	D1D	D1E	D1F	D2A	D2B	D2C	D2D	D2E	D3A	D3B	D3C	D3D	D3E	D4A	D4B	D4C	D4D	D4E	D4F
D1A																						
D1B	12.96																					
D1C	14.21	10.70																				
D1D	11.15	14.36	10.01																			
D1E	19.52	8.79	16.89	8.76																		
D1F	11.08	16.33	12.31	14.83	12.48																	
D2A	10.27	12.68	7.43	7.76	10.47	7.00																
D2B	15.79	12.98	14.55	11.48	11.41	14.08	13.70															
D2C	11.29	8.33	12.89	9.03	11.00	10.47	11.62	14.61														
D2D	7.53	9.09	11.03	7.67	7.82	7.28	10.76	9.34	10.90													
D2E	5.88	8.13	9.79	10.72	8.40	8.60	8.54	9.24	7.33	6.89												
D3A	11.33	12.39	12.00	9.94	11.18	12.33	15.22	18.14	12.22	8.08	8.09											
D3B	13.91	8.92	16.43	10.26	17.19	12.22	8.77	15.33	12.25	10.34	6.82	11.08										
D3C	15.69	9.32	20.24	9.13	21.38	10.78	10.75	15.98	11.85	9.78	7.28	9.83	16.42									
D3D	12.79	12.88	16.08	9.93	11.03	20.66	8.32	13.96	10.45	10.26	7.88	7.32	11.98	10.28								
D3E	9.75	11.81	12.76	14.19	9.03	12.53	7.22	9.54	4.68	7.75	6.55	7.49	12.20	9.41	10.26							
D4A	14.08	11.76	8.32	9.38	8.56	8.63	8.52	11.18	5.75	7.84	9.29	10.50	8.27	7.86	7.70	8.97						
D4B	9.70	16.63	9.51	8.65	6.80	10.95	7.19	11.24	7.89	7.59	6.35	8.13	9.40	6.84	10.64	7.87	9.18					
D4C	9.56	13.52	10.04	9.70	8.24	11.49	5.76	10.10	7.39	6.07	7.05	7.68	9.08	7.61	10.59	7.35	8.61	14.92				
D4D	7.03	8.40	9.52	8.27	7.82	9.41	5.18	8.72	7.93	5.86	7.43	6.32	7.16	7.54	10.95	5.19	6.71	9.66	10.21			
D4E	15.18	14.38	12.60	15.80	9.65	13.00	9.94	12.09	7.04	7.38	10.66	9.62	11.89	10.02	10.84	12.25	15.92	10.79	10.36	10.61		
D4F	14.37	11.81	10.33	10.80	11.04	9.16	12.03	10.84	6.97	6.66	10.12	11.47	8.86	8.53	9.78	10.03	13.76	12.58	10.14	14.04	17.33	

APPENDIX L

UNCORRELATED CFA MODEL RESIDUAL COVARIANCE MATRIX

	D1A	D1B	D1C	D1D	D1E	D1F	D2A	D2B	D2C	D2D	D2E	D3A	D3B	D3C	D3D	D3E	D4A	D4B	D4C	D4D	D4E
D1A																					
D1B	-0.01																				
D1C	0.00	-0.04																			
D1D	-0.04	0.05	-0.04																		
D1E	0.06	-0.12	0.05	-0.10																	
D1F	-0.08	0.05	-0.01	0.06	-0.04																
D2A	0.69	0.72	0.55	0.59	0.65	0.56															
D2B	0.76	0.70	0.74	0.69	0.67	0.68	0.01														
D2C	0.70	0.58	0.74	0.62	0.69	0.67	-0.01	0.01													
D2D	0.59	0.62	0.70	0.60	0.59	0.58	-0.01	-0.04	0.04												
D2E	0.45	0.61	0.60	0.73	0.56	0.66	0.03	0.02	-0.05	0.02											
D3A	0.69	0.71	0.71	0.67	0.68	0.72	0.70	0.80	0.71	0.59	0.61										
D3B	0.75	0.58	0.81	0.68	0.81	0.66	0.58	0.74	0.70	0.69	0.50	0.02									
D3C	0.77	0.63	0.83	0.64	0.83	0.68	0.66	0.77	0.71	0.66	0.50	0.01	0.02								
D3D	0.72	0.72	0.79	0.68	0.70	0.85	0.61	0.71	0.67	0.68	0.63	-0.05	0.01	-0.03							
D3E	0.66	0.70	0.64	0.77	0.63	0.74	0.56	0.62	0.37	0.52	0.56	0.02	-0.04	-0.01	0.07						
D4A	0.75	0.69	0.61	0.65	0.61	0.62	0.62	0.67	0.48	0.53	0.58	0.59	0.53	0.58	0.59	0.63					
D4B	0.65	0.68	0.58	0.62	0.54	0.70	0.55	0.66	0.58	0.52	0.48	0.58	0.58	0.55	0.62	0.60	-0.04				
D4C	0.59	0.65	0.57	0.64	0.55	0.70	0.47	0.60	0.55	0.46	0.58	0.56	0.55	0.47	0.63	0.57	-0.04	0.09			
D4D	0.55	0.59	0.59	0.62	0.51	0.65	0.43	0.56	0.53	0.50	0.59	0.47	0.50	0.51	0.70	0.46	-0.11	0.00	0.02		
D4E	0.77	0.75	0.73	0.79	0.66	0.74	0.67	0.69	0.54	0.50	0.62	0.56	0.65	0.66	0.68	0.72	0.07	-0.06	-0.05	-0.01	
D4F	0.70	0.69	0.59	0.69	0.61	0.63	0.71	0.66	0.52	0.49	0.67	0.69	0.56	0.55	0.66	0.66	0.02	-0.03	-0.07	0.05	0.01

APPENDIX M

CORRELATED CFA MODEL RESIDUAL COVARIANCE MATRIX

	D1A	D1B	D1C	D1D	D1E	D1F	D2A	D2B	D2C	D2D	D2E	D3A	D3B	D3C	D3D	D3E	D4A	D4B	D4C	D4D	D4E	D4F
D1A																						
D1B	-0.049																					
D1C	-0.057	-0.103																				
D1D	-0.085	0.004	-0.107																			
D1E	0.042	-0.144	0.013	-0.129																		
D1F	-0.119	0.008	-0.071	0.009	-0.055																	
D2A	0.029	0.079	-0.111	-0.055	0.003	-0.1																
D2B	0.03	-0.009	0.012	-0.022	-0.04	-0.053	0.001															
D2C	0.043	-0.065	0.077	-0.019	0.049	0.012	0.03	0.009														
D2D	-0.027	0.018	0.079	-0.002	-0.008	-0.04	0.011	-0.063	0.067													
D2E	-0.156	0.016	-0.013	0.134	-0.034	0.053	0.013	-0.035	-0.058	-0.004												
D3A	0.01	0.041	0.026	0.001	0.014	0.033	0.061	0.094	0.073	-0.004	0.017											
D3B	0.058	-0.097	0.11	-0.001	0.128	-0.034	-0.072	0.018	0.05	0.087	-0.103	-0.069										
D3C	0.054	-0.063	0.113	-0.059	0.133	-0.037	-0.014	0.027	0.043	0.031	-0.124	-0.108	0.017									
D3D	-0.006	0.017	0.065	-0.03	-0.013	0.124	-0.07	-0.035	-0.005	0.047	0.005	-0.215	-0.059	-0.134								
D3E	0.017	0.067	-0.009	0.136	-0.001	0.092	-0.049	-0.049	-0.233	-0.045	-0.002	-0.114	-0.074	-0.077	-0.047							
D4A	0.087	0.05	-0.054	0.003	-0.038	-0.039	0.048	0.038	-0.088	-0.002	0.046	0.009	-0.061	-0.026	-0.021	0.079						
D4B	-0.005	0.042	-0.076	-0.017	-0.098	0.046	-0.016	0.033	0.008	-0.015	-0.051	0.002	-0.003	-0.057	0.007	0.052	-0.06					
D4C	-0.034	0.037	-0.056	0.028	-0.062	0.074	-0.074	-0.007	0.003	-0.047	0.074	0.01	-0.007	-0.106	0.041	0.044	-0.052	0.11				
D4D	-0.056	-0.001	-0.015	0.031	-0.084	0.045	-0.096	-0.022	0.007	0.015	0.102	-0.062	-0.039	-0.051	0.139	-0.04	-0.114	0.02	0.043			
D4E	0.046	0.037	0.001	0.078	-0.058	0.012	0.032	-0.011	-0.097	-0.089	0.032	-0.081	0.002	-0.011	0.005	0.117	0.019	-0.076	-0.06	-0.018		
D4F	0.013	0.01	-0.103	0.016	-0.07	-0.06	0.109	-0.003	-0.083	-0.067	0.113	0.081	-0.061	-0.091	0.011	0.087	0.02	-0.001	-0.038	0.083	-0.001	

APPENDIX N

MPLUS CFA SPECIFICATION CODE

Uncorrelated CFA Model

```
TITLE: FFT CFA Uncorrelated
DATA: File is FFT for Mplus.csv;

VARIABLE:
Names are
                    schooid StatusID AdminID
                    D1a D1b D1c D1d D1e D1f
                    D2a D2b D2c D2d D2e
                    D3a D3b D3c D3d D3e
                    D4a D4b D4c D4d D4e D4f;

Usevariables are
                    D1a D1b D1c D1d D1e D1f
                    D2a D2b D2c D2d D2e
                    D3a D3b D3c D3d D3e
                    D4a D4b D4c D4d D4e D4f;

Categorical are
                    D1a D1b D1c D1d D1e D1f
                    D2a D2b D2c D2d D2e
                    D3a D3b D3c D3d D3e
                    D4a D4b D4c D4d D4e D4f;

ANALYSIS:

MODEL:

Domain1 by  D1a* D1b D1c D1d D1e D1f;
Domain2 by  D2a* D2b D2c D2d D2e;
Domain3 by  D3a* D3b D3c D3d D3e;
Domain4 by  D4a* D4b D4c D4d D4e D4f;

! Set scale on factors
Domain1@1; Domain2@1; Domain3@1; Domain4@1;

!specify no factor covariance
                    Domain1 with Domain2@0;
                    Domain1 with Domain3@0;
                    Domain1 with Domain4@0;
                    Domain2 with Domain3@0;
                    Domain2 with Domain4@0;
                    Domain3 with Domain4@0;

OUTPUT: sampstat stand (STDYX) Res mod;
PLOT: Type = plot1 plot2 plot3;
```

Correlated CFA Model

```
TITLE: FFT CFA Uncorrelated
DATA: File is FFT for Mplus.csv;

VARIABLE:
Names are
                    schooid StatusID AdminID
                    D1a D1b D1c D1d D1e D1f
                    D2a D2b D2c D2d D2e
                    D3a D3b D3c D3d D3e
                    D4a D4b D4c D4d D4e D4f;

Usevariables are
                    D1a D1b D1c D1d D1e D1f
                    D2a D2b D2c D2d D2e
                    D3a D3b D3c D3d D3e
                    D4a D4b D4c D4d D4e D4f;

Categorical are
                    D1a D1b D1c D1d D1e D1f
                    D2a D2b D2c D2d D2e
                    D3a D3b D3c D3d D3e
                    D4a D4b D4c D4d D4e D4f;

ANALYSIS:

MODEL:

Domain1 by  D1a* D1b D1c D1d D1e D1f;
Domain2 by  D2a* D2b D2c D2d D2e;
Domain3 by  D3a* D3b D3c D3d D3e;
Domain4 by  D4a* D4b D4c D4d D4e D4f;

! Set scale on factors
Domain1@1; Domain2@1; Domain3@1; Domain4@1;

Domain1 with Domain3@.90;

OUTPUT: tech4 sampstat stand (STDYX) Res mod;

PLOT: Type = plot1 plot2 plot3;
```

APPENDIX O

PERMISSION TO CONDUCT ORGANIZATIONAL RESEARCH

August 29, 2013

RE: Permission to Conduct Research in

Dissertation Topic: Examining the Validity of a State Policy-Directed Framework for Evaluating Teacher Instructional Quality: Informing Policy, Impacting Practice

As you may know, I am currently a doctoral candidate in the Leadership and Innovation (L&I) program within the Mary Lou Fulton Teachers College, Arizona State University. This fall, I am entering into the dissertation phase of my studies and would like your permission to conduct research within the ASU's L&I program emphasizes the conduct of scholarly research within organizational settings for the purpose of informing and improving innovations, program, and interventions. My on-going work with leadership team to design and implement the computational framework supporting the new teacher evaluation system aligns with these goals.

Specifically, my research examines the implementation of Arizona's newly mandated teacher evaluation framework. I am interested in examining the construct validity underlying the state-policy framework originally specified in Senate Bill 1040 as interpreted by the State Board of Education. Districts throughout the state are engaging in efforts similar to to fully comply with this mandate. I am hoping that results from my research will not only help inform/improve development of system but also contribute to the broader state policy context.

My research questions focus on two distinct areas of inquiry: the characteristics of the quantitative measures (achievement & rating scores) utilized to represent teacher instructional quality and the latent "Instructional Quality" construct that the quantitative measures attempt to capture. The analytic activities focus on exploring the reliability and association characteristics of the Value-Added models used to estimate student academic growth along with the rating measures obtained from evaluator observations of teacher professional practice. The latent construct investigation involves examining stakeholder's perspectives of (1) the meaning of the term *Instructional Quality*, and (2) the processes of quantitatively representing this construct using the state-prescribed components.

Research Activity: To conduct my research, I am requesting permission to do the following activities:

1. Access and analyze the Value-Added (student growth) and related student demographic information used to estimate the statistical evaluation models maintained in the district's research-assessment data system
2. Access and analyze teacher professional practice ratings maintained in the district's evaluation system database
3. Conduct one-on-one interviews with a random sample of selected teachers, principals, and district office staff that are connected with the evaluation framework. During this activity, I would also ask the identified stakeholders to complete a short survey related to teacher evaluation.

Voluntary Participation: In keeping with establish guidelines and ethics for conducting human subjects research, staff participation in interview/survey activities will be completely voluntary based on informed consent. To ensure this, I will provide each selected participant detailed written communication on the purpose and intent of the study, assurances of confidentiality, and use of the data in the research process.

Confidentiality: All of the information will be protected and confidential. No individual teacher or stakeholder's name will be released in any form including publication of my final dissertation, conference presentations, or any subsequent publication. Confidentiality of information will be protected from loss, theft, or other unexpected release by utilizing codebooks and cross-referencing identifiers unique to this study. Codebooks and reference identifiers will be kept in separate electronic and physical locations from the raw data. Student and/or teacher names will be excluded from this project's data files. In addition, no reference to the will be made in any publication or presentation and the employment status of this researcher will not be explicitly linked to the district.

I hope you will consider my request to conduct research within the . I would be glad to answer any questions you might have. In addition, I will be glad to provide more detailed documentation of the research design, research questions, and analytic plan for your review. I can also provide you a copy of ASU's Human Subjects Research IRB report and approval.

Thank you for your consideration.

Edward F Sloat, Doctoral Candidate
Leadership & Innovation
Mary Lou Fulton Teachers College
Arizona State University

Approval to Conduct Research within the District:

My signature below represents approval to conduct research within the within the framework outline in this communication. It is understood that no personally identifiable information will be released in any form. In addition, no connection of the data, results, or interpretive findings will made to the

_____ Date: 8/29/13

Superintendent

APPENDIX P

INSTITUTIONAL REVIEW BOARD APPROVAL LETTER

ASU Knowledge Enterprise Development

Office of Research Integrity and Assurance

To:	Keith Wetzel FAB
From:	for Mark Roosa, Chair pm Soc Beh IRB
Date:	07/24/2013
Committee Action:	**Exemption Granted**
IRB Action Date:	07/24/2013
IRB Protocol #:	1307009409
Study Title:	Evaluating the Validity Evidence for Systems of Teacher Instructional Quality (TIQ) Implemented Ur Policy-Directed Evaluation Framework

The above-referenced protocol is considered exempt after review by the Institutional Review Board pursuant to Federal regulations, 45 CFR Part 46.101(b)(1) .

This part of the federal regulations requires that the information be recorded by investigators in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects. It is necessary that the information obtained not be such that if disclosed outside the research, it could reasonably place the subjects at risk of criminal or civil liability, or be damaging to the subjects' financial standing, employability, or reputation.

You should retain a copy of this letter for your records.

APPENDIX Q

ANALYTIC SKILL SETS LEVERAGED FOR STUDY

I. Data Processing and Application Programming

A. **Microsoft SQL Server 2012**:

- Database construction, management, & processing; Microsoft Visual Studio 2012; SQL Server Management Studio; extensive T-SQL Programming for data manipulation and processing.

… These tools served as a foundation for constructing, storing, and manipulating the quantitative data used to compute teacher performance measures. This included development of student demographic and achievement information assembled across three academic school years and the assembly of teacher professional practice ratings resultant from principal-evaluators based on the Danielson Framework. Microsoft SQL Server and related tools and utilities served as the primary environment by which to integrate highly diverse raw data tabulations into cohesive formats suitable for use in statistical and computation application systems.

B. **Data Integrity/Information Sources:**

- Integration of source data within organization's core information systems (Infinite Campus, Visions); development and integration of organization's student assessment data warehouse

… All of the district's elemental student information originates within the district's student information system (Infinite Campus). The underlying data structures are highly complex, composed of hundreds of interconnect tables and associations. In addition, teacher employee information is housed within a separate system (Visions). It was necessary to utilize each of these systems to extract and assemble the elemental information feeding into the analytic environment.

II. Statistical Analysis Tools and Environments:

1. **Statistical Package for the Social Sciences (SPSS Version 21):**
    - *Statistical Analysis*: Regression, Mixed (Multi-Level) Models, (various) advanced statistical analysis, factor analysis, extensive data file/table manipulations & processing, application interface with Microsoft SQL Server
    - *Command Programming*: Advanced syntax construction, sequential syntax batch processing, file structures, and file processing

… SPSS served as the primary statistical processing environment including estimation of multi-level value-added models, exploratory factor analysis, data transformations, and descriptive data analysis. Activities required integration between SQL Server and SPSS utilizing advanced programming syntax.

2. **SPSS/AMOS (Version 20):**
    - AMOS modeling environment, Structural Equation Modelling, Confirmatory Factor Analysis, output analysis

… Initial exploration of the confirmatory factor analytic models was completed using AMOS including model fit estimates and graphical analysis of model specifications. However, this program was not able to satisfactorily account for bias resulting from data exhibiting non-multi-variate normal distributions. Later in the study, the MPLUS application environment was utilized for selected statistical modeling purposes.

3. **Mplus (Version 7.11)**
    - Command syntax development, Mplus modeling environment, Structural Equation Modelling, Confirmatory Factor Analysis, Exploratory Factor Analysis, adjustments for nested data structures, output analysis

…The MPLUS application was used to adjust for the non-multi-variate normal characteristics of the Danielson professional practice rating information. MPLUS was one of only a small number of analytic environments that utilized adjustment estimators to correct for bias introduced by data non-normality. These adjustments were applied to the final confirmatory and exploratory factor models reported in the study.

4. **HLM – Hierarchical Linear and Non-Linear Modeling (Version 7.0): Multi-Level Modeling:**
    - HLM modeling environment, input data file construction, model specifications, output analysis

… The HLM application was utilized to initially estimate multi-level value added models of student academic growth. This program is a de facto standard for estimating models

using hierarchical (nested) data common in educational settings. However, advanced programming limitations and restricted output file structures limited its efficient use in a large-scale production environment. As such, the SPSS application was used to estimate the collection of multi-level models presented in this study. Importantly, the HLM program and the extensive literature base for interpreting multi-level output served as verification of the SPSS-generated output.

III. Additional Analytic Tools & Environments

5. **HyperResearch - Qualitative Analysis Tool (Version 3.5.2) – Qualitative Information Processing:**
    - Source data construction/organization, extensive file processing/partitioning, coding organization, filtering, extraction, file extraction formatting for integration into Microsoft Word, code analysis (graphics, frequency, relationship mapping)

…The HyperResearch application served as the primary software environment of examining the qualitative information used in this study. All codes, code groups, annotations, and links to exemplars were constructed using HyperResearch. In addition, its data manipulation capabilities were utilized to substantively restructure the information for use in more comprehensive analysis and reporting.

6. **Microsoft Word: (Data Analysis)**
    - Extensive content manipulation, formation & segmentation; use of editing tools, conditional coloring, track changes, bookmarking, macro programming;

… In later stages of the research, Microsoft Word was used as a primary tool for conducting additional rounds of narrative analysis and data organization. Use of conditional data coloring, commenting/annotation, concept bookmarking, track changes became essential tools in the analysis, data organization, and writing activity. Extensive use of macro programing including custom scripts for navigating, searching, identifying, and formatting made the process of working with hundreds of pages of qualitative data and writing far more manageable and efficient.

7. **Website Design and Data Management**
    - Constructed a cloud-based data location to contain HTML and related documents for linking via on-line blog site.

…As part of the research activity, a web site was constructed to document all of the various components, background information, intermediary analysis, code books, and supporting statistical information underlying the study. The intent was to provide readers with information not inclusive in the formal dissertation publication and to make each part of the research activity more accessible.

8. **Blog Sites**
    - Established two on-line blogging sites:
        i. one for use in recording personal reflections of research activity in support of research questions

ii. the second to organize dissertation resources and information in order to web-enable the dissertation activity and provide comprehensive data to interested readers

- (Note: the resource blog site was constructed to place

... An important element in the research activity was the journaling activity meant to record personal experiences and reflections. The early activity utilized a hand-written journal. Part way through the study, a web-based blog site was established so that entries could be made electronically. This made the process of reviewing, searching, and codifying journal entries much more efficient. A secondary intent of establishing the blog site is to begin discussing issues related to teacher evaluation, accountability, and related scholarly research in an open forum (however, personal entries made during the course of the study will remain unavailable to the general public).