An Argument-Based Validation Study of the Teacher Performance Assessment in Washington State

HENNING, ANGELA,SUE

Abstract

This study examines the validity assumptions of the Teacher Performance Assessment (TPA) using data collected from teacher candidates, mentor teachers, university supervisors and university faculty in two programs at one university during the 2012 field test in Washington State. Applying the work of Michael Kane (2006) on argument-based validation, this study developed interpretations and assumptions of TPA test score use using the following five inferences: Construct Representation, Scoring and Evaluation, Generalization, Extrapolation, and Decision Making. This multi-method study utilizes survey, case study, and test score data. The overarching research question that guided the study was "Is the TPA a valid measure for determining teacher readiness?" The overall findings suggest that the operationalized construct of readiness is stable but scores are not generalizable across populations and guidance was not in place regarding score meaning and use prior to the field test. Low correlation between the TPA and university instruments provided divergent evidence for the use of TPA scores, indicating that decisions made based solely from TPA scores may not be reliable.

Keywords: Teacher Performance Assessment, Teacher Preparation, Assessment, Validity

An Argument-Based Validation Study of the Teacher Performance Assessment in Washington State

A dissertation submitted to the School of Education at Durham University

in partial fulfilment of the requirements for the degree of

Doctor of Education

A. S. Henning

May 2014

Table of Contents

List of Abbreviations

AACTE	American Association of Colleges of Teacher Education
ABV	Argument-Based Validation
AL	Academic Language
CA	California
CAEP	Council for the Accreditation of Teacher Preparation
CT	Connecticut
ELL	English Language Learner
EoM	Errors of Measurement
GRD	Graduate
IA	Interpretive Argument
IHE	Institution of Higher Education
INTASC	Interstate Teacher Assessment and Support Consortium
NB	National Board
KSJ	Knowledge, Skills, and Judgments
PESB	Professional Educator Standards Board for Washington State
PPA	Performance-Based Pedagogy Assessment for Teacher Candidates
RQ	Research Question
SCALE	Stanford Center for Assessment, Learning, and Equity
ST	Student Teaching
SV	Student Voice
TPA	Teacher Performance Assessment
UG	Undergraduate
VA	Validity Argument
VE	Validity Evidence

List of Tables

List of Figures

Statement of Copyright

Acknowledgments

I would like to thank my director, Dr. Per Kind, for his invaluable training, patience, and guidance in planning, implementing and writing this study. There are so many individuals who deserve a special thank you for their assistance. Ms. Kaitlyn Rebbe, research assistant extraordinaire, it is clear that your motto is to go above and beyond. Dr. Janine Darragh, my colleague and collaborator, I owe you much for your work as a second interviewer, a second pair of eyes, and a discussion partner. Give me a D.O.N.E! To Margo Long, Stacy Hill and Dr. Brian Davenport, my Dissertation Support Group, because sometimes highlighters are the perfect tool and a coffee break is the best strategy. To Dr. Debbie Tully, Dr. David Cherry and Dana Stevens I offer my thanks for guidance and feedback on various stages of this project. I am grateful to Dr. Dennis Sterner and Dr. Betty Williams whose leadership created an environment in which both the product and process of this work could take place. I am fortunate to have been mentored by educational leaders of your caliber. Credit goes to Stanford Center for Assessment, Learning and Equity for providing permission to use excerpts of the Teacher Performance Assessment. Thanks to Chairs Coffee House, you are an academic's version of *Cheers*.

I would also like to thank the mentor teachers, university faculty, and especially the teacher candidates and supervisors featured in this study, for their willingness to share the experiences that made this study possible. You know who you are and I hope you feel I have represented your voices well.

To Linda Graham, Melanie Van den Hoven, Michal Rosenberger and Tracy Adams, encouragement can seem a small gift to give but it's an enormous one to receive, thanks for your cheer. Most importantly, Dr. Brian Henning, thanks for your faith, inspiration, laughter, and partnership. You believe me capable of what I do not yet think possible and so I dream. May we walk and talk all our days. And to my daughters, Hope and Nora, you've never known a life when your mother was not pursuing her doctorate. The best is yet to come.

Chapter One

Overview of the Research Problem

Assessing teacher candidates' readiness to teach is an essential task for professional quality and effectiveness. In most US states, candidates for teaching licensure are evaluated during their "student teaching" semester to determine their readiness to teach and abilities as an effective practitioner. Such assessment practices are not controversial. However, determining just what criteria define "effective teaching" and "teaching-readiness" can be contentious. This has led to differences in teacher preparation between states. Definitions of teacher effectiveness have come under criticism as US students' test scores lag behind other countries, leading to political intervention from both the federal government and states to improve teacher quality. In an effort to maintain control over the profession and standards for teacher quality, preparation, and effectiveness, state and national educational institutions such as the Stanford Center for Assessment, Learning, and Equity (SCALE), Council for the Accreditation of Educator Preparation (CAEP), and the American Association of Colleges for Teacher Education (AACTE) have endorsed the adoption of a national standard instrument to determine teacher readiness called the Teacher Performance Assessment (TPA).

This recommendation, in combination with states' mandates to adopt a new instrument for determining teacher readiness, is in response to a movement for more uniformity in defining standards for teacher preparation which would assist in evaluating the effectiveness of teachers. Because researchers in educational assessment and measurement have for some time shown that performance assessment is appropriate for measuring teacher readiness, adoption of a performance-based tool to determine teacher readiness is not provocative. Any summative assessment given during the student teaching term should aim to collect evidence supporting the qualification of the candidate to adequately *perform* the expectations of the field. However, performance assessments currently practiced have differed in the ways they have measured performance. Thus a new, parallel instrument is needed. One such assessment adopted by twenty-nine states is the Teacher Performance Assessment (TPA).[1] The purpose of the TPA is to establish whether candidates are ready to teach within their subject-specific performance-domain. Ultimately, in Washington State, where this study was conducted, the TPA will determine whether a candidate will become a licensed teacher. Without such a license, or certificate, a teacher cannot be gainfully employed in the public school system. The ultimate consequence of failure on this exam is the denial of the professional livelihood of a trained teacher-candidate. It is high-stakes.

More research is needed to support the validity and reliability of the TPA as a high-stakes assessment. One challenge of the TPA is that it differs based on candidate placement and endorsement areas. It is designed to "focus on subject-specific pedagogy and use evidence drawn from an authentic experience teaching a group of students" (SCALE, 2012). To date, the TPA has 27 different "handbooks", each with its own subject specific performance requirement. In addition, differences in individual student teaching placement sites and partnerships, as well as differences in the support provided during the testing preparation and completion, increase the need for explicit research in the claims that SCALE has successfully "create[d] and deliver[ed] a reliable and valid performance assessment system for enhancing the quality of teachers in the United States" (SCALE, 2012). The problem of divergent candidate experiences, and whether those differences impact test scores, must be addressed if the TPA is to generate the most accurate diagnostic information for licensure decisions.

As a teacher-educator, I became interested in this assessment in the fall of 2011, soon after its adoption in WA. Charged with implementation of the TPA in the undergraduate program during the optional pilot and mandatory field test, concerns quickly emerged about assessment procedure

[1] The TPA was renamed in Fall 2013 to the edTPA. However, this study was conducted during the Spring 2012 term and will use the name of the assessment as it appeared at that time.

and scoring, how to best prepare and support candidates for the TPA, and, more importantly, about the reliability and validity of the TPA as an assessment used for high-stakes licensure decisions for novice teachers. A quick review of the literature before the field test (spring 2012) indicated divergent evidence and no published validation research around the intended summative purposes of the TPA test scores (see Chapter two). In response to these concerns, this study investigates the validity of the TPA. An argument-based approach to validity is applied as a framework to guide validation judgments and to justify score-based interpretations and uses (Kane, 2006). In this multi-methods research project, participants were surveyed throughout student teaching and six candidates shared their experience in case studies. In addition, mentor teachers and university supervisors, who were partnered with these candidates, prepared evaluations and observation reports to ascertain teacher readiness. Interpretations of, and uses for, the TPA were validated using multiple instruments and from diverse perspectives. The argument-based evidence provided in this study will ultimately add to the body of literature surrounding the TPA as a high-stakes tool for determining teacher readiness.

Argument-Based Approaches to Validity

Validity is the most basic and essential factor in the development and evaluation of an examination, most especially one with high-stakes outcomes such as the TPA. According to the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999), validity “refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9). Using the work of Lee Cronbach (1971, 1988), Ernest House (1977), and Samuel Messick (1989), Michael Kane (1990) argues that judgments of a test-score’s meaning are associated with a chain of interpretive arguments, which include the assumptions and inferences inherent within “the proposed interpretation and uses of test scores” (Kane, 2004, p. 136). Validation efforts “focus on empirical checks of the inferences and assumptions in the interpretive argument” (Kane, 2004, p. 144). Therefore, validity is “an argument construed by an analysis of theoretical and empirical evidence instead of a collection of separate quantitative or qualitative evidence” (Park, 2012, p. 73). Focus is on the validation of the interpretation of a test-score and not a specific test item, task, or instrument. Each time the test is administered is a separate, unique event that impacts its interpretation but its validity "is never proven; it is always subject to change" (Kane, 2004, p. 144). These inferences, taken together, form the interpretive argument which is then examined based on the plausibility of the assumptions that underpin them. The more plausible the assumptions, the more valid the test-score interpretation.

Kane's structure for an argument-based approach to validity is the framework of this study. Using Kane's (Kane, 2004) two-stage process, this study first sets out to identify teacher readiness, as it is defined in the TPA, through the test tasks and criteria, including standards of readiness proficiency for each of these tasks. This first stage also includes the development and articulation of the interpretive argument by clarifying which inferences, claims and assumptions reinforce TPA interpretations. These assumptions lead to a collection of proposed claims or inferences for the way in which the TPA would be used. This incorporates the test maker's (SCALE) specified and stated purpose of the TPA.

In the summative stage of the study, the interpretive arguments are evaluated using the sources of validity evidence. As mentioned above, credibility of the validity argument is examined in relationship to the plausibility of the assumptions that support each inference, or claim, and are used to frame the study research questions. The validity of the interpretive argument is stronger when each type of evidence supports the inferences and assumptions regarding test score judgments and uses (Kane, 2006). A description of the validity framework appears in the next section.

Research Framework

Following Kane's methodology, the research framework has two parts: the interpretive argument (IA) and the validity argument (see RQ below). The IA in this study constructs and supports arguments for score-based interpretations for determining teacher readiness in the TPA. Primarily, the key interpretation guiding the research process is the argument that *TPA scores provide a measure of relevant teacher readiness*.

To justify this interpretation, the following four claims (inferences) must be proven sound:

1. The TPA is relevant and aligned.
    - The criteria of the TPA represent, to the same depth and breadth, the state standards for teacher readiness (basis of state accreditation requirements for teacher preparation programs).
    - Criteria on TPA are aligned to researched and established expectations of teacher readiness.
2. The TPA is fair.
    - Test scores across all identifiable candidate groups have comparable interpretations with respect to certification area.
    - All identifiable and relevant candidate groups receive equitable treatment from (or with respect to) the assessment.
3. The TPA is based on adequate levels of proficiency.
    - The TPA retains an adequate level of rigor in the proficiency levels of each of the 15 sub-traits.
    - Judgments of candidate proficiency are set using researched and established methodology.
4. The TPA is consistent.
    - Candidate scores do not depend upon assignment to a particular scorer, handbook, university, placement site, triad partnership (candidate/mentor/supervisor), or student teaching length.

- Candidate TPA scores are equally as reliable (or better) an indicator of achievement of teacher readiness as mentor and supervisor observations of candidate achievement and readiness to teach.
- Scores of teacher readiness on the TPA correlate to scores provided by mentors and supervisors partnered with the candidate during student teaching.

Finally, in order to determine the plausibility of the four inferences, the following assumptions regarding performance evidence guided the creation of the IA:

1. The tasks in the TPA represent the particular performance and skills that can be used to base a decision on teacher readiness.
2. The performance criteria on the TPA that measures teacher readiness should conform to the state standards for teacher preparation in any state for which that tool was adopted.
3. Performance traits on the TPA are related to performance of the same traits measured in other assessments of teacher readiness.
4. The criteria, rubrics, procedures, and scores derived from the TPA are generalizable across different candidates and handbooks.
5. TPA proficiency does not depend on factors beyond the candidate’s control. The criteria, rubrics, procedures, and scores derived from the TPA are generalizable across testing sites, placements and placement length.

The first and second assumptions postulate that the TPA evaluates the knowledge, skills and abilities of candidates for the purpose of diagnosing their readiness to enter the professional field. Proficiency and readiness to teach according to the TPA is defined by a candidate who can provide evidence of ability in:

(a) planning for instruction, assessment, academic language, and the specific students they teach; *and*

(b) instruction in such a way as to engage and deepen student learning; *and*

(c) analysis of assessment data to improve teaching and student learning; *and*

(d) reflective practice to identify areas of teaching strengths and weaknesses.

In WA, requirements for the preparation and licensure of education professionals are articulated and evaluated by an advisory body called the Professional Educator Standards Board (PESB). PESB's role is to ensure that all educators prepared in Washington, "are competent in the professional knowledge and practice for which they are certified. Have a foundation of skills, knowledge, and attitudes necessary to help students with diverse needs, abilities, cultural experiences, and learning styles meet or exceed the state learning goals. Are committed to research-based practice and career-long professional development" (PESB, 2013). In order to test these first two assumptions, the criteria from the PESB for teacher preparation will be compared with the definitional criteria articulated by SCALE in the TPA tasks and rubrics. If this document analysis reflects alignment between these two groups of descriptors of teacher readiness, a theory-based inference would be supported.

It is important to distinguish two terms sometimes conflated in the literature of teacher preparation: readiness and effectiveness. Effectiveness is the domain and implicitly involves a scale of measurement with novice at one end and expert at the other. Just as National Boards outline the expectations of effective *expert* teaching, the validity argument in this study will determine whether the construct of the TPA similarly outlines developmentally appropriate expectations for effective *novice* teaching. For this reason, an assumption of this study is that the theoretical construct of any assessment of teaching intended during student teaching, limits and interprets the domain to only knowledge, skills, and judgments appropriate for novice teachers. One of the requirements for developing expert professional skill is time spent engaged in professional activities; expertise requires practice. Recent articles in the study of professional expertise suggest that skilled performance is acquired through targeted training and extended opportunities for deliberate practice (Billett, 2012; Ericsson & Ward, 2007; Ericsson, 2004). Ericsson and Ward (2007) write, "In domains where expert performance is measurable, acquisition is gradual and the highest levels are only attained after 10 years of intense preparation—even for the most ''talented'' (p. 346). The old adage "practice makes perfect" has been challenged by some cognitive psychologists and should be replaced with "10,000 hours of practice makes progress" (Gladwell, 2008; Csıkszentmihalyi, 1990). Throughout this study, the terms "effective" and "readiness" are distinguished because it is important to identify the differing domain expectations for novice teaching verses that of expert teaching. The assumption that expertise requires theory in practice is embedded in this ABV study.

The third assumption examines criterion-related validity and the extent to which the TPA test scores are related to the scores of teacher readiness provided by other measures of assessment during student teaching. Candidate TPA scores should be as equally as reliable an indicator of achievement of readiness as mentor and supervisor observations of candidate achievement and readiness to teach. Rather than simply seeking out convergent evidence among these different sources, this study views divergent evidence as possible additional factors to better understand the target construct being assessed. Correlation between scores on the TPA with a scale completed by both mentor and supervisor at the mid-point and summative point during the student teaching term was calculated using Multi-Trait, Multi-Method (MTMM) analysis. If these two sets of scores are highly-correlated, an assumption that the TPA is a reliable measure of teacher readiness can be determined. A low-correlation, on the other hand, does not necessarily mean that the TPA is unreliable. Low-correlation may be a result of the diverse purposes for which the two measures were developed, used, and scored. Whereas because the TPA is rated by scorers who have had no contact with candidates, students, placement sites or the university that prepared the candidate, the supervisor and mentor scores are influenced by personal connections and relationships with these groups.

The fourth and fifth assumptions address the potential impact of various random errors associated with the conditions under which participants completed their TPA. Because the TPA emphasizes subject-specific pedagogy and evidence obtained from “authentic experience teaching a group of students” (SCALE, 2012), differences in placement sites, levels, partnerships, and lengths (test method effects) are critical factors that may jeopardize valid interpretations of scores across differing licensure subjects (TPA handbooks). It may also increase the chances for construct-relevant or irrelevant errors in test scores. Candidate scores should not depend upon assignment to a particular scorer, handbook, university, placement site, the mentoring relationship, or the student teaching length. If the experiences and scores derived from the TPA are consistent across handbook subjects, placement sites and levels, successful and unsuccessful partnerships and differences in student teaching term lengths, the generalizability assumption will be supported. The reliability issues associated with these differing factors will be examined. If bias exists, it will undermine the validity score interpretations.

Research Questions

The second stage in ABV is the validity argument. The purpose of the validity argument is to validate the interpretations of TPA scores as a measurement of teacher readiness using multiple-method data sources and diverse perspectives. An argument-based approach to validity provides the guiding framework in the form of research questions and justifies the score-based interpretations. The five assumptions addressing the differing claims of the interpretive argument are subsequently used to formulate the nine central research questions of this study:

Inference 1: CONSTRUCT REPRESENTATION

1. Do the four TPA tasks represent the six categories relevant to the intended construct (teacher readiness)?

Inference 2: SCORING/EVALUATION

2. Are the scoring procedures sound and reliable?

3. Are the rubric score levels achieved by the candidate actually representative of what that candidate performed on the TPA? (Does candidate performance correlate to the assigned test score?)

Inference 3: GENERALIZATION

4. Are the score levels achieved on the rubrics a true representation of a candidate's performance? In other words, are the scores a candidate earned consistent and generalizable with other samples of that candidate's teaching performance?
5. Does poor performance on the TPA imply a lack of adequate mastery of the construct?
6. How generalizable are the criteria, rubrics, procedures, and scores derived from the TPA across different candidates and handbooks?
7. Does TPA proficiency depend upon factors beyond the candidate's control? How generalizable are the criteria, rubrics, procedures, and scores derived from the TPA across testing sites, placements and placement length and programs?

Inference 4: EXTRAPOLATION

8. Do TPA test scores provide reliable indicators of a readiness to teach? Are the TPA scores, as a whole, a true measurement of teaching ability?

Inference 5: DECISION MAKING

9. Is guidance in place so that all stakeholders know what scores mean and how the outcomes will be used?

Significance of the Study

This study makes significant contributions to the research on the Teacher Performance Assessment and will have direct implications for teacher preparation practices. Four research areas are of particular relevance: (a) identification of the readiness construct, (b) importance of thoughtful implementation of performance assessment for beginning teachers, (c) evidence of the TPA as an opportunity for teacher growth and evaluation, (d) implications for practice and programmatic revision in teacher preparation, and (e) use of argument-based validation in performance assessment with high-stakes licensure consequences.

First, data provided from this study enables researchers and test developers to better understand the relevance and inferences of the TPA. Despite much research in the area of performance assessments for teacher licensure, only one study to date has attempted to validate use of TPA scores (SCALE, 2013). Contributing to the literature, this study uses MTMM to correlate scores for identified standards measured by the TPA with other instruments of teacher readiness. The findings derived from these analyses enrich theories of teacher readiness by examining the plausibility of test score usage for an accepted measure of that construct. Any weaknesses found in the validation argument provides direction for TPA review.

Second, this study confirms the importance of performance assessment for review of teacher readiness by examining as one model of performance assessment (TPA). Although not the case in Washington, some states' licensure assessments are limited to multiple-choice tests that measure only content or pedagogical knowledge through written scenarios. This limited approach prevents a thorough investigation of teacher readiness across the nation and fuels concerns about teachers' practical effectiveness. Because the TPA evaluates key performance expectations of practicing teachers in the field, few researchers can counter the importance of the core tasks performed by candidates in the TPA. However, while the political climate of data-driven decision-making around teacher quality measurements urgently pushes for conformity in the profession, this study responds to and contributes to the call for research and more careful review before the adoption and implementation of high-stakes performance assessments such as the TPA.

Third, this study fills a gap that exists between earlier studies of teacher performance assessments for teacher licensure and the TPA. Despite sharing similar structures, the TPA is a different assessment from its predecessor, the Performance Assessment for California Teaches

(PACT) (Chung, 2005) and, as such, deserves its own validation study.[2] Despite some significant differences between them, most research cited to support the validation of the TPA is actually, (1) based on the PACT and, (2) has either emphasized qualitative methods to ask questions about the experience of the beginning teacher (S. Newton, 2010; Chung, 2005; Wei, 2010; Pecheone & Chung, 2006), or has concentrated on defining the construct of teacher readiness (Darling-Hammond, 2012; Darling-Hammond, 2010; Whittaker & Young, 2002). This study expands the scope of previous research by critically evaluating claims made about TPA reliability and validity for licensure decisions. Using qualitative case studies and quantitative surveys, this study examines the extent to which the TPA is experienced as a learning tool, in addition to a summative teaching event, in order to validate its use as a performance-based assessment in the developmental experience of student teaching.

As an ever-growing number of states adopt and implement the TPA for decisions of teacher readiness and licensure, teacher preparation programs across the country will be revising content taught and procedures for the preparation of their candidates, not just for success on the TPA, but based on the standards of teacher readiness defined by the TPA. Using qualitative focus group interviews and quantitative surveys, implications for reform in teacher preparation and for programmatic changes needed for candidate success are discussed.

Finally, because few ABV studies have been conducted, and none address high-stakes performance assessment in teacher readiness, this study contributes to the literature on ABV practice.

Study Overview

There are five chapters in this research study. Chapter one details an overview of the research problem, the framework for the argument-based approach to validity used in the study including an outline of the five assumptions that guided the development of the research questions and process. Chapter two reviews the literature on performance measures that define and assess teacher readiness

[2] In 2012-2013 the TPA handbook and scoring criteria were still undergoing significant revision.

and the argument-based validation framework. Chapter three describes the study methodology and outlines participants, instruments and data collection, and analysis procedures applied in the study. Chapter four presents the data analysis outcomes and reports the findings of the study relative to the research questions. Finally, chapter five synthesizes the research and discusses areas of future research, specifically the possible implications for the preparation of teachers whose license will be granted or denied based on their TPA scores and the use of ABV as a validation methodology.

Chapter Two

Review of Literature

Teaching is a multi-faceted and complex professional skill. Therefore, a variety of components comprise the construct of teacher readiness including content pedagogical knowledge, professional skill and judgments, and the ability to apply both of these in a relational real-world setting. Like other professions with licensure and certification requirements, professional teaching distinguishes between novice and experienced practitioners (NBPTS, 2013; Kane, 2004; Kane, 1994).

This chapter will describe the standards, licensure and certification process for practicing teachers in Washington (WA) where the performance assessment adopted for licensure decisions is the Teacher Performance Assessment (TPA). Research on, and validation studies of, performance assessments similar to the TPA will be described. Then, the TPA will be defined and situated within the standards-based, performance assessment movement. Finally, justification is provided for the use of an argument-based validation approach.

Standards for Teaching

Initial Licensure Standards and INTASC

In recent decades, national standards for teacher preparation have been developed and adopted by many states in the US using a framework from The Interstate Teacher Assessment and Support Consortium (INTASC). INTASC is a consortium of state education agencies and national educational organizations. Since 1987, INTASC has been dedicated to the “reform of the preparation, licensing, and on-going professional development of teachers” and “guided by one basic premise: An effective teacher must be able to integrate content knowledge with the specific strengths and needs of students to assure that all students learn and perform at high levels” (CCSSO, 2013). In 2011, INTASC released its latest standards for beginning teacher licensing, assessment and development. These 10 standards focus on a “Common Core of Teaching Knowledge.” Within each standard, specific knowledge, disposition, and performance criteria are articulated in three quality levels or "progression indicators" (CCSSO, 2013) (see **Appendix A**).

Certification and NBPTS

Experienced, licensed teachers can seek certification (Shimberg, 1981). Certification is granted at the national level by the National Board for Professional Teaching Standards (NB) who have developed a framework separate from that of INTASC to describe teaching effectiveness and quality. In 1989 the NB published the document *What Teachers Should Know and Be Able to Do* describing "Five Core Propositions for Teaching" which has been compared to "medicine's Hippocratic Oath - setting forth the profession's vision for accomplished teaching. The Five Core Propositions form the foundation and frame the rich amalgam of knowledge, skills, dispositions and beliefs that characterize NB Certified Teachers (NBCTs)" (NBPTS, 2013).

NB "has a clear vision of what highly accomplished teaching means, and it thinks that the identification and recognition of high accomplished teachers can contribute to improvements in the quality of education" (Sackett, 1998, p. 126). Many studies have examined the NB certification process, and positively reviewed its effects on the candidate and on teaching (Sato, Chung, & Darling-Hammond, 2008; Darling-Hammond, 2006; Humphrey, Koppich, & Hough, 2005; Darling-Hammond & Loewenberg Ball, 1998). Boards are expensive and time consuming. If the candidate qualifies, they complete an 18-24 month evidence-based, student-centered, reflective process culminating in the submission of an extensive teaching performance assessment in the form of an eportfolio. Certification is granted in one teaching subject area per submission and must be renewed every 10 years. While licensure in the state of practice is required, board certification is highly valued but not mandatory.

Initial Licensure in Washington and the PESB

WA initial residency teacher license requires that candidates have successfully completed a bachelor's degree from a state accredited institution of higher education (IHE),[3] pass an item response exam of basic skills called the West B, and a computerized item response exam of subject-area knowledge called the West E. Until 2012, candidates were also required to pass a performance assessment called the Performance-based Pedagogy Assessment of Teacher Candidates (PPA) during a student teaching internship. In 2010, WA Senate Bill 6696 required the PPA be replaced with another performance assessment. This led to the adoption of the TPA.

Teacher preparation programs are accredited by the Professional Educator Standards Board (PESB). Unlike standards boards in other states, PESB did not adopt the INTASC teacher standards, opting instead to develop their own. Most of the PESB standards align with those articulated by INTASC. PESB conducts site visits to determine how well each IHE in WA meet these criteria. Only accredited IHE are authorized to recommend candidates for an initial residency license. Unlike licensure, teaching certification is achieved at the national level and WA recognizes the NB for the advanced certification of teachers.

Performance Evaluations of Teaching Effectiveness and Readiness

Professional organizations such as INTASC, NB, and PESB "developed standards based not only on what teachers needed to know, but also on what they needed to be able to do" (Sandholtz & Shea, 2011, p. 40). These standards became the basis for designing assessments that could measure whether a candidate was ready to teach. Standards for teaching performance are not contentious. However, the best way to *evaluate* candidate knowledge, disposition, and performance differs and informed experts disagree (Youngs, Odden, & Porter, 2003).

[3] These requirements changed in 2014. This reflects standards during the field test. Teachers with a valid teaching license from another state, and at least three years of teaching experience, can also apply.

The rise of performance assessment corresponds to an evaluation movement advocating criterion-based, or standards, assessment rather than norm-based assessment (Kane, Crooks, & Cohen, 1999; Taylor, 1994). Many educators and researchers questioned decisions based on relative or comparative measurement. The goal is to make sure all candidates meet standard, regardless of whether their peers also meet standard (Taylor, 1994). Grant Wiggins (1989) critiqued the use of norm-referenced tests to assess whether a student met a standard. Instead, he proposed designing and using performance assessments as more authentic assessments of learning. His call convinced many educators to adopt performance assessment to evaluate student understanding and rapidly spread from classroom practice to teacher preparation.

Building on curricular changes in teacher education programs that emphasized knowledge, skills, and dispositions in terms of situated learning and constructivism, professional preparation promoted a view of teachers who “must adapt their teaching to meet the diverse and changing needs of students in their classrooms” (Sandholtz & Shea, 2011, p. 40). This “variability of context, combined with complexity of teaching, has shifted the view of the teacher to ‘a thinking, decision-making, reflective, and autonomous professional’” (Sandholtz & Shea, 2011, p. 40). Advocates of teaching, as a highly-skilled profession, argue that both learning and readiness to teach is *contextual.* Therefore, objective pencil-and-paper measurements do not evaluate teachers’ practice in context are no longer sufficient for licensure decisions which led to the adoption of performance assessments, or assessments that include tasks that “elicit complex demonstrations of learning and measure the full range of knowledge and skills necessary” to address the construct (PARCC, 2010, p. 35).

Teacher learning in training is now widely perceived to involve real-world problem solving and the application of teaching practices that adapt to meet the changing needs of students in authentic classrooms. In the 1980s, Georgia, Florida and Texas were early adopters of teacher-performance assessments. However, these assessments were too standardized and focused on a uniform set of behaviors and strategies that all teachers should perform, regardless of the situational context of their teaching placement (Youngs, Odden, and Porter, 2003). Therefore, a new generation of performance assessments were designed that applied observation of teaching practice embedded within the classroom context where candidates can demonstrate that they were responsive to the learning needs of every student and reflect on their professional knowledge and skills, specifically in the context of observed taught lessons. Schools of education in WA have an established history of evaluating teachers and recommending licensure based upon observed evidence of knowledge *in practice,* this expectation was imbedded in the PPA adopted in 2004.

Because they are more authentic, performance assessments are considered more valid for licensure decisions than tests that focus only on knowledge, or on simulated practice, with multiple-choice or essay questions (Darling-Hammond & Snyder, 2000). Performance assessments of teacher readiness have been promoted as stronger in content validity (Popham, 2005; Popham, 1990), more effective in shaping teaching habits of mind (Fenderson, 2010) and a more authentic approach to assessing candidate knowledge, skills, and judgments than standardized, multiple-choice tests which may over-simplify professional teaching activities or offer multiple "right" responses for various teaching contexts (Cochran-Smith, 2003; Darling-Hammond, 2001).[4] Performance, as a teacher readiness assessment method, has many benefits, depending on established validity, reliability, and

[4] Conversely and controversially, Sackett, Borneman, and Connelly (2008) found that the predictive measures of success from standardized testing used for employment and decisions in higher education were strongly "supported by the preponderance of the evidence" (p. 225). Upending some of the major assumptions about high-stakes, standardized tests, they write, "We offer a very positive appraisal for the evidence (a) that tests of developed abilities are generally valid for their intended uses in predicting a wide variety of aspects of short-term and long-term academic and job performance, (b) that validity is not an artifact of SES, (c) that coaching is not a major determinant of test performance, (d) that tests do not generally exhibit bias by underpredicting the performance of minority group members, and (e) that test taking motivation mechanisms are not major determinants of test performance in these high-stakes settings" (p. 225).

fairness. The aim of performance assessment is to “replicate what candidates encounter in a real work situation and determine competence by judging their performance in the actual tasks and activities” (Sandholtz & Shea, 2011, p. 40). One benefit is that these assessments are linked to professional teaching standards that represent a high degree of consensus around the domain of effective teaching (Arends, 2006; Wiggins & McTighe, 2000). As mentioned above, unlike traditional assessments, another benefit of performance assessment is that it measures evidence of teacher practice. This is a more direct method to evaluate teacher readiness (Kane, Crooks, & Cohen, 1999). “The focus shifts from determining a candidate’s possession of knowledge and skills to determining the way in which a candidate uses his or her knowledge, skills, and dispositions in teaching and learning contexts” (Sandholtz & Shea, 2011, p. 40). As predictors of success in work settings, direct methods of assessment have been thought to provide stronger evidence than indirect tests (Uhlenbeck, Verloop, & Beijaard, 2002). Decisions made about the quality of candidates are less subjective because the data is based on more credible evidence (Arends, 2006). Finally, in addition to the data they provide for decision-making, performance assessment can provide opportunities for formative professional development and offer education programs valuable feedback on their strengths and weaknesses for programmatic improvement (Sandholtz & Shea, 2011, p. 40; Schultz, 2002; King, 1991).

Performance Assessments: A Few Concerns

The benefits of performance assessment are tempered by a number of concerns. Sandholtz and Shea (2011) note that “when performance assessments are used for multiple functions such as credentialing decisions, accreditation, program improvement, and candidate learning, additional measurement challenges arise” (p. 41). Primarily, these concerns arise when performance assessments are used for high-stakes, summative decisions. In these cases, “the overarching challenge is ensuring validity, reliability, and fairness of the measures” (p. 41). When high consequences accompany a performance assessment, there are additional burdens for inter-rater reliability and alignment between the assessment procedures and the teaching contexts (Arends, 2006; Kane, 1994). Key issues for validation studies involve "balancing competing demands and ensuring that one function does not dominate and lessen the value of the measure for the other functions" (Sandholtz & Shea, 2011, p. 41). Candidates also express concerns that their personal lives and health suffer during the period where they prepare for and complete these performance assessments (Okhremtcouk, et al., 2009; Sackett, Borneman, & Connelly, 2009).

In addition to these concerns, performance assessment requires significant levels of human and financial resources both on university programs and the individual candidates. Zeichner (2003) notes that limited resources may cause teacher preparation programs to superficially implement the assessment procedure. When accountability stakes are high for programs and candidates, a concern is that the result turns "performance based teacher education into a purely mechanical implementation activity that has lost sight of any moral purpose" (Zeichner, 2003, p. 502). This leads to what some have called the hidden curriculum of performance assessment. The emphasis in teacher preparation is "shifting to alignment and compliance, thus limiting the way teaching is represented in the curriculum, inhibiting consideration of other perspectives, and avoiding issues related to values and philosophical choices" (Sandholtz & Shea, 2011, p. 41). The concern is that only those standards aligned to the performance assessment will be taught, reducing the curriculum by removing all but what is tested. Young, Odden, and Porter (2003), in a review of state policies for teacher licensure, found that, in 2002, only nine of fifty states were using performance assessment for licensure decisions. They suggest that this was because of the high costs of implementation, and that the time consuming nature of the assessment would have a negative effect on the teacher supply. They also found there were questions about their validity, reliability, and fairness (Young, Odden, and Porter, 2003).

Some have suggested that this practice will decrease academic freedom, which could impact research, scholarship, and inquiry or, paradoxically, loss of important traditions, practices or methodologies. Peck, Gallucci and Sloan (2010) use the term "negotiation" to describe faculty engagement with the "new and demanding state policy mandates" for performance assessment (p.

9). They write that the WA mandate was “perceived by faculty and staff to intrude strongly on the integrity of local program values and practices” (p. 9). Though the authors could not answer whether the policy change was “a good thing or a bad thing” they did feel that, by changing the “focus” from compliance to “inquiry,” they “served to deflect what they perceived to be potential negative outcomes” (p. 11). Nevertheless, Peck, Gallucci and Sloan describe a difficult transition for faculty. While they are convincing that teacher preparation was improved based on PACT data, it is not clear that the negotiation they describe was fully open to multiple perspectives. The program may have also too easily dismissed legitimate concerns from colleagues whose educational paradigm, research interests, scholarship, and course materials had to be significantly adapted based on the changes negotiated. Performance assessments, especially those that are standardized for a large number of participants and contexts, may exclude aspects of teaching that are important, but not easily measured (Arends, 2006) and this will have a trickle-down effect in reducing the curriculum to only what is required for the assessment. The hidden curriculum of performance assessment can potentially eradicate important differences that, while maybe not always welcomed, do contribute to a scholarly discussion of teaching as a practice and a field, and that are part of the purpose of a university in our society (Newman, 1907).

Performance Assessments for Accountability

Despite, or perhaps because, these concerns did not outweigh the benefits for advocates, evaluating teacher ability through performance assessments is not viewed as controversial. Likewise, it is hardly controversial to find that teacher education reform would be the result of data from high-stakes testing. By the turn of the century, evidence in support of performance assessment convinced many state legislatures to adopt performance assessment measures as a part of the initial teacher licensure process (Youngs, Odden, & Porter, 2003; Taylor, 1994). This was true of the WA Legislature and PESB, when the PPA was adopted.

Mandates for performance assessment are connected to policymakers and public perceptions of poor teacher quality. Public trust in education has remained low since 1983, when the highly critical and influential *A Nation at Risk* report was first published. Negative views of education and teacher effectiveness have only increased, despite national reforms implemented by the federal government for K-12 students, the publication of *The Nation's Report Card* by the National Assessment of Education Progress, and open-access to data released from *No Child Left Behind* testing, showing improvement in many schools across the country. Critical views have extended to the IHE that train teachers. For instance, the federal *Higher Education Act* (HEA) now asks that IHE evaluations be partially based on graduates' performance and test scores. The American Association of Colleges for Teacher Education (AACTE) (2011) made the following recommendations for teacher preparation:

1. The federal Teacher Quality Partnership grant program should be "revamped" to focus on preparing and supporting teachers, principals, school administrators, and other key educational personnel (pp. 6-7).
2. Teacher quality levels should be established to distinguish between "qualified and effective" (p. 7). The TPA would be tied to the process by which teachers would become "qualified" (pp. 8-9).
3. Teacher evaluation should be based on multiple data measures including "impact on student learning, classroom observations, peer reviews, and school-wide progress" (p. 8).
4. The federal government should "streamline" current accountability provisions in the HEA to enforce the policies and close "low-performing and at-risk programs" (p. 11).
5. The federal government "must invest in statewide data systems" to track the effectiveness of teachers across the continuum from candidate to expert (p. 13).

This "Data Quality Campaign" is intended to both "ensure accuracy and fairness" and also "improve the quality and effectiveness" of teacher preparation programs (DQC, 2013).[5] As seen above, AACTE is endorsing federal plans to link funding and access to resources to a national teacher performance

[5] In 2013, the Obama administration announced that they would like to extend those criteria to include graduates' earning potential and salaries (Stewart, 2013).

measure, the TPA, among other measures. Teacher preparation institutions, accreditation bodies, and schools find themselves in a perennial position of defending their practices, even as those practices continually evolve to improve test scores. Accountability has become central to any discussion of the assessment of teacher readiness.

Diane Mayer (2005) discusses several ways in which the TPA is seen as a measure of professional accountability, using an established framework developed by Darling-Hammond (1989) (p. 177). She argues that performance assessments can "strengthen the framework" of an "accountability system" in education but that "a model of accountability that values responsiveness to individual students is at odds with a model that is based on the premise that all students can and should achieve, which implies that if they do not, it is either the fault of the students or the teacher" (p. 178). Mayer's overall message is that a performance assessment for making teacher licensure decisions is, ultimately, a part of a "policy bargain" being made between those experts in the field of education and the public to "re-open discussions" and "instill public confidence" in teachers' ability, quality, and professional judgment (p. 180). Stephen Sawchuk (2012) connects this policy bargain to TPA adoption, "there has been a wave of policy interest in teacher education" and it is "pretty clear that a lot of teacher educators … have been increasingly pressed to come up with alternatives that can reliably measure teaching competence, and the Teacher Performance Assessment appears to be the main tool that many are pinning their hopes on" (Sawchuk, 2012).

Performance Assessments of Teacher Readiness

The TPA is one piece of the current reform movement asking candidates to prove their qualified status on a subject-specific, high-stakes performance exam as a part of initial teaching licensure. Since the 1980s, policies and programs in California (CA) and Connecticut (CT) have served as an unofficial national "pilot test" for performance assessment of teacher readiness.[6] In these

[6] Other states were also moving toward performance portfolios for measuring teacher readiness. These portfolios gathered evidence from the multiple field experiences and focused on "Process (reflective),

states, teaching performance assessments were developed with the purpose of ensuring candidates' ability to connect practice to student learning (CTC Website, 2009).

Between 1990 and 2003, many studies were conducted using CA and CT licensure programs and performance assessments as the context (McCormick, 2001; Darling-Hammond and Snyder, 2000; Darling-Hammond and Macdonald, 1999; Mitchell, et al. 1998; Stone, 1998; Lomask, et al., 1997; Rearick, 1997; Pecheone & Stansbury, 1996; Lyons, 1996; Shulman, 1992). At the same time, researchers were collecting data to understand candidate score differences between subjective and objective assessments. Vollmer and Creek (1993) investigated this relationship and found that candidates who had the ability to achieve on standardized tests may not have shown the same high scores on more practical, performance tests. One study focused on the CA performance assessment, the PACT, as *formative* for candidates, for university faculty, and for programmatic development, reported highly positive results for practice and reflection (Chung, 2005). This was not universally the case. Other studies found the PACT to be prescriptive, time consuming, and a "hoop to jump through" rather than an opportunity for professional growth (Moir, 2002). As mentioned before, studies found that the cost burden for "complex performance assessments often impose unfeasible resource demands" (Sackett, 1998, p. 128). The literature is nearly unanimous that the resources required for performance examinations are extensive and can prove limiting.

Validity Studies for Performance Assessment

Surprisingly, despite the call by many to make sure that the common exam for novice teachers be both "valid and reliable," the only validation study published on the TPA is the Summary Report produced by SCALE in November 2013 (SCALE, 2013). While often declared "valid and

Produce (lessons and teaching materials), or Performance (teaching experiences) evidence" (Goodman, Arbona, & Dominquez de Rameriz, 2008, p. 29). Portfolios from other states were seen as significantly different enough from the TPA, or had no validity or reliability data available, and therefore did not contribute to this study.

reliable" by proponents,[7] the studies linked to such statements (when there are citations) are for the California PACT assessment, California CalTPA (Riggs, Verdi, & Arlin, 2009), or the Connecticut BEST exam. While these assessments are similar to the TPA, especially the PACT, they are not the same.

In addition, because not all California schools adopted the PACT, those studies conducted on its reliability and validity have been small in terms of supporting a national assessment for accountability. Other than SCALE's Summary Report (2013), as of January 2014, no other validation study could be cited for the TPA or its 2013 iteration, the edTPA. The lack of studies examining the TPA validity in relation to teaching outcomes, or reliability data for the instrument and scoring process, demonstrates a need for further study. There are numerous mentions of pilot data collected for validation purposes so it may be that other studies are currently underway. Both the *edTPA Summary Report* and the studies available for those performance exams that are similar to, but not identical to, the TPA are discussed below.

SCALE summary report. SCALE published the only known TPA validation study in November 2013.[8] It is important to note that this study uses data from the edTPA, which differs from

[7] E.g., "A recent development that will significantly strengthen accountability for teacher preparation programs and reflect candidate' readiness for the classroom is the creation of a nationally available, valid, and reliable teacher performance assessment (TPA). … A few additional performance assessment models exist, but they are not as widely used as TPA, nor do most of them have the reliability and validity that TPA and PACT do" (AACTE, 2011, p. 9).

[8] Though the report suggests that the studies were conducted over the course of multiple years, all scores and data was derived from one administration of the edTPA. Note: the TPA and edTPA are two different versions of the assessment. In Fall 2013, SCALE renamed the assessment, the edTPA, when it published a revised iteration. It is not clear whether or how the data collected from the 2012 field tests of the TPA was used. For instance, was it compared to the 2013 edTPA?

the TPA. For instance it includes only three tasks (planning, instruction, and assessment).[9] This truncated report provides a historical summary of the development process, the assessment procedure, and validity and reliability data. By analyzing roughly 4,000 submissions (33%) from the spring 2013 term,[10] SCALE suggests that the edTPA has strong construct validity, inter-rater agreement,[11] and high correlations with job analysis studies (pp. 17-24). While acknowledging some implementation issues, no threats to validity were shared. SCALE is expected to offer a full technical report in 2014.

PACT. As noted, Connecticut and California were early adopters of a mandated performance assessment. There are several studies of the initial licensure exam, the PACT, developed primarily as a direct evaluation of a candidate's teaching for credentialing decisions. One of the additional purposes of PACT is to serve as a formative, professional learning experience for candidates. Lastly, PACT was designed to provide evidence for programs to understand their strengths and weaknesses to use for program improvement. In 2007, a technical report was published by the consortium summarizing the validity and reliability studies from the pilot

[9] "The Stanford Center for Assessment, Learning, and Equity (SCALE) is the lead developer of the edTPA, and Stanford University is the sole owner of the edTPA" (p. preface). The researcher requested permission to examine the edTPA version of the assessment but did not receive a response. Therefore, a full comparison of these two instruments could not be conducted here. The task and rubric numbers in the summary report do not align to the TPA, making it difficult connect the data collected in this study and the summary report.

[10] According to SCALE, standard setting occurred in August 2013. Participants in this study were selected from the group that preceded standard setting (SCALE, 2013, pp. 1-3).

[11] According to SCALE, 10% of all submissions are randomly selected to be double scored by Pearson (p. 23), though it is not clear whether this is 10% of the 4,000 submissions selected for the study or 10% of the full participation group of 12,000 candidates. Therefore, it is not possible to confirm inter-rater agreement based on the data provided in this report.

(Pecheone & Chung-Wei, 1997). Based on that report, and the endorsement of the growing consortium of universities, the PACT was approved for use by the California Commission on Teacher Credentialing.

Okhremtcouk et al. (2009) used a mixed method survey to analyze the effects of the PACT on student teaching, university coursework, instructional practice, classroom management, personal time, and candidate perceptions of the level of support required for success. Their study discusses three findings. First, the PACT was overly time consuming, cutting into the time available to the candidate to focus on instruction, development, and university coursework. Second, the PACT *did* contribute to candidates' perception of their professional growth as a teacher. Finally, IHE and placement site support networks and mechanisms are essential to candidate success. The authors write, "one of the most useful findings of this study is *the local factor*: how school placements impact student teachers ability to complete PACT" (Okhremtcouk, et al., 2009, p. 59). The outcome of their investigation suggests that those IHE with a PDS model in place are better equipped to provide support for candidates completing the PACT.

Beyond Okhremtcouk, other studies examine the formative value of the PACT. Ruth Chung-Wei provides a context for the value of the PACT as a developmental learning tool for candidates (Chung, 2008; 2007; 2005; Darling-Hammond, Chung Wei, & Johnson, 2009; Pecheone & Chung, 2006). Darling-Hammond, Chung-Wei, and Johnson (2009) have used these studies to successfully argue that "current research suggests that there are many teacher characteristics and abilities which, in combination, predict teaching effectiveness" (p. 631).

As mentioned above, a preliminary study of one year of pilot data (2003-2004) on the PACT was conducted by Pecheone and Chung-Wei (2007). They report preliminary findings from a two year pilot with thirteen IHE programs.[12] Pecheone and Chung-Wei found discernible patterns in

[12] Pecheone and Chung (Wei) published the initial pilot data in their 2006 article and then a full study for California that was pivotal in the CCTC endorsement of the PACT (2007). Because the data shared in both of

student performance demonstrating high levels of achievement in instructional planning. For the portion of the PACT that was double scored during the pilot, the study found high levels of inter-rater reliability. The pilot data also confirmed other studies of the PACT as a significant actor in the formative development of teachers. Pecheone and Chung-Wei found that candidates in urban settings believed their settings to mandate teaching decisions in such a way as to be too limiting to succeed on the PACT. The data confirms that those in urban settings who reported these limitations were associated with lower test scores (p. 29). Pecheone and Chung-Wei's study of the pilot data from thirteen programs (in one IHE) is often cited as evidence of the validity of the PACT. Their work is the most important validity evidence for the PACT in the literature, to date. However, some have questioned whether that study is enough to support national adoption of a similar instrument without its own record of reliability. Ann Berlak (2010) writes that "the key question is whether PACT scores accurately and objectively measure quality teaching. That PACT assessments are neither reliable nor valid is certain to become widely apparent in the next decade" (Berlak, 2010).

In fact, studies of the PACT as a summative assessment are less common. Stephen Newton (2010) conducted a predictive validity study of the of PACT for Stanford (the developer of PACT), presenting the relationship between beginning teacher's scores on PACT and their later teaching effectiveness as measured by value-added achievement gains in their students' English Language Arts test scores. This study examined student test scores from first and second year teachers who taught students in the upper elementary grade levels. The study found that the PACT was highly correlated (.58 to .66) with at least one of four value added measures (S. Newton, p. 12). Newton writes, "for each additional point a teacher scored on PACT, her students averaged a gain of one percentile point per year on the California Standards Tests as compared with similar students" (p. 12). The summary finding confirmed "the validity of the PACT as a measure of teacher quality, and as

the report and the article are from the same study and explicitly connected by the authors, I discuss these two sources as one study.

a useful tool for evaluation of candidates and as a way to provide feedback to teacher education institutions" (p. 13). It is notable, however, that this investigation was small in participant numbers (only 14 teachers and 259 students participated). Ducker, Castellano, Tellez and Wilson (2013) examined the internal structure of the Elementary Literature Teaching Event in the PACT finding high reliability coefficients and domain-based structures but poor evidence for the task-based structure. Another study compared university supervisor predictions and candidate scores and found that predictions did not match score performance (Sandholtz & Shea, 2011). Unlike the formative value of the PACT, an examination of the literature demonstrates there are mixed views of the PACT as an assessment that can provide data for licensure and accountability decisions.

Several articles focus on the programmatic value of the PACT. Darling-Hammond (2006) concluded that the PACT was an integral part of assessing program outcomes at Stanford and helped to identify areas for attention across institutions (p. 131). Stanford faculty, Ira Lit and Rachel Lotan (2013) agree in their examination of the PACT and the dilemmas that this high-stakes performance assessment created for individual candidates and for different programs within an institution. The central dilemmas studied are "managing the conflicting values of the formative nature of the work of educators and the summative imperatives of high-stakes assessment" and "reconciling the contribution of high-stakes assessment to curricular coherence and alignment of practice, on the one hand, and the program's perspective of offering a range of competing theories and fruitful practices, on the other" (Lit & Lotan, 2013, p. 55). They find that differences across programs create a "balancing act" of implementing PACT and working with candidates as they complete it. Like Peck, Gallucci, and Sloan (2010), Lit and Lotan found that as professionals and teacher trainers "respect for the professional judgment of a faculty member and the program's commitment to support intellectual diversity were pitted against the demands and priorities of high-stakes assessment" (p. 70) and that this can lead to homogenization. Lit and Lohan write, "because of its high-stakes nature, the Teaching Event becomes something of an albatross-an experience to worry and fret over, rather than an opportunity for thinking, reflecting, and improving on one's practice" (p. 65). While their article is overwhelmingly positive and supportive, these are dilemmas indeed. Reading between the lines, the take-way for validation investigations is that the procedural conditions by which the PACT (and its off-shoot, the TPA) is performed will necessarily be varied. This will impact the inferences that can be made from those PACT test scores

TPA: Creation, Promotion, Adoption

The legislature in California approved multiple measures of teaching performance assessment (i.e., CalTPA, PACT). The version created by SCALE (PACT) was touted as a clear national leader by key educational assessment and policy advocates in the US (i.e., Arne Duncan, Linda Darling-Hammond, and AACTE President, Sharon Robinson). Darling-Hammond (2010) writes that the overarching goal is to provide “a system of reliable, valid, and nationally available performance assessments–from a teacher’s point of entry through the development of accomplished teaching” (p. 3). Such a system would *begin* with “a common tool for assessing novices” (p. 3). For this reason, researchers in California began to undertake the ambitious project of developing one performance assessment to evaluate beginning teacher competence. This one assessment, if used nationally, would provide data on teacher effectiveness that would allow constituents to compare states to each other, compare programs within a state, and candidate growth in the course of their career. The TPA aspires to be that common assessment; a national test of teaching quality and readiness that applies both the NBPTS and INTASC standards. The instrument will be described below. For now, the context for which the assessment was created and adopted is discussed.

The TPA is the first step in a continuum of teacher performance assessments that follow a teacher from the initial licensure through to NB certification (Darling-Hammond, 2010, p. 13). Darling-Hammond writes, “By 2015, a national system of teacher performance assessments will be available for use in policy decisions, ranging from initial licensing to professional licensure and advanced certification” (p. 12). She further clarifies that, “This set of assessments can be used not only for personnel decision making over the course of the teaching career, but also for guiding teacher development and for evaluating and improving teacher education, mentoring, and professional development programs" (p. 32). Such "a reliable and valid system of performance assessments based on common standards would provide consistency in gauging teacher effectiveness, help track educational progress, flag areas of need, and anchor a continuum of performance throughout a teachers career" (pp. 32-4).

That the data from these common assessments would be used for accountability purposes is made plain. "The aggregated data will ultimately be used for program accreditation to provide a basis for deciding which programs should be encouraged, improved, or closed if they cannot improve enough" (pp. 17-18). Darling-Hammond describes seven purposes for the TPA:

1. *Programmatic Improvement*: TPA data can be used to "flag program needs, guide improvements, and track progress" (p. 23).
2. *Decision-Making*: states, districts, schools and IHE can use data to make decisions about "recruitment, employment, professional development, career development" and to report and track data about teacher outcomes to make decisions about their career path (p. 23).
3. *Accreditation*: TPA outcome data can be used to "leverage significant improvements in preparation programs, especially if accreditors adopt an expectation that programs must show a specific level of performance" to maintain or receive accreditation. Darling-Hammond recommends 70% (p. 23).
4. *Mentoring*: TPA outcome data can "guide more effective mentoring for beginning teachers" and shape the process and expected outcomes of a teacher's probationary period (p. 24).
5. *National License*: High scorers on the TPA could be offered a National Teacher License and other recruiting incentives that would facilitate greater teacher mobility to high-need areas (p. 24).
6. *Common Framework*: states, school districts, and preparation programs share "a common framework for defining and measuring a set of core teaching skills that form a valid and robust vision of teacher effectiveness, reflecting both teacher practices and student learning" (p. 24).

7. *Accountability and Policy*: a continuum of TPA can "contribute to the development of a more coherent and comprehensive national policy environment for teacher licensure, recruitment, and in-service evaluation, and ultimately to a more effective national agenda for improvement of teacher quality" (p. 24).

It is clear that policymakers' desire comparative data on teacher effectiveness and that the need for accountability is driving researchers' development of instruments to provide reliable scores for decision making.[13]

Recognizing the need to control the definition, process, and assessment conversation for teacher preparation has led many states/IHE to set aside concerns about performance assessment and endorse an identical measure that collects comparable data that can be used nationally by teacher trainers, educational researchers, and policymakers for decision making. As of November 2013, twenty-nine states and the District of Columbia had adopted the TPA. These high-stakes uses and consequences mean that TPA scores can potentially follow a candidate from the point of licensure throughout their entire career. As a part of the continuum that Darling-Hammond outlined, the TPA is intended to provide data for this purpose. The high-stakes consequences of performance also apply to preparation programs, districts, and states whose accreditation, funding, and local decision-making ability may be called into question if TPA scores do not meet expectations. It seems reasonable to assume that future federal funds will be tied to this data as well. The accountability decisions Darling-Hammond describes depend upon affirming one very important question: Is it valid to use TPA scores for such consequences?

[13] Darling-Hammond (2009; 2006; 1998) argues that experts in the teaching profession are best prepared to define the construct of teacher readiness and should take the lead in developing assessment measures. Cap Peck in *Rethinking Schools* (2010) writes, "PACT represents the most significant attempt to date by teacher educators to take control over the evaluation of our own profession" (para. 3).

Performance Assessments for Licensure and the Standards-Based Movement

Despite broad support in educational circles for performance assessment, many psychometrians, test creators, and researchers remain wary. Literature in educational measurement before 2000 indicates wide-spread concerns about the need for test-score reliability and standardized procedures and testing conditions (Shavelson, Baxter and Pine, 1992; Suen and Davey, 1990). Some question whether performance assessment can adequately cover the construct (Haertel, 1999; Feinberg, 1990). Concerns about resource use abound including time, costs, and scoring needs (Cole, 1988).Though performance is an important part of the movement to standards-based assessment, standards-based evaluation was delayed because validation efforts for performance assessment tools did not always provide data that test scores measured student learning of the standards (Kane, Crooks, & Cohen, 1999). Validity researchers realized they could not employ the same techniques used for norm-based, objective tests to performance assessment because the construct, scoring, generalization, reliability, and judgment inferences were changed by the assessment type (Taylor, 1994, p. 242). The methods used in norm-based tests focused on quantitative models, correlation studies, factor analysis, and other statistical analysis methods. Taylor notes, “The quantitative methods now used to gather evidence for the reliability and validity of assessments within the framework of the measurement [objective, normed testing] model will not function for the standards model” (p. 246). Linn, Baker and Dunbar (1991) called this a paradigm shift that would require new ways to gather evidence, evaluate growth, and include cycles of feedback and revision. In the world of assessment and validation, this shift was not simply a matter of degree, but of a new kind.

A review of the literature since 2000 indicates that performance assessments for use as high-stakes licensure examinations still needs more study (Sandholtz & Shea, 2011; Rennert-Ariev, 2008; Arends, 2006; Delandshere & Arens, 2001). Goodman, Arbona and Dominguez de Rameriz (2008) review performance assessment teaching licensure exams and find that “studies examining the relation between teacher scores and state-mandated tests and school achievement could not be found" (p. 25). They argue that "there is a need to further examine the reliability and validity of measures teacher education programs use to evaluate their teacher-candidates" (p. 37). However, other studies have found that teacher preparation can make a positive difference in relationship to school student achievement (Darling-Hammond, Holtzman, Gatlin, & Helig, 2005). Goodman, Arbona and Dominguez de Rameriz (2008) find that the assumption of "validity is based largely on their construct validity along with a faculty member's professional judgment" (p. 37). In other words, "authentic measures usually require candidates to exhibit in applied settings the knowledge, skills, and dispositions thought to be required for effective teaching. They are, therefore, assumed to be a valid measure of teaching effectiveness" (p. 37). Despite this call for more research, national adoption of the TPA by 2015 is underway with a plan to study it later. The purpose of this study is to examine the 2012 WA TPA field test to determine whether the data from this assessment can support making consequential licensure, accreditation and policy decisions from TPA scores.

The Teacher Performance Assessment Handbook

The TPA is a high-stakes performance exam used to make decisions about teacher readiness. Performance assessments, or any single assessment, cannot possibly cover all of the aspects of a multi-faceted profession. For instance, workplace skills such as punctuality, professional email writing, and maintaining a tidy workspace are likely not evaluated on a performance assessment, especially not one that is high-stakes and standardized. Because performance assessments can only focus on a limited number of aspects of teaching, desired outcomes affect the areas of focus (Stansbury, 1998, p. 16). As stated before, there are multiple intended outcomes for the TPA test scores. Its primary purpose is to provide scores for use in licensure decisions. Secondarily, scores will be used for IHE accountability. Thirdly, the TPA should offer formative growth through learning-in-practice.

These outcomes are repeated by the TPA creator, SCALE. The main purposes for the TPA are described on their website as four-fold:

1. To provide teacher preparation programs with data to make programmatic improvement decisions. Ostensibly, among other things these improvements could impact course sequence, curriculum, faculty assignment, internship length and placement.
2. To provide decision making bodies with data to support licensure decisions for candidates preparing to enter the profession.
3. To provide accreditation bodies, such as CAEP, with data to support accreditation decisions.
4. To provide an assessment system which would improve upon the current quality of teachers prepared in the US (SCALE, 2013).

In the handbook, the purpose of the assessment is communicated to candidates as a "nationally available assessment of readiness to teach for novices" (SCALE, p. 1). Candidates are told to provide "TPA evidence [that] will demonstrate your current abilities, knowledge and skills as a beginning teacher on your way to becoming a highly accomplished teacher" (p. 1). Teacher readiness is demonstrated by evidence that is:

> focused on student learning and is designed around the principles that successful teachers apply knowledge of subject matter and subject-specific pedagogy, develop and apply knowledge of their students' varied needs, consider research/theory about how students learn, and reflect and act on evidence of the effects of their instruction on student learning. (p. 1)

"Assessments signal to teachers and students what is most important" (Whittaker & Young, 2002, p. 46). Candidates complete the TPA in four discrete tasks. On the TPA, the term "task" is used to describe similar teaching activities that have been grouped together; these are performance objectives. Each task has corresponding rubrics, or scoring guides, which detail scoring criteria, or list the expectations and required evidence a candidate must provide, and a scorer must locate, to determine levels of proficiency to rank performance ability. While these four tasks are in all TPA handbooks, the requirements are further specified by subject area or professional discipline. The

TPA is "subject-specific with 27 separate versions for Early Childhood, Elementary, Middle Childhood and Secondary" that correspond with state "licensure areas" (p. 1). For instance, a candidate who is certified in secondary science will be required to take the TPA in that subject area and it will differ in content expectations from that of a candidate pursuing elementary literacy.

Candidates provide multiple types of evidence to demonstrate readiness.[14] Evidence is provided in two basic forms: artifacts and commentaries. Artifacts include documents that explain the context, learning segment, and student work. These would include lesson plans, assessments (both completed and not completed), video clip(s), and/or notes for lectures. Commentaries are candidate responses to specific questions posed within each category of the TPA, called "prompts." Each of the four tasks has its own commentary. Commentary responses are meant to provide the rater with a better understanding of the candidate's thinking before, during, and after the learning segment. Candidates are told that "although your writing ability will not be scored directly, commentaries must be clearly written and well-focused" (p. 2).

Candidate readiness in WA is measured in six categories. Academic Language (AL) and Student Voice (SV) are categories with professional activities that bridge the four tasks. AL and SV have separate guiding question(s), performance level requirement(s), and evidence. However, candidates submit evidence for both AL and SV across the four tasks, rather than discretely. For instance, articulation of the language demands, vocabulary and discipline specific syntax for the learning segment should be included in the planning task, highlighted in the lesson plans, observed during instructional video clip(s) and/or evaluated in the assessment instruments. For this reason, AL and SV are considered "embedded" categories. To summarize, the TPA is a discipline-specific

[14] SCALE identifies candidate TPA evidence as "specific records of practice (evidence)" which "consist of artifacts of teaching (lesson plans, video clips of instruction, student work samples) and reflective commentaries" all of which are meant to "justify the professional judgments underlying the teaching and learning artifacts" (SCALE, 2012).

instrument with multiple measures (six categories and four tasks), each with its own demands for evidence (see **Appendix B** and **C**).

Construct Assessment and Task Alignment

The WA TPA evaluates the construct through fifteen rubrics. Each rubric has a title or “central focus” and a “guiding question.” The guiding question is the construct or performance expectation for candidates. Thus, there are fifteen performance expectations. The majority of the rubrics include only one evaluation criteria. Six of the fifteen (40%) rubrics involve two criterion. Two rubrics (13%) have three criterion. Each rubric is aligned to specific artifacts and commentary evidence, some of which is performance-based, and includes five scoring levels. These levels are the rubric criteria. Rubric criteria specify the extent to which a candidate must prove readiness. SCALE developed each of the rubric levels with shared definitions for “below expectation” (levels 1 and 2), “at expectation” (level 3), and “above expectation” (levels 4 and 5). The passing level is considered scores that earn an “at expectation” ranking. For this reason, it is clear whether candidates have “passed” a rubric but it is up to each adopting state to define what “pass” means for the assessment as a whole (all rubrics and tasks together). At the time of the 2012 state-wide field test, no passing definition for the TPA had been determined.

Evaluating the Assessment

As discussed above, traditional approaches to assessment validation do not always align to performance assessment. However, the design of the TPA allows for a natural comparison between the components of traditional assessment and the TPA, as a performance assessment. The design of this study applies an Argument-based approach to validation which will be discussed in the next section. Before moving to validity, it is important to summarize what is understood about the TPA. Table 2.1 describes the traditional components of an assessment and applies those to the TPA, as a performance assessment.

Table 2.1

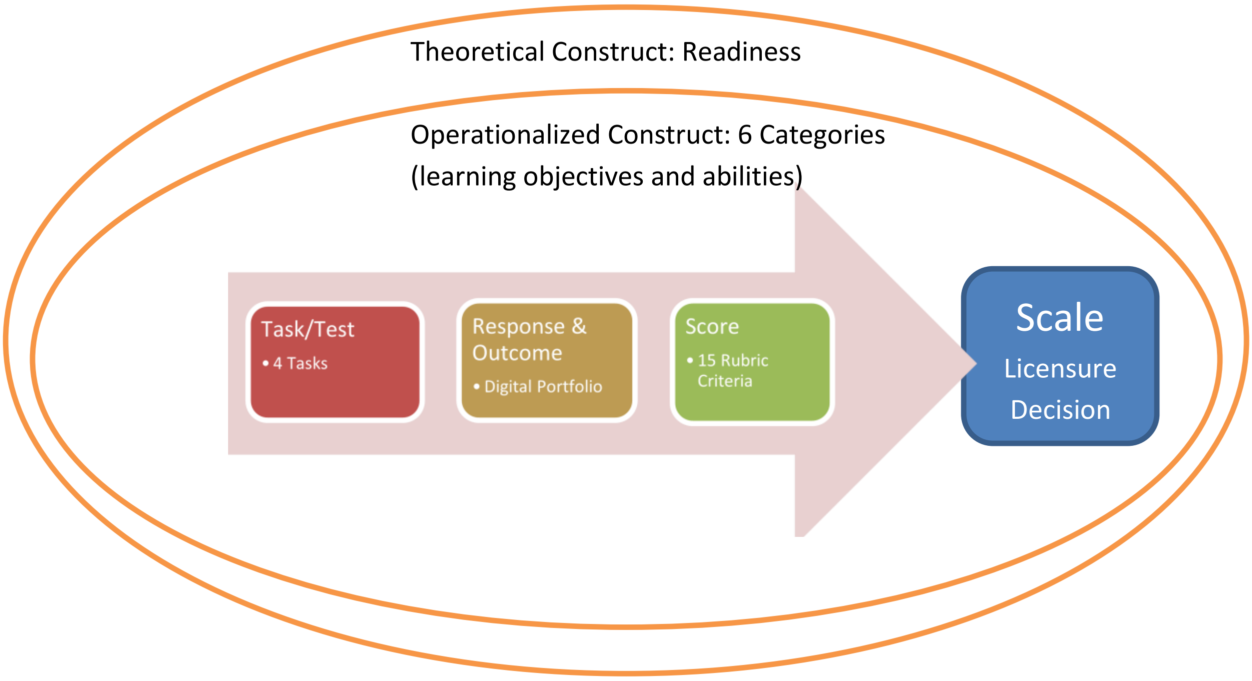
Assessment Components Aligned to TPA

Assessment Component	**TPA**
THEORETICAL CONSTRUCT	Readiness to teach
OPERATIONALIZED CONSTRUCT	6 Categories (learning objectives and abilities)
TASK/TEST	4 Tasks (3-5 Lesson Taught Segment)
RESPONSE & OUTCOME	Digital Portfolio: 8 Evidences in 2 Types - Artifacts and Commentaries for each of the 4 tasks
SCORE	15 Rubric Criteria with 5 Rubric Levels that identify levels of performance at "below expectation," "at expectation," and "above expectation."
Scale	Licensure Decision: WA Cut scores for licensure were not yet determined at the time of the field test. These were adopted in November 2013.

Another way of representing this information is to present the relationship between the broader construct of readiness (theoretical construct), the construct of readiness as it is defined in the TPA (operationalized construct), the Tasks that define what readiness looks like (Test), the candidate responses demonstrating their ability on those tasks (Response/Outcome), the score of those responses (Score) and then the (Scale) decisions made from those test scores (see Figure 2.1).

Figure 2.1

Assessment Components Aligned to TPA


Theoretical Construct: Readiness
Operationalized Construct: 6 Categories (learning objectives and abilities)
Task/Test
• 4 Tasks
Response & Outcome
• Digital Portfolio
Score
• 15 Rubric Criteria
Scale
Licensure Decision

The content being assessed on the TPA is representative of the current understanding of the professional activities of teaching. The TPA was designed to align with INTASC standards for teachers. There are ten INTASC standards, each with performance indicators. Standards one through seven align directly with the TPA requirements. Some of the performance indicators in Standard eight are met by the TPA while others are not. Standards nine does not align with the TPA. Similarly, WA PESB standards for teaching were aligned to the TPA. Nine of the thirteen Standard V requirements can be measured using the TPA. Because the INTASC standards are widely regarded to be the national standards of teacher preparation, such close alignment between the TPA and INTASC standards demonstrates that the TPA construct definition for teaching readiness sufficiently reflects the views of experts in the field. When responding to the research questions of this study, the construct of teacher readiness is that operationalized by the TPA. Whether such an articulation of teacher readiness is sufficient is a crucial question of construct validity and will be addressed in Chapter four.

Validation

A clear understanding of validity, its meaning, its methods, and its results, is a central concern of any validation study. Definitions of measurement and validity have “evolved over time” and each “different version has conveyed different implications for validation practice and the appropriate interpretation and use of results” (P. Newton, 2012, p. 6). Newton (2012) writes that the “classic definition” of validity refers to the “property of a test that characterizes its suitability as a measuring instrument” (p. 3). In other words, “the test was valid” (p.3). Because this definition did not take into account how measurement was modified by procedural fluctuations, the classic definition has been labeled inadequate and even “naïve” (Cronbach & Meehl, 1955).

Cronbach’s and Meehl’s seminal work on validation shifted validation practices away from the classic definition, where validly belonged to the instrument, to one where validity belonged to the interpretation. The classic definition was problematic because it did not account for measurement differences that occur as a result of variations in a) procedures, b) guidelines and administration of the assessment, c) contexts in which the assessment might occur, d) characteristics of the test-taker, and e) uses of results (Cronbach & Meehl, 1955). These differences led to a definition of validity that would encompass the quality of the measurement procedure, in its entirety. Cronbach (1971) writes, “The phrase *validation of a test* is a source of much misunderstanding. One validates not a test but an interpretation of data arising from a specified procedure” (p. 447). Over the next 50 years, researchers’ sought a consensus definition of validity, one that identified validation as involving the test purpose and, specifically, different types of inferences made around the consequences of that test.

The most current, accepted, definition of validity is published by the American Educational Research Association (AERA) in a report called *Standards for educational and psychological testing*, often referred to simply as *Standards.* One of the first set of *Standards* for psychological testing, beyond that of the classic definition, presented validity in four categories: content validity, predictive validity, concurrent validity and construct validity (P. Newton, p. 4). Distinguishing between these different forms of validation helped researchers to situate their various, and often divergent, inferences and choose validation methods that best addressed the types of validation questions required by the inference. By 1974, the *Standards* were revised to three forms of validation: content, criterion-related, and construct. These three types became known as the “trinity” (P. Newton, p. 4) and were widely criticized as allowing validity researchers to “pick and choose” from the types of validation methods, selecting only the type that would best support interpretations. Critics such as Guion (1980) and Messick (1975) argued, therefore, that a new definition was needed. The evolution of validation studies led researchers (Kane, 2006; Messick, 1989; Loevinger, 1957) to argue that the concept which unified all forms of validity was construct validity, or the “focus on evidence needed to support (and challenge) the theory underlying the test score interpretation” (Shepard, 1997, p. 424). Messick is credited with the presentation of a unified validity framework through a multi-faceted view of test validation.

In his influential chapter, Messick (1989) differentiates between evidentiary and consequential bases for validity. Messick's uses the following table to describe validation.

Table 2.2

Messick's Facets of Validity Framework

	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct validity & Relevance/utility
Consequential Basis	Value implications	Social consequences

An evidentiary basis borrows from the earlier definitions of construct validity described by Cronbach and Meehl wherein construct validity includes a "network of associations for propositions" that "lead to predicted relations among observables" (p. 299). Some of what is observed can be referred to as "criteria" and it is the criteria which are under investigation in order to determine construct validity (pp. 299-300). Evidentiary basis includes construct validity, which, as an umbrella concept, also includes criterion-related and content validity. Validity investigations should consider test use inferences which examine meaning (relevance and utility). The second basis by which to determine validity is the consequential basis. Consequential interpretations address both the intended and unintended consequences of test use. These include the value implications that follow from the intended purposes and those test use inferences which have social consequences. Here it is important to note that Messick is primarily interested in consequences that were *intended* by the assessment (P. Newton, 2012; Sackett, 1998; Messick, 1995).

Messick's chapter is one of the most cited and authoritative references on the topic of validity. His work achieved two important milestones in defining and understanding validation investigations. These include: (a) settling the debate around a consensus view of validity which takes construct validity as its unifying conception and (b) articulating a vision of validity which includes not just test score meaning but also consequences of test use (Shepard, 1997, p. 423). Messick's view of validation implies that when undertaking a validity investigation, a full set of questions are required including more than *just* construct validation. Messick's interpretation of validation later became part of the definition described by the *Standards* (1999).

Not all validity researchers agree with Messick (Davies & Elder, 2005; Bellack & Herson, 1984). W. James Popham (1997) disputes that social consequences of test use should be considered a "'facet,' 'aspect' or 'dimension'" of validity (p. 13). Popham writes that there are three problems with a view of validation which include consequences and test uses. First, a clear definition (i.e., simpler) would be better understood and used by educational practitioners (p. 9). In other words, the problem with a complex view of validation is that it will complicate whether educators view "numerical results as 'valid'—that is, as accurate" (p. 10). Second, he states that "cluttering the concept of validity with social consequences will lead to confusion, not clarity" (p. 9). Popham writes, "Messick's 1989 validity framework did, indeed, cut and combine evidence so that social consequences became a key facet of validity. It's just that the price to be paid for doing so is far too high" (p. 10). A view of validity that includes social consequences moves the focus away from evidence and "make[s] it impossible for run-of-the-mill educators to understand what measurement validity is really about" (p. 12). Keeping consequences out of validation investigations will keep the central focus (and conclusions) on the "*accuracy of score-based inferences*" (p. 10). Popham argues that, "test-use consequences should be systematically addressed by those who develop and utilize tests, but not as an aspect of validity" (p. 9).

Paul Sackett (1998) agrees that consequential validity has "come to refer inappropriately to any situation in which unintended negative consequences are observed" (p. 121). Sackett clarifies that the concept of consequential validity, as introduced and explained by Messick, is not "controversial" (p. 120). However, the way in which it has been used by researchers has moved the bar to "consequences *per se*, even if not linked to test flaws" and it is that stretch in practice that becomes controversial (p. 120). His claim is that researchers have to avoid the idea that a test is "valid for me and not for valid for you" (p. 121). To maintain the integrity of validation investigation, especially for high-stakes tests, there is little room for the ambiguous. In summary, the overarching concern is that, by applying consequences "the concept of validity loses its stature as the most important consideration in test development and use" (p. 121).

Robert Brennan (2001) disagrees that academics and teaching professionals would be harmed by a more complex and robust definition of validity. He writes, "It is particularly unfortunate when it is stated or implied that academics have no understanding of the practical realities of testing, and/or those who work for testing companies or federal/state/local agencies are less committed to technical standards" (p. 17). However, despite the complexity of validation, most educational assessment textbooks and educational practitioners still use the classic definition when discussing validation (P. Newton, 2012). Shepard (1997) discusses this problem in her thorough evaluation of the practice of validity. She suggests that test manuals and texts portray construct validity as "not understood," "too complex" or that "its demands are perceived to be too complex" (p. 407). The problem with limiting teachers' exposure to validation practice, or making it sound as though it were so complex as to be nearly impossible for the practicing teacher to engage, is that it gives "practitioners permission to stop with incomplete and unevaluated data" (p. 407) or, worse, a lack of understanding about the meaning of test scores and the decision-making process that surrounds them. Application of validity to professional educational practice requires that the definition be shared with all "members of a profession since it functions as a principle, or guarantee, underwriting the exchange of goods between developers and publishers, on the one hand, and policymakers, users, and members of the public, on the other, most obviously those who are actually assessed" (P. Newton, 2012, p. 6). The problem, as Stephen G. Sireci (2007) suggests, is that many stakeholders, such as policymakers, still view validity as the property of the test. When this is the case, misunderstandings about what validation can "prove" and not prove around test score use could influence stakeholder judgments. This discussion illustrates that the need for a consensus definition of validation is important, not just for validity investigators, but for all stakeholders who utilize and rely upon valid test score judgments.

Shepard convincingly defends incorporating consequences and uses into the definition of validity and validation investigations, a call from measurement researchers for what is often referred to as the unitary concept or consensus definition. She writes, “side effects, which are the unintended consequences of a test used for its intended purpose should not be mistaken for the effects of test misuse” (Shepard, 1997, p. 8). In order to address the concerns of critics, especially those like Sackett, Shepard recommends the work of Michael Kane (1992) who “suggested ways to use an argument-based approach to prioritize validity questions and thereby reduce the burden of validity studies (Shepard, 1997, p. 8). Sireci (2007) points out that the unitary conceptualization of validity has not provided enough guidance as to how to engage in validation studies but that Kane’s approach helps to remedy that problem and that it “has been endorsed” because it “provides an example of how validity can be characterized in a general way, without referring to a construct” (p. 478). While not solving all of the issues surrounding the unitary view of validity, Kane’s approach to validation helps to:

> prioritize validity questions and gather evidence to defend the way test scores are currently used. It acknowledges that validity can never be unequivocally established but also that we need to put forth enough evidence to make a convincing argument that the interpretations made on the basis of test scores are useful and appropriate. (p. 480)

Kane’s validation model, called the Argument-based Approach to Validity, will be discussed in the next section. This framework was developed within the historical context just described and is meant to address some of the earlier concerns about use of validation methods.

Michael Kane (2013; 2009) suggests that “we could decide to make such a change in our terminology and usage” to simplify the concept of validity, but “this would not absolve us of the obligation to develop a reasonable basis for our test-based interpretations and decisions, whatever they are" (2009, p. 45). Kane argues that theory of validity must relate to the practice of validation. Defining validity without consequences and uses would be simpler, but "it would not simplify the basic task of justifying the interpretations and uses of test scores" (p. 45).

Like Kane, Edward Haertel (1999) also advocates a view of consequential validity which addresses both the value implications and social consequences of any interpretation. He writes, "validation should be a process of constructing and evaluating arguments for and against proposed test interpretations and uses" in which "the overall argument is only as strong as its weakest premise" (Haertel, p. 5). Haertel addresses an important consideration for the validation of the TPA in that "as test developers and publishers, sponsors, administrators, and users, we by and large have a practical, and often an economic, interest in supporting and defending tests and testing programs" (p. 6). When examining consequences and uses, this reality can shape the data. As professionals in an age of accountability measured by high-stakes tests, it is more important than ever to understand what validity is, when it is claimed, and to carefully review the source of validation evidence for potential bias.

The battle to limit the definition of validity to a simpler notion without consequences was lost. In 1999, two years after Popham's article, the latest iteration of the *Standards* defined the aim of validity as *construct* validity through interpretation of test scores (1999). The explicit recognition is that "test scores are used or interpreted in more than one way" and that "each intended interpretation must be validated" (AERA, APA, & NCME, 1999, p. 9). Messick's definition of validity was codified in the *Standards*, which now included consequential uses, both value implications and social consequences. Though the question of score meaning, distinct from, or in addition to, score consequences is still contested; it is now well accepted that investigators should consider the consequential basis when clarifying test inferences and test uses. In fact, "issues of meaning and interpretation have become central" to validity (Maxwell, 1992, p. 280).

A consensus definition of validity is still a work in progress. A definition of validation that hinges on interpretation requires that each purpose and use for the test results be evaluated on its own merits. It is important to note that the current version of the *Standards* (1999) states that validity refers to test score interpretations, but not necessarily uses. *Standards* posits that an explicit statement of the interpretation be linked to a rationale for the proposed use (p. 9). In other words, test use alone, without a connection to interpretations, are not clearly the purview of validity. The question for validation design is whether the omission of "uses" was deliberate or accidental. Paul Newton argues that judgments about validity, which extend beyond the procedure and into justification, are "tantamount to giving it [validity] up for adoption" (p. 14). Newton suggests that consequential evidence cannot even be synthesized into judgments about score meaning. The core assertion is that claims about the legality, economics, and ethics of score judgments that are not for validity researchers to make (p. 14). This is one of the debates around the definition of validity that will need to be addressed in the next iteration of the *Standards*. Some recent claims (Borsboom, Cramer, Keivit, Scholten, & Franic, 2009; Borsboom, 2005; Borsboom, Mellenbergh, & van Heerden, 2004) refute the view of validity as one that should include test uses. Borsboom and Mellenbergh (2007) suggest that uses of the test are outside of the test-makers control. They argue that if the stakeholder uses the test to measure data on a group for which the test was not intended, it is not an issue of validation. Borsboom has gone further to say that the overarching interest in validation needs to return to "how the test works" (2009, p. 149) and not its interpretations. But this view remains a minority one in discussions about validation.

The focus of this study is to accumulate evidence to support the assumptions and inferences of the TPA. Because this instrument was already adopted by WA, the study was not likely to be helpful in determining a policy direction with regard to its use. However, the consequences of the TPA are high for candidates, their programs, districts and states and a validation model that would allow for a complex review of the procedures as well as consequences of use were sought. Models from both Messick (1995) and Linn, Baker, and Dunbar (1991) did not offer a methodology to address some of the most essential TPA inferences, notably those differences in consequential accountability measures. The definitive consequences and uses of the TPA are still being determined and developed by policymakers and educators, and this is an important ongoing process to remain relevant to the evolving social practice of teaching. The TPA is a complex instrument with even more complex inferences, including multi-construct interpretations and multi-level constructs. Choosing a method to validate score inferences and interpretations requires an argument that can clearly and adequately identify score meaning. When the uses and consequences of the exam are as clearly stated, and as high-stakes, as those of the TPA, and the consequences of the scores are a matter of disparate and diverse interpretations, consequences and uses should be a part of the validation of the decision-making process. Users and stakeholders will likely view any conclusion as one that determines the plausibility of these uses.

Whether an instrument can reasonably be said to provide data on its stated uses for the purpose of decision-making is the central question of validity. Are the decisions derived from TPA scores appropriate? In particular, are the decisions made from the TPA data sound and plausible? The TPA is an example of a single test with multiple purposes, inferences, and uses. It is both a multi-construct and multi-level exam and its consequences are high-stakes, not just for the test-taker, but for stakeholders who may have never come into contact with the test-taker. It is important to note that any purpose for the TPA includes an accountability measure to policymakers and the public. Therefore, the validation framework selected needs to apply a definition of validity that can produce claims for accountability decisions and these claims need to mean what their stakeholders believe them to mean.

Argument-Based Validation

Given that performance assessments are increasingly used, often for high-stakes decision-making, and that identified issues with their use continue to be problematic, it is important to be as clear as possible about the consequences of any adopted performance assessment and about the types of evidence needed to validate decisions made from their scores. To help researchers develop practical, informative and “do-able” validation investigations that incorporate the right kinds of evidence, Michael Kane (1990) developed the Argument-Based Approach to Validation Studies (ABV). An ABV investigation starts with the premise that inferences and uses of test scores have to be part of the evaluation of an instrument. This approach to validation was developed to provide a “realistic and pragmatic framework” to link “kinds of evidence needed to validate a test-score interpretation to the details of the interpretation” (Kane, 1990, p. ii). Such a practice is implicitly endorsed through the *Standards* (1999) definition and discussion of validity. As an approach to validation investigations, ABV “explicitly associates validity with the plausibility of various assumptions and inferences involved in the interpretation” (p. 2). Such guidance is critical because validity depends on use of the correct procedures and evidence, based on the assessment inferences. Creating a model that structures the process turns an abstract concept into a practical guide.

Kane’s ABV is an argument of practical reasoning. Arguments of this nature tend to be interdisciplinary and address “practical affairs” (Kane, 1992, p. 527). Practical arguments focus on clarity, coherence, and plausibility using parallel lines of evidence and counterarguments. Evaluating practical interpretive arguments requires theory based inferences, observation of the measurement procedure, generalization, and extrapolation. Investigators are often seeking triangulation and looking for patterns and redundancies in the data. “This view of validity is conceptually very simple, but it can be demanding in practice, because it requires that test developers, users, and validators achieve a high degree of clarity about what they are doing, and this can involve a lot of serious analysis and hard work” (Kane, 2009, p. 50). Kane often describes validity as simple, but complex. Validation studies are made even more complex when, “the test scores are used for multiple purposes and have a complex interpretation (e.g., in terms of constructs, like verbal aptitude or intelligence), with many built in assumptions and associations” (p. 50). The more complex the study, the more clarity is required. An ABV investigation involves several steps meant to clarify the inferences around test scores and then collect evidence to support or challenge their plausibility.

ABV is a two-step process. The first step, called the Interpretive Argument (IA), specifies the proposed interpretations and uses of scores (Kane, 2001, p. 4).[15] The second step is the Validity Argument (VA) which, "evaluate[s] the overall plausibility of the proposed interpretations and uses" (p. 4).[16] In his early work, "An Argument-Based Approach to Validation" Kane (1990) describes ABV as an "attempt to move toward a technology for validation" (p. 37)

ABV is an approach to validity, not a type of validity. It can apply any type of validation evidence needed to address the construct. It is important, therefore, when constructing the IA, to identify evidence required for different interpretations. Descriptive interpretations discuss test scores as an "estimate of some variable for the examinee being tested" without specifically identifying test score use (Kane, 2002, p. 32). Decision-based inferences are those that involve decisions about the test-taker. Both are value judgments about the test-taker and the test. Most decision-based inferences begin with a descriptive inference. As is the case with the TPA, is it possible to have wide-spread professional agreement on the descriptive inference (teaching standards) but not on the decision-making inference (accountability measures). The IA and VA must carefully articulate and separate these two types of inferences in order to determine whether there is a challenge to validation.

Inferences are based upon assumptions. Kane distinguishes two types of assumptions that typically influence inferences. The first type are semantic assumptions. Semantic assumptions trigger

[15] Also described as "formative stage" (Kane, 2004; December 1990) and the "developmental stage" (2009).

[16] Also described as "evaluative stage," (Kane, 2004) "summative stage" (1990), and "appraisal stage" (2009).

semantic inferences that draw conclusions about descriptive variables present in or from test scores (p. 33). Semantic assumptions "make claims about what the test scores mean" (p. 33). Policy assumptions are those claims about the consequences of test scores. Policy assumptions are the basis of policy inferences. Policy inferences involve the leap from test scores to conclusions to decisions to judgments. In other words, policy inferences are the decision-rules for test scores. Any evaluation of policy inferences involves an evaluation of the consequences and uses of the test scores. ABV for the TPA involves both descriptive and policy inferences and will, therefore, also examine the consequences of the accountability and licensure decisions based on those test scores.

Once the purpose of the assessment, including its proposed consequences, have been clearly stated in the interpretive argument, a design is made to collect the evidence required to validate that interpretation. To do so, Kane discusses a chain of inferences, or reasoning, with links to represent significant validity and reliability questions central to any study. When using ABV, the researcher should focus on the weakest links in the argument (Kane, 2006, 2004, 2002, 1999, 1994). Typically, these weak links become obvious in the development of the IA, although sometimes they are revealed in the validation process. Weak and strong links are predictable *based on the assessment type*. For instance, all performance assessments share a weak link in generalizability across test scores because of variability in the assessment procedure. Likewise, all objective tests share a weak link in that they are less likely to mimic performance in practice. Performance assessments of teacher readiness (professional competence) ask teachers to apply their knowledge, skills, and dispositions of teaching practice to an actual teaching encounter. The direct observation of performance is the most authentic and preferable for evaluating teacher readiness (as compared to simulations and objective tests). Therefore, when adopting performance assessment, actions should be taken to minimize the predictable and problematic nature of the known weakest links in the chain of inferences.

Measuring teacher readiness is “complicated, but our models need to be simple in order to be manageable” (Kane, 2011, p. 12). Using ABV, and applying the chains of argumentation requires “at least three inferences: evaluation, generalization, and extrapolation” (Kane, 1992, p. 169). Evaluation entails a judgment of the performance to assess whether it was good, bad or in between. In order to evaluate observations of a candidate’s performance, the criteria for judging its quality will be specified and likely include a “focus on effectiveness and efficiency” and “on the avoidance of any harm” (Kane, 1992, p. 169). The score given is a “direct observation of performance” in one moment in time (p. 171). For instance, if a candidate scored 325 correct answers out of 516 possible questions on an exam, with knowledge of the scoring criteria, one would know how the candidate’s performance was scored. Making meaning, and decisions, from that score involves something beyond the specific observation of that performance and assessment procedure. This is the generalization inference and it is basically an inference of reliability. Generalizability assumes that on future performances on the same assessment, the candidate would earn roughly the same score, even if taking the test in a different setting with different professional encounters. The score alone cannot provide its meaning; assigning a score is independent of providing meaning to that score and both of these actions are separate inferences in the IA. The first inference, assigning a score, is an evaluation inference. The second inference, assigning the score a meaning, is a generalizability inference. Finally, score judgments must be situated within the frame of the testing objective. The third inference, extrapolation, assumes that the score can be applied to a broader, or even different, situation. Extrapolation inferences are significant because the purpose of giving a test is to provide data to support a decision around a testing objective. In the case of the TPA, the testing objective is to determine whether a candidate is ready for teaching practice. If the TPA scores cannot be extrapolated beyond the context in which the test was conducted (student teaching), the decision for licensure will not be sound because one cannot credibly know whether the candidate is ready to teach in any context in which the license applies.

Chains of Evidence: ABV and Performance Assessment. Kane refers to performance assessment as “direct observation” assessment (1992). The TPA is a performance assessment and, as such, relies on evidence of observed performance. Observations of professional performance in real classrooms, and with real students, is the most direct approach to determining professional competence (Kane, 1992, p. 172) and the extrapolation inference is likely to have high-fidelity. However, as is the case with all assessment methods, the inferences made from performance assessments may still be problematic. In several articles, Kane (2004, 1999, 1992) examines the chain of inferences from evaluation to extrapolation for performance assessment and warns that measures must be taken to protect against challenges in the evaluation and generalization inferences. In the validation of performance assessments of professional competence, “extrapolation is usually the strongest link . . . evaluation can be a problem, and generalization is almost always a problem” (Kane, Crooks, & Cohen, 1999, p. 173). It is important to note that when conducting a validation inquiry using ABV, seeking out the weakest links in the IA is vital. If the validity challenges discovered in the validation argument can be addressed, these inferences can be improved and strengthened. Kane describes how the weak links in the IA are central to the VA in this way:

> It is convenient, but ultimately misguided, for advocates of performance testing or high-fidelity simulations to ignore issues of generalizability and scoring problems, just as it is convenient but misguided for advocates of objective testing to focus their attention on the objectivity of scoring and on the generalizability of the resulting scores. We all like good news and feel some inclination to shoot the bearer of bad tidings. But in evaluating assessment procedures, it is important to play devil’s advocate. Claims about the validity of performance tests and high-fidelity simulations cannot be accepted without evidence indicating that the scoring is defensible and that the results are generalizable, no matter how realistic, natural, or authentic the assessment. (Kane, 1992, p. 181)

It is particularly important not to conflate what Sackett calls “two distinct problems: evaluating a person and evaluating a work product” (Sackett, 1998, p. 118). To avoid this problem, one of the first links in the ABV chain of inferences is evaluation.

Evaluation: Kane notes that “the assignment of scores to performances in real practice settings involves some serious problems” (1992, p. 172). First, evaluation requires consensus with regard to “best practices” and pedagogy and where experts disagree there is likely to be errors of measurement and a potential for incoherence in the definition of the operationalized construct. It is likely that scorer disagreement about the action a candidate should have taken will result in difficulty when scoring “quality.” The advantage of the TPA as an assessment of teacher readiness is that it asks for a sample of teaching practice in the context of actual teaching. The TPA tasks ask candidates to work in complex, realistic situations. However, these classroom-based settings demand the most of scorer judgment and “pose the greatest difficulties in evaluating performance” (p. 172). Inter-rater reliability is a prerequisite for trustworthy scores and this requires intense, costly, and time consuming scorer training and monitoring. Similarly, variability in the authentic settings that may be present for different candidates sitting the same assessment makes it necessary to create more generalized criteria that can account for a “wide range of situations that may arise in actual practice” (p. 172). The more general the criteria, the more judgment is involved both for the candidate in interpreting those criteria and for the raters in scoring the sample using that criteria. Specific rubrics and better trained and experienced raters reduce bias and subjectivity, but Kane notes that “these problems cannot be completely eliminated” (p. 172).

Generalization: One area in which critics and proponents agree about performance assessment is that it is expensive to administer and to score. As stated before, the strength of the TPA is that it is an authentic assessment. Unfortunately, observing performance in actual classroom settings is inconvenient, time consuming, and expensive. These factors influence the generalizability of scores. Generalizability inferences are influenced by the samples of performance because they are typically (a) rather short in length or small in number, (b) collected over a limited period of time, and (c) limited to the number of contextual factors present in that given sample. In addition, because the time spent scoring observed performances is extensive, the number of raters is typically low (one or two). Simultaneously, the level of control the candidate has over the situation in which they will be evaluated is often very limited (by placement location, curriculum taught as determined by the district or test due date for the learning segment, and classroom population). Candidates are often placed in a student teaching classroom not of their selection but based on the availability of the mentor teacher, school administrator, and school district. Allowing for the level of variation in the contexts with which two otherwise equal candidates will complete the TPA requires that the rubric criteria used to evaluate each candidate be general enough to account for those variations. The very need for this level of generalizability will create “substantial errors of measurement” (p. 172).

Adding to the problematic nature of generalizability inferences in performance assessment is the concern that the sample may not be representative (or reproducible). The logistics of student teaching dictates that the choice of encounters for candidates is determined by their placement, which is determined by the university they attend and the district where they are placed. Taken together, the validity challenges around generalization likely constitute a weak link in the TPA because it is a performance assessment. Determining readiness based on a small sample of observed performance drawn from a single setting (or one lesson) where students are likely to have much in common with each other (they are assigned schools based on their geographic proximity) to a universe of encounters generalizable to the larger domain of “teaching effectiveness” is problematic and may represent a serious threat to validity if not properly addressed (Kane, 1992, p. 173).

Extrapolation. Of the three central links in the chain of inferences, extrapolation is likely to be the strongest for performance assessment. Observation of performance for assessment purposes in authentic, real, actual settings in which that professional performance is practiced makes an inference of teacher readiness highly probable, in theory. Despite this, there can also be validity challenges present in extrapolation inferences of performance assessment. Kane notes that simply observing performance can cause “subtle and not so subtle influences in the quality” (p. 173). The TPA, which requires that candidates record and submit timed sections of their taught segment, adds an additional element of observational influence: the video camera. Students may be accustomed to having other adults (besides the teacher) in the classroom. Certainly, the school administrator would have periodic visits, but other guardians, school board members, teachers and interested parties are often welcome to visit classrooms. Not so typical is the introduction of a video camera. Some students may perform better simply because of the recording out of solidarity with the candidate or because of other conditioning, while other students’ behavior may deteriorate if they become self-conscious or seek additional attention. Class participation may be enhanced or non-existent simply because a video camera is turned on in the back of the room. While students can be conditioned over time to accept the presence of the video camera, it is likely to still have an impact on the candidate who is intently aware of being observed and evaluated. These issues constitute challenges to the extrapolation inference for any performance assessment that requires recorded teaching and an electronic submission.

Errors of Measurement and Standardization. Some of the links in the chain of inferences introduce errors of measurement (EoM). EoM are the differences between a measured value and a true value. Typically, EoM are not errors in the assessment instrument nor the design study. However, they influence the data and how it might be interpreted. For these reasons, EoM cannot be totally avoided, but they can be reduced. There are two types of EoM, systematic and random. “*Systematic errors* are constant across some set of scores (e.g., all scores for a particular person or occasion or test form) and are therefore potentially predictable” (Kane, 2011, p. 15). Random errors, conversely, are based on statistical fluctuations that exist when measured values are inconsistent and these cannot be predicted. One way EoM are minimized is through standardization.

Kane (2011) discusses the importance of standardization and its limitations in reducing random and systematic errors. Higher degrees of standardization have a way of decreasing random error because standardization seeks sameness by reducing differences in the testing procedure. However, standardization can increase systematic error because performance assessments are conducted in unique contexts that are harder to generalize across a set of scores. Paradoxically, standardization in the testing procedure can promote fairness but, in some cases, that can also decrease fairness. When it comes to standardization more is not always better. “Standardization is less effective for absolute interpretations” and can “increase the overall error in some cases, especially for absolute interpretations (p. 18). However, “standardization tends to promote fairness and the appearance of fairness,” important factors when considering a high-stakes assessment (Kane, 2011, p. 17). Kane suggests that standardization works well in many contexts, but not all. In some cases, standardization practices can prove problematic, especially in large scale educational accountability assessment systems such as the TPA, where there is a rub between the nature of the type of assessment selected and the need for high levels of standardization in the assessment procedure (p. 26).

In addition, systematic errors associated with standardization tend to increase when the test consequences are high-stakes. For instance, accountability comparisons across programs and states are likely to be influenced by how well programs prepared candidates to take the test, as separate from preparing candidates to teach. When significant differences occur in test preparation and procedures, the “resulting systematic errors can be especially serious for the kinds of absolute interpretations (e.g., in terms of achievement levels) typically employed in accountability systems” (p. 26). Some standardization is necessary, but it is not the fix-all for errors of measurement (Kane, 2001). In fact, efforts to standardize a performance assessment may reduce its benefits as an authentic instrument because “even modest levels of standardization are difficult to implement in real practice situations, and these efforts may tend to make the performance assessment somewhat artificial and contrived” (Kane, 1992, p. 172). As observations of performance become more standardized, they can be seen as less representative and accurate. Procedures to improve standardization are important for reliability, efficiency, scoring, and, especially, fairness, but may contradict the benefits of authentic, context-specific and unique characteristic of performance assessment. The role of standardization, whether there is enough or too much, should be clarified in the evaluation, generalizability and extrapolation inferences and included as a part of the evidence collected in the VA for performance assessments.

Validating Measures of Performance. All assessments have stronger and weaker links. Performance assessment, because of its strength as more easily extrapolated to the larger domain of teacher readiness, is preferred for decisions of teaching licensure and certification in the US. Cizek (2001) points out "there is simply no way to escape making decision about students" (p. 21) and decisions about which teachers should be licensed and which should not is one of the important decisions to be made if we are to preserve an educational system that promotes democratic activism and equity for all learners. The stakes are high and somewhat problematic. Goodman, Arbona and Dominquez de Ramirez (2008), state the challenges when they write, "if the test results are not valid with respect to evaluating candidates, they create the potential for individuals to be certified who really do not exhibit competence in authentic settings" (p. 26). However, "another concern is that failure to pass these high-stakes, minimum-competency tests could eliminate otherwise qualified candidates from the teaching profession" (p. 26). Democracy in the US is dependent upon active, participatory citizens who can read, write, analyze, and defend positions on issues. We must know that our teachers can develop learning environments in which student achievement leads to, at minimum, the maintenance of the status quo. In validation terms, "the end point of our chain of inferences consists of conclusions about readiness for practice in an area, and this end point is fixed, in the sense that we do not want to change our conception of readiness for practice or our standards of performance just to make the development of the assessment procedure easier" (Kane, 1994, p. 141).

It seems reasonable to determine readiness to practice in a specific professional field, "in terms of the extent to which a candidate is prepared to manage the situations that arise in practice" (Kane, 1994, p. 139). LaDuca (1994) describes this as "professional encounters" but most licensure or certification exams do not ask that the samples of performance on these tasks be samples of actual practice (Kane, 1994). In this way the TPA is unique. For instance, we do not ask lawyers to demonstrate their readiness to practice law by winning an actual court case. There are many reasons why high-stakes, authentic performance assessment have not been more widely adopted, including difficulties in designing studies that can provide data that verifies the judgments and decisions around test scores. Kane describes the issues in asking for specific, real-life samples of actual practice. The central concern is that there is "great variability in clients and their problems" (p. 139) and we see this in teaching as well. Candidate placements in different districts which practice different levels of scripted curriculum, schools with differing levels of socio-economic status, and diverse teachers and supervisors with different levels of experience both teaching and mentoring novices (Moir, 2013; Moir, 2012; Sandholtz & Shea, 2011; Darling-Hammond, 2006) present variations in the testing procedure. These concerns need to be a part of the validation design. For the purposes of this study, Kane's definition of validity, which includes the uses and decisions that follow from the interpretation, was adopted. Robert Brennan (2001) notes that the work that Kane has published "offers the greatest hope for bridging the gap" between the seemingly simple theory of validity, and the complex practice of validation (Brennan, p. 12). Using Kane's ABV model allows for questions about how context impacts performance to be a part of the study design, which makes it preferable for high-stakes, large-scale performance assessments of teacher readiness. As a model, ABV has been successfully applied to defend the plausibility of assessment inferences in several validation studies.

Several studies have applied ABV to language acquisition assessment (Chapelle, 2011; Chapelle, Enright, & Jamieson, 2010; Ryan, 2002), and several validation studies of high-stakes, large-scale tests in education have applied ABV. Shaw and Crisp (2012) used Kane's (2006) framework to examine International A level Physics exams in the UK (Shaw, Crisp, & Johnson, 2012). Gotch and Perie (2012) apply Kane's model to local assessment systems using school districts in Pennsylvania state (Gotch & Perie, 2012). Bennett, Kane and Bridgeman (2011) use a modified ABV framework to evaluate the use of formative evaluations as a part of a summative assessment system designed by the Partnership for Assessment of Readiness for College and Careers and the Smarter Balanced Assessment Consortium. In the case of the Bennett study, a modification to Kane's original 2006 framework was made suggesting that the inclusion of a Theory of Action (Bennett, 2010) would augment validity arguments, specifically looking at formative assessment. Most of these tests have been designed for a P-12 audience, as a part of the testing systems adopted for public education. None of these tests were performance-based. At the time of writing this study, no ABV study of the TPA, or its predecessors the PACT, BEST, or Texas Portfolio, had been published.

Data to support the validity of assessments based on standards requires the collection of several types of evidence and requires multiple types of research. Because professional assessments of readiness take as a part of their construct the predictive nature of the test scores, ongoing research must be conducted. These should include: collections of candidate work to understand the processes by which they understand and practice pedagogy, questionnaires to understand the breadth of candidate thinking, and interviews to understand the depth of candidate thinking. In addition, studies of different sub-groups, contexts, and time-periods must be undertaken to better understand the consequences of the assessment and any biases present. The evaluation of standards, using performance assessment, introduces more variables in the testing experience and, unlike measurement of objective norm-based test comparisons, requires that validation practices used incorporate more of the candidate experience. Validators have, since Cronbach and Meehl, understood that there is a distinction between type of test and the validation method used. The validation model selected influences our ways of thinking about the learner, the tasks, the construct, and measuring validly and reliability (Kane, Crooks, & Cohen, 1999; Taylor, 1994).

Limitations of ABV. Even among those proponents of a broad, unified, definition of validity, ABV is not without its detractors. Several researchers have questioned the practicality or even the prescriptive nature of ABV (Schilling, 2004; Briggs, 2004). Haertel (2004) does not object to

ABV, though he does take issue with the notion that ABV can verify direct observation assessments for certification (p. 194) without also demanding that content validity be present in the model. Similarly, Talbot and Briggs (2007) argue that, when applied, Kane's validation approach can lack "substantive knowledge of the phenomenon being measured" (Talbot & Briggs, 2007, p. 207). Mainly the objections arise because Kane's approach takes the IA as the theory to be validated, and does not demand that a separate content construct be identified unless it is already part of the IA. Talbot and Briggs suggest that two "amendments" to Kane's approach be added. These include "(1) a clearer distinction between assumptions and inferences in the formation and evaluation of the interpretive argument; (2) breaking up the interpretive argument into what the authors describe as elemental, structural and ecological pieces" (Talbot & Briggs, 2007, p. 205). They go on to say that "each of these pieces is then associated with specific methods for gathering the evidence needed to support a validity argument" (p. 205). This builds on an earlier argument from Briggs (2004) suggesting that Kane's model include a third step, design validity, implemented before the development of an interpretive argument. Kane (2004) responded to these arguments reiterating that, while complicated, interpretations and uses of test scores deserve to be clearly stated in an effective validation framework that "can foster dialogue among stakeholders on the merits of specific inferences and assumptions" (Briggs, 2004, p. 199) but that the content construct need not necessarily be the overriding theory. Kane addressed most of these criticisms in his piece in *Educational Measurement* (2006) which clarified the purposes and approach of ABV and addressed errors in previous applications of ABV that did not incorporate the construct with enough clarity.

Using ABV for the present study is not without its disadvantages. First, in order for ABV to successfully offer the plausibility of the IA, the interpretive argument must be stated very carefully and clearly. However, this process is time-consuming and often difficult because the assumptions underpinning the uses of a test can be general and implicit. Kane acknowledges this when he discusses the development of the interpretive argument as a "formative" stage that is "likely to be stated in very general terms" and that "[m]uch of the interpretive argument tends to be left implicit on the assumption, presumably that the details are self-evident or unimportant" (2004, p. 141). It may take several iterations in the research process before an interpretive argument can be fully understood. It is not simple. "Getting agreement on the interpretation and use is not always an easy task and may require extended negotiations among stakeholders" (p. 142).

Secondly, a validator who is not also the test creator or a stakeholder in the "negotiations" above can find ABV a frustrating approach to validation studies because articulating the interpretations, assumptions, and uses for the instrument requires "getting inside the head" of the test-maker. While the interpretive argument, the assumptions, and the inferences are often stronger when developed by a validator who can see outside the assessment tunnel, if the test-maker is not accessible (or the assessment procedures, the scoring, and the test-maker are represented by separate organizations) and the validator must rely on written documents (or press releases) to determine parts of the interpretive argument, ABV is made more difficult and less applicable. This is especially true if suggestions for improvements are not sought by the test-maker.

Finally, even when an interpretive argument is stated, it is not always clear that ABV offers a "methodology" or design for a validation study. ABV requires that the researcher take some leaps to get from the development of an argument to a study design. In addition, as a comprehensive approach to validation, responding to the assumptions and the questions they provoke for each inference from each occasion of the test is a limitless endeavor that, if not carefully managed, can consume the researcher.

Advantages of ABV. There are several advantages of Kane's ABV framework for the present study. The advantages to applying the ABV model to performance assessments for professional judgments include:

1. It is highly tolerant (Kane, 1992, p. 534). As a model, it can be used for any type of assessment or type of evidence. The TPA as an instrument is complex and requires a validation argument that can simultaneously look both broadly and narrowly. Because the TPA involves a variety of stakeholders and has a high-stakes impact on a large and diverse

population, the flexibility of ABV was necessary and key in its selection. Likewise, there is no preference made to include a specific type of data which allowed for a multiple-method design incorporating both correlational studies and personal narrative, both of which are central to understanding the proposed uses and value of the TPA in WA. Rather than a focus on one type of evidence, data is judged based on how well it responds to the inferences, assumptions, and uses as described in the interpretive argument (p. 534).

2. It does not "have to be associated with formal theories" (p. 534) which also allows it to be associated with any theory.
3. It "provides a basis for deciding on the kinds of evidence needed to validate a particular interpretation" (1990, p. 37).
4. It allows for a way to gauge progress preventing a never-ending, unwieldy investigation (Chapelle, 2011; Kane, 1992).
5. It increases the likelihood that the validation study results can lead to improvement in educational measurements, especially the assessment procedure for the assessment studied. It is productive with an aim to improve the procedure in order to make it more plausible (Kane, 2006).
6. "The term argument emphasizes the existence of an audience to be persuaded, the need to develop a positive case for the proposed interpretation, and the need to consider and evaluate competing interpretations" promoting a long-term view of validation practice that incorporates the needs of multiple, diverse stakeholders (1990, p. 37).

Finally, ABV is a preferred framework because the data and research can ultimately lead to the improvement of the instrument, rather than simply a critique of its uses. Because the TPA is a mandated instrument in WA, the overall goal of this study is not to eliminate or replace the TPA but to confirm that its claim to measure teacher readiness is strongly plausible and if/where it is not, to offer data to improve the TPA. Such an approach to validation studies can help to eliminate a bias either in support or against its use (Kane, 1992, p. 534). Because it can be used with high-stakes assessments, assessments that are a part of a larger accountability system with multiple stakeholders, and performance assessment, the ABV framework is ideally suited for an examination of the TPA.

Conclusion

Teacher educators have established a set of criteria to determine professional competence and readiness to teach. Over the past 30 years, an emphasis on evaluating those standards by using performance assessment has led to the widely acclaimed and adopted assessment for determining readiness, the Teacher Performance Assessment. Washington, with its emphasis on these shared criteria, as well as Student Voice and equity, has adopted the TPA for use in determining whether a teaching license should be granted. This practice has led to questions about the validity and reliability of the consequences and uses of TPA test scores, both for candidate licensure and for accountability of teacher training programs.

When it comes to validation studies that address consequences and uses, there are "two kinds of questions to ask of any procedure: *can* it be used as intended (a technical question) and *should* it be used as intended (an ethical question)" (P. Newton, 2012, p. 14). This study suggests that, in some cases, dissecting the consequences from the purpose of the assessment is a complex task. How do we validate an assessment whose evaluation question includes a construct of accountability, itself a consequence of test score use? Validation as a contribution and practice depends upon a shared definition of what validity means and upon what such studies should focus. Any validation inquiry based upon a weak theory (IA) and too narrow a definition of validity will likely leave the end user without a true measure of whether their test-score-based-decisions are appropriate in the context in which they are being used. The benefit of ABV is that it allows the validator to select the evidence that is most needed (that most challenges the weakest links in the chain of inferences) based on the theory of validation required for those inferences based on the type of assessment studied and the use of the scores from that assessment. ABV applies the most challenging consequences, best types of evidence, and most robust validation inquiry based on the inferences *unique* to that assessment procedure.

Kane reminds his readers that "The users of test scores have the primary responsibility for evaluating the decision that they adopt" (Kane, 2009, p. 62). As this is the case, users also have a responsibility to select the theory and model that will provide them with the most robust validation study. Because the TPA is a high-stakes performance assessment of teacher readiness, itself a fairly complex construct, ABV provides the best model for this undertaking. The TPA is a high-stakes measure for many diverse constituents and users. Therefore, it is even more important that the inferences and uses of TPA test scores be based on a validation study and not "beg the question" of validity and reliability (Kane, 2009).

This chapter outlined the standards currently used to determine teacher readiness in the US and the movement to assess those standards using performance assessment. The strengths and challenges of performance assessment, both as a measure of assessment and as an assessment in a validation study, were discussed. The TPA, as the performance assessment examined in this study, was described. Finally, a brief overview of the history of validity, its definition and practice, and the current method of Argument-based validation inquiry were described. In the next chapter, the research design will be specified and connected to the methodology and process employed to address the primary research question: Is the TPA a valid measure for determining teacher readiness?

Chapter Three

Research Design and Methodology

This chapter discusses the research design, methods, and process used for this study. The literature outlined in Chapter two demonstrated the need for scholarship addressing the central research question of this study: Is the TPA a valid measure for determining teaching readiness?

An Argument-based approach to validation (ABV) developed by Kane (1990, 1999, 2002, 2004, 2006) was used as the study theoretical framework. The first section of this chapter reviews assessment validity, ABV and the work by Shaw and Crisp (2012) as the framework to guide the reporting of the validation argument evidence. The second section provides the interpretive and validity arguments. The validity argument, research questions, population and setting of the study, as well as sources of validity evidence, are outlined in the third section. For each of the phases of data collection, the sources, sample size, and questions addressed are described. Data collection methods are explained and connected to the argument framework. Finally, methods of data analysis and the limitations of the methodology are addressed.

Validity and Validation

The basic premise of validity is that a test should measure what it claims to measure. Tests do not exist in a vacuum but as a part of a system of feedback used to make decisions and claims about the test-taker. If a test does not measure what it claims to measure, it is not of value to its stakeholders who will use the scores to make decisions. This is the case for all testing, but decisions made from high-stakes measures have more serious consequences such as the TPA, which will be used to determine professional licensure. Validating the use of TPA scores—that is, the consequences and decisions made from the TPA scores— is the objective of this project. Such validity questions are important for test-makers and constituents, but the higher the stakes, the more imperative it is that the instrument claims are strongly plausible.

Determining test validity has evolved significantly from a focus on criterion-related evidence to content-related and then construct-related evidence (Kane, 2004, pp. 136-8; Kane, 1990, p. 8).

The current edition of the *Standards* describes validity as "the most fundamental consideration in developing and evaluating tests" (AERA, APA, NCME, 1999, p. 9). It is important to note that even the *Standards* are consistently evolving and the latest iteration includes a serious discussion of the role that fairness plays in test validity (Xi, 2010; Kane, 2010, Kolen, 2010).

Argument Based Approach to Validation

Michael Kane (1990, 1992, 1999, 2004, 2006) claims that validation should be an argument of the plausibility of the uses of test scores based on two types of argument (1990, p. 9). These include an interpretive argument (IA) which lists "the assumptions and inferences involved in the interpretation of test scores" and "the evidence to support the interpretive argument," which is the validity argument (VA) (pp. 9-10). The IA dictates the types of evidence needed to argue for plausibility (validation) (p. 9). Kane (1992) describes the validation argument as "an approach to validity rather than a type of validity" (p. 534). He recommends a methodology or "measurement procedure" flexible enough to allow for all three types of validity evidence (criterion, content, construct) and that evidence used by the validator is selected from inferences, assumptions, and weakest links in the IA (see Chapter two).

Cambridge Approach to Structure ABV

Shaw and Crisp (2012), in their ABV study on GCE A levels in the UK, propose moving from existing frameworks toward a structure for the argument of assessment validation (p. 167) which I will call the Cambridge Approach (from the home institution of its authors). Adopting Kane's terminology of "links" to describe the validation argument process, Shaw and Crisp describe five central inferences for any assessment validation. These include Construct Representation, Scoring, Generalization, Extrapolation, and Decision Making. The present study has adopted those five inferences and their description. Application to the TPA will be discussed later in this chapter (see Framework). However, the nature of performance assessment is that it measures professional activities in professional contexts and so it is likely to have high extrapolation and weak generalization evidence. For this reason, emphasis will be placed on the evaluation and generalization links in the chain of evidence.

"It is now a widely accepted view that validity is concerned with the appropriateness or correctness of inferences, decisions, or descriptions made about individuals, groups, or institutions from test results" (Shaw, 2012, p. 162). For this reason, the advantages and applicability of ABV as an approach for validating the uses of the TPA were clear and this is the framework that guides the study, research questions, evidence collected and used, and, ultimately, the determination of the plausibility of TPA claims.

Validity of the ABV Framework

A method for analyzing validity has its own validity as a research method. In this study, the type of and data collected allowed for method validation through both triangulation and complementarity. Triangulation (Denzin & Lincoln, 2005; Hammersley & Atkinson, 1995; Campbell & Fisk, 1959) involves using results from different methods to find areas of convergence and corroboration (Greene, Caracelli, & Graham, 1989). The rationale is that there is an increase in validity when bias (method, researcher, context, or theory) is countered by confirmation of data from multiple, diverse evidence sources. Complementarity "seeks elaboration, enhancement, illustration, clarification of the results from one method with the results from another method" (Greene, Caracelli, & Graham, 1989, p. 259). Complementarity increases meaningfulness through interpretation by building on the strengths of methods employed while counteracting method bias (Greene, Caracelli, & Graham, 1989). The interpretive argument in ABV requires a range of diverse evidence for each inference in the argument. Multiple methods and analysis are meant to assist the researcher by offering a "preponderance of evidence" to interpret for each inference in the IA and also serves to counteract method and researcher bias increasing the validity of the method.

Framework of the Study: The Methodology of ABV

An IA is developed from "observed performances to conclusions and decisions" (pp. 141-2) and "is never proven; it is always subject to change" (p. 144). Kane (2002) writes that the first step of an "effective validation requires a clear statement of the proposed interpretation" (p. 32). This interpretation involves four core inferences: first, there is an "evaluation" of the observed performance and the assignment of a score to that performance; then, a "generalization" or conclusion is drawn from the performance about how that performance relates to all possible performances in similar circumstances; the third step involves "extrapolating" that conclusion or viewing it in relationship to the domain or construct in which the test was designed to measure; and, finally, a "decision" is made about the meaning of that conclusion (p. 31). In the case of the present study, these inferences are:

1. EVALUATION/SCORING: A candidate's TPA performance is given a score for each rubric
2. GENERALIZATION: The score is used to form a conclusion about overall performance
    a. Conclusion1: The candidate passed the TPA
    b. Conclusion2: The candidate failed the TPA
3. EXTRAPOLATION: The broader domain of teacher readiness is derived from the conclusion
    a. Extrapolation1: The candidate is ready to teach
    b. Extrapolation2: The candidate is not ready to teach
4. DECISION-MAKING: the meaning of the inference concerning teacher readiness is determined by stakeholders
    a. WA State: Given the candidate's readiness to teach, a license is granted or denied
    b. Accreditation Bodies (state and national): Given the total number of candidates ready to teach, a program is preparing candidates well or poorly
    c. Federal Government: If only candidates ready to teach are certified, then WA teachers are ready to be effective
    d. All: Teachers hired are ready for professional practice and effective educators.
        i. if the TPA measures teacher readiness to teach, and
        ii. if only those teachers who pass the TPA are given credentials, and

iii. if only credentialed teachers are hired,

iv. then teachers hired are ready and effective.

Thinking through these inferences also includes determining which assumptions, sometimes implicit, are embedded within each step and determining "the weakest part of the interpretive argument because the overall argument is only as strong as its weakest link" (Kane, 1999, p. 15). Some assumptions can be taken as "given" and do not have to be defended. Other assumptions are essential components of an interpretive argument and cannot be omitted. In his article, *Validating Measures of Performance*, Kane (1999) applies his ABV model to performance assessments and concludes that most validity studies should focus on evaluation, generalizability and extrapolation because strengths and weaknesses in these areas are typically "make or break" issues for assessment systems. Generalizability is likely the weakest link because performance assessments target specific skills in very specific settings and are harder to apply beyond the testing criteria and standards (1999, p. 15; 2002, p. 40). However, because the TPA is a performance assessment that is also high-stakes, consequences involve policy decisions, inferences, and assumptions (2002, p. 33) that may or may not be justified by the use of the test scores (p. 34), which may lead to challenges to the decision-making inferences. Finally, because validity and fairness are "intertwined" (2010, p. 179) and "closely connected" (p. 181), inferences and assumptions about the fairness of the uses of the TPA will also be a focus of this study.

Interpretive Argument

The IA that framed the study appears below (see Chapter one). Bolded assumptions were considered primary when applied to the present validation argument.

Table 3.1

Interpretive Argument

Inference	Assumptions and Warrants Justifying the Inference
Construct Representation: Teaching Effectiveness / Teacher Readiness <ins>*Domain of performance:*</ins> *Effective Teaching*	**The tasks in the TPA represent the domain of performance and skills of teacher readiness.** *Taking these assumptions as given:* Effective teaching is of interest to stakeholders and candidates Performance assessment on effective teaching has positive influence on curriculum for teacher preparation and readiness Performance assessment is the most valid and reliable way to measure teaching effectiveness and readiness
Evaluation / Scoring (Target Domain) <ins>*Target Domain:*</ins> *Observed Performance is representative sample of performance domain (TPA definition)*	**Criteria used to score the performance are appropriate and have been applied as intended.** Observed performance on the TPA can be considered performance in target domain (representative sample of effective teaching and teacher readiness). Scores indicate whether candidates can perform adequately on the tasks presented to them. TPA scores reflect the *level* of teacher quality and readiness on the assessment tasks. Therefore, a low score on the TPA indicates that the candidate cannot perform the tasks included in the assessment (Kane, 1994) The performance occurred under conditions compatible with the intended score interpretation (judgment/decision/consequence) in terms of the candidates' level of skill.
Generalization *Universe of Target Domain*	**The tasks/scores adequately sample and reflect performance on all possible and relevant tasks for target domain.** Scores include a representative sample of performance from the target domain. Interpretation of scores will emphasize levels of skill in sub-domains or tasks. All candidates are provided equal opportunity to succeed. **(Procedural fairness) The test is procedurally fair. The same rules are applied to everyone in more or less the same way. All test-takers are consistently treated essentially the same way, using the same (or equivalent) procedures and rules. If modifications are necessary (or required by law), they are applied in ways that create an equitable testing procedure.** (Substantive fairness). Test design and criteria are reasonable in the context of the test procedure. No substantively new content was introduced by the test, for the test, or if so, equity in opportunity to learn that content was provided before tested. Score interpretations and decision rules are reasonable and appropriate for all test takers and sub-groups. Funding sources are appropriate to testing purpose.

Extrapolation *Target Domain AND Performance Domain*	The skills assessed are necessary (if not sufficient) for effectiveness in the performance domain. **The knowledge, skills and judgment required on the performance assessment are essential for effective teaching and teaching readiness in real-world practice. As such, an interpretation and use of the results supports decisions about candidate future teaching effectiveness. *Interpretation of scores allows for a judgment about the candidate's readiness to perform as an effective teacher.*** Anyone who performs well on the assessment should also be able to perform well in the target domain. Proficient scores on the TPA tasks likely reflect teaching readiness. Anyone who performs poorly on the assessment should also perform poorly in the target domain. Poor scores on the TPA tasks likely reflect a lack of teaching readiness.
Decision-Making	Uses of scores are clear. **Uses of scores are appropriate.** Funding sources are clear and appropriate.

Validating the interpretive argument. When the IA has been done correctly, "the validation effort can focus on empirical checks of the inferences and assumptions in the interpretive argument" (Kane, 2004, p. 144). Judgment is needed to determine which of the inferences and assumptions will require "close scrutiny," "further support" or "trade offs" (p. 144). Thus, the second phase is where "proposed interpretation and uses are critically examined" (p. 137). Kane explains that "The argument based approach to validation adopts the interpretation as the framework for collecting and presenting validity evidence and explicitly associates validity with the plausibility of the various assumptions and inferences involved in the interpretation" (1990, p. 2). Using the interpretive argument above, a validation argument (VA) was crafted creating questions for each inference assumption and what evidence would support or threaten validity. The VA appears below.

Using ABV as a Framework for the TPA: Research Questions

The core question that framed this project is whether the TPA is a strongly credible measure for determining teaching readiness. In order to answer this question, an IA was developed and then a VA that lists the sub-research questions and types of evidence needed in order to answer this question. The validity arguments become the research questions. Emphasis on generalizability is reflected in the number of research questions and evidences required for these inferences**.**

Validity Argument (research questions):

Inference 1: Construct Representation

1. Do the four TPA tasks represent the six categories relevant to the intended construct (teacher readiness)?

Inference 2: Scoring/Evaluation

2. Are the scoring procedures sound and reliable?
3. Are the rubric score levels achieved by the candidate actually representative of what that candidate performed on the TPA? (Does candidate performance correlate to the assigned test score)?

Inference 3: Generalization

4. Are the score levels achieved on the rubrics a true representation of a candidate's performance? In other words, are the scores a candidate earned consistent and generalizable with other samples of that candidate's teaching performance?
5. Does poor performance on the TPA imply a lack of adequate mastery of the construct?
6. How generalizable are the criteria, rubrics, procedures, and scores derived from the TPA across different candidates and handbooks?
7. Does TPA proficiency depend upon factors beyond the candidate's control? How generalizable are the criteria, rubrics, procedures, and scores derived from the TPA across testing sites, placements and placement length and programs?

Inference 4: Extrapolation

8. Do TPA test scores provide reliable indicators of a readiness to teach? Are the TPA scores, as a whole, a true measurement of teaching ability?

Inference 5: Decision Making

9. Is guidance in place so that all stakeholders know what scores mean and how the outcomes will be used?

These questions were used to determine the evidence required to address the VA for the TPA and guided the data collection steps. If the uses are consistent with the inferences, assumptions, and questions above, the TPA will be plausible. Evaluating this claim involves evaluating these five inferences using the evidence for and challenges to validity. Applying the Cambridge Approach Framework, Table 3.2 will be used to determine the outcome of the validity argument.

Table 3.2

Framework for Determining the Plausibility of the Validity Argument

Claim	**Evaluation**
How appropriate are the intended interpretations and uses of TPA test scores?	
Interpretation: Scores provide a measure of relevant teaching readiness.	

Research Timeline

After the initial pre-study, data collection occurred over the course of one semester, or four months, during the field test conducted in spring 2012. Surveys were distributed electronically via candidate, supervisor, mentor and faculty email. All participants had access to university or district internet and computers. Case study subjects were interviewed four times, once each phase, from February through late May 2012. Candidate group interviews occurred in April, near the end of the term, after the submission of the TPA and just before the conclusion of student teaching (ST) after which candidates leave the university setting and become difficult to contact. Supervisors and mentors were interviewed prior to the TPA experience, after the submission of the TPA, and at the conclusion of ST. University faculty members were surveyed and interviewed at the end of the term in early May (see **Appendix D** for timeline).

Research Design Overview

In order to address the questions in the VA, two categories of evidence were collected. The first is evidence from the TPA. This includes the handbook, which lists the directions, expectations, and scoring criteria, and the scores that resulted from candidate submissions. A second category of evidence was collected to supplement and explain the assessment. This category will be described as "investigation" data. Investigation evidence involved the collection of multiple qualitative and quantitative sources (case studies, mentor evaluations, supervisor evaluations, and university methods) to provide evidence of plausibility of TPA scores to determine teaching readiness. Both TPA and investigation data was collected in four phases. Constructing a series of validity arguments, including assumptions and inferences, determined the types of data to be used. Both categories of data (TPA and investigation data) were analyzed and synthesized using quantitative (correlational and generalizability studies) and qualitative (coding) methods. Table 3.3 contains a summary of the research questions, participants, instruments/data, and procedures over the four phases.

Table 3.3

Study Design

Phase 1: Pre-TPA			
	Instruments	**Participants**	**Procedures**
1	Case Study Interviews	n=6 • 3 Undergraduate • 3 Graduate ○ 2 Primary ○ 2 Intermediate / Middle School ○ 2 High School	• Semi-structured individual interviews • Audio recording and interviewer notes • 2 interviewers (3 each)
2	Pre-Experience Survey	n= 47 (90%) 25/26 (96%) Undergraduate 22/26 (85%) Graduate	• Electronic survey distributed via email
3	West E[17] scores for correlation	n=52 (100%)	• Scores collected from OSPI[18] website and recorded in spreadsheet
4	Survey for supervisor and mentor	Supervisor n=16 (100%) Mentor n=59 (79%)	• Electronic survey distributed via email

[17] The West E is an electronic, multiple-choice endorsement area certification exam in WA. At Sterner, candidates must pass the West E prior to ST.

[18] The Office of the Superintendent for Public Instruction (OSPI) is the agency that oversees K-12 public education in WA, including teacher certification (Office of the Superintendent of Public Instruction, 2013). For more information, see http://www.k12.wa.us/.

Phase 2: Taught Segment / Writing Up TPA			
	Instrument	**Participants**	**Procedures**
5	Case Study Interviews	n=6 • 3 Undergraduate • 3 Graduate ○ 2 Primary ○ 2 Intermediate / Middle School ○ 2 High School	• Semi-structured individual interviews (debriefing after each lesson taught and after draft submissions) • Formal and informal observations of candidate teaching TPA learning segment • Audio recording and interviewer notes • 2 interviewers (3 each)

6	TPA Experience Survey	n=43 (83%)
		n=25 (96%) Undergraduate
		n=18 (69%) Graduate
			• Electronic survey distributed via email

7	Survey of TPA Draft Task 1 / Drafts of TPA	n=23 (89%) Undergraduate	• Electronic survey distributed via email
8	Observe candidates during the writing up phase	n=26 (100%) Undergraduate	• Audio recording and observation notes
9	TPA Writing Surveys	Undergraduate: Survey 1 n= 20 (77%) Survey 2 n= 21 (81%) Survey 3 n= 20 (77%)	• Electronic survey distributed via email
Phase 3: Post-TPA Submission			
	Instrument	**Participants**	**Procedures**
10	Case Study Interviews	n=6 • 3 Undergraduate • 3 Graduate ○ 2 Primary ○ 2 Intermediate / Middle School ○ 2 High School	• Semi-structured individual interviews • Audio recording and interviewer notes • 2 interviewers (3 each)

11	Post-Submission Candidate Survey	n=46 (89%)
		n=26 (100%) Undergraduate
		n=23 (89%) Graduate
			• Electronic survey distributed via email

12	Teacher Candidate Progress Report (candidate, supervisor and mentor)	n=135 Reports 78 Undergraduate Reports • n= 20 Undergraduate Candidate • n= 26 Undergraduate Supervisor • n= 32 Mentor 57 Graduate Reports • n=0 Graduate Candidates • n=26 Graduate Supervisor n=31 Graduate Mentor	• Participant has choice to complete by writing on hardcopy OR electronically

13	Mentor and Supervisor Survey	n= 14 (88%) Supervisors
		n= 24 (32%) Mentors
			• Supervisor coordinates survey collection and submits to investigator

Phase 4: Post Student Teaching Term			
	Instruments	**Participants**	**Procedures**
14	Case Study Interviews	n=6 • 3 Undergraduate • 3 Graduate ○ 2 Primary ○ 2 Intermediate / Middle School ○ 2 High School	• Semi-structured individual interviews • Audio recording and interviewer notes • 2 interviewers (3 each)

15	Post-ST Candidate Survey	n=39 (75%)
		n=25 (96%) Undergraduate
		n=14 (54%) Graduate
			• Electronic survey distributed via email

16	Teacher Candidate Final Progress Report (candidate, supervisor and mentor)	n=91 reports
		n=34 Undergraduate Mentor
		n=26 Undergraduate Supervisor
		n=31 Graduate Supervisor
			• Participant has choice to complete by writing on hardcopy OR electronically

17	Mentor Final Survey	n=28 (37%)	• Electronic survey distributed via email
18	Supervisor Observation Tool	n=130 26 Undergraduate with 5 reports each	• Created electronically and completed on hardcopy triplicate throughout the semester. One copy given to researcher at the end of the term , one given to mentor and one given to candidate at the conclusion of the observation date.
19	Group Interview #1 (TPA)	n= 16 8 Undergraduates 8 Graduates	• Semi-structured interview with up to ½ of the participants, groups of 4-5. • Interviews conducted by primary, secondary investigators/research assistant. • Interviews will be audio/video recorded, interviewer will take notes.
20	Student Teaching Debrief Interview	n=30[19] Undergraduate only	• Structured Interview with 4-10 participants each. Questions are asked verbally and recorded on the hardcopy interview form. • Supervisor conducts the interview and records the data.
21	Student Teaching Field Placement Evaluation	n=26 Undergraduate only (Graduate program did not participate in this aspect of the study)	• Survey completed in class on paper and submitted to investigator

[19] Because these interviews were conducted by supervisors and in anonymous groups, the researcher was unable to know individual responses to remove comments from the four non-participating candidates.

22	Candidate Professional Growth Plan (PGP)	n=26 Undergraduate only	• Survey created / completed electronically and submitted with PGP • Investigator reviews and notes areas of correlation
23	TPA Scores and Feedback Reports	n=52 (100%)	• Raw scores will be electronically submitted to the IHE and input in spreadsheet
24	Faculty Survey	n=22/40 (55%)	• Electronic survey distributed via email
25	Faculty Group Interview	n=11 (28%)	• Structured Interview with 4-10 participants each. Questions are asked verbally and recorded.
26	Supervisor Group Interview	n=16 (100%)	• Structured interview with 4-10 participants. Questions are asked verbally and recorded.
27	Document Analysis	n/a	• Review PESB standards • Review TPA handbooks • Review TPA samples

Specific to the inferences, assessments, and uses of the TPA, each data set connects to the validity argument, described in Table 3.4. An asterisk (*) identifies use of TPA data. Generalization and scoring are the weakest links in the chain of evidence and the focus on the validation study.

Table 3.4

TPA Interpretive Argument and Data Sources

Inference	**Primary Data Sources**	
	Qualitative Data	Quantitative Data
Construct Representation	Expert Consensus TPA Handbook* Document Analysis*	
Evaluation / Scoring	Case Study Interviews Surveys	TPA scores* Descriptive analysis* Factor analysis* West E scores Surveys
Generalization	Case Study Interviews Surveys	TPA scores* MTMM* Generalizability study* Surveys
Extrapolation	Surveys Interviews	Mentor and Supervisor Evaluations Surveys
Decision-Making	Document Analysis* Interviews	

Study Sample

The credentialing program at the suburban, private, religious “Sterner University” was selected for a number of reasons. First, this university is consistently recognized as an innovator within the state, participating in most voluntary state-wide pilots in teacher preparation. Second, administrative leadership in the School of Education (SOE) at this university remained steady for more than twenty years, eliminating change factors from leadership fluctuations. Third, this university’s SOE was accredited by both the PESB and National Council for the Accreditation of Teacher Education (NCATE) the spring prior to the pilot, receiving the highest accolades from both organizations and a “recommendation without revision” accreditation rating. Lastly, Sterner has several distinct licensure programs which follow different models of preparation, especially during ST, from a year-long placement to a six week placement. Finally, Sterner places candidates in multiple school districts across WA. These variations mimic the diversity throughout the state in the licensure programs available (all of which must prepare candidates to meet the same standards for licensure, one of which is the TPA).

Ethical Considerations

There were some conflicts between the investigator’s role in conducting research and her need as an instructor not to violate the principle to “do no harm” (Hammersley & Atkinson, 1995). The ethical issues involved in asking candidates to simultaneously undergo a high-stakes internship, assessment, and study were identified and addressed, as much as possible, through anonymity, confidentiality, and informed consent. Prior to the study, participants were informed of their options for involvement and opportunity to opt out at any time. The anonymity of participants during the term was respected by asking a third party to introduce the project, secure consent, and keep records until research concluded (and candidates graduated). In addition, case study participant anonymity was protected through multiple investigators for the case study interviews. Some stress is acknowledged and unavoidable. To reduce stress, interviewers used flexibility in scheduling of interviews, locations of interviews, and accommodating surveys deadlines. Participants were aware of the researchers' role, the purpose of the study, and the meaningfulness of their participation. The investigator was in dialog with mentors, supervisors and candidates to minimize additional stress for participants due to research requirements. Specific candidate protections were required by Sterner in the IRB process and are described below.

Teacher Candidates

During the 2011 academic semester, three candidates volunteered to be a part of the researcher's pilot in a pre-phase of this study. Each candidate was asked to complete the TPA, two interviews with the researcher, and to provide feedback to the rising cohort of candidates in a video-taped forum. In the middle of the pilot, one of the three candidates opted not to participate due to the strain of the TPA on the ST experience. In fall 2011, all UG candidates completed TPA, were surveyed, and interviewed. Faculty at Sterner were trained to score TPA. A sub-group of scorers were also interviewed. While helpful to the researcher, ultimately, these TPA samples and the feedback from participants proved less beneficial because the assessment underwent significant revision between September 2011 and January 2012.

During the 2012 spring semester, all candidates in WA were required to pilot the TPA in core endorsement areas. At Sterner there were thirty-one traditional undergraduate candidates, thirty graduate candidates, and thirty non-traditional undergraduate candidates placed in ST internships. The director of the non-traditional program chose not to have that group participate. Due to IRB requirements at Sterner, all subjects were recruited by a third-party. Candidates were invited to volunteer in a short presentation, letter introducing the project, and consent form. Those interested in the case study marked a check-box on the consent form. Following university guidelines for internal research, the third-party research assistant collated and logged consents and kept the master-list. She also selected from the volunteers three graduate and three undergraduate candidates for the case study based on placement levels (elementary, middle, high) within each group. Until the conclusion of the term, the investigator was unaware of which candidates gave permission, beyond the three graduate candidates interviewed for the case study.

Candidates participating in ST were determined based on programmatic criteria independent of the study, which included the accomplishment of programmatic benchmarks, successful completion of coursework, recommendations from faculty advisors and field teachers, and an interview with the Director of ST[20] (undergraduate) or Elementary/Secondary Coordinator (graduate). Four undergraduate candidates opted not to participate. In addition, one undergraduate candidate took a medical leave in the middle of the semester and was removed from the study. Four graduate candidates opted not to participate. A total of fifty-two candidates (85%) participated.[21]

Sterner SOE, along with all teacher-preparation programs in WA, piloted the TPA in the following subject-specific areas: Elementary Math, Elementary Reading, Secondary Social Studies, Secondary Math, Secondary Science, Secondary English, Theatre Arts, Secondary World Languages, and Music Performance. It was not possible to randomly assign subjects into piloting and control groups because all candidates were required to pilot the instrument.

Further, because of the way that Sterner programs were designed and organized, candidates' program coursework and ST requirements varied by cohort groups, some of whom remained together as a cohort. It was not possible, therefore, to control for coursework, the way in which candidates were introduced to the tool, or how much exposure each candidate had to the TPA prior to the study. Although undergraduate and graduate candidates at Sterner take similar coursework, the courses were taught by different instructors and may have been offered in different academic semesters. Sequencing of courses was also sometimes different. In addition, some programs and cohorts had longer ST placements.[22]

[20] The Director of Student Teaching is also the study investigator.

[21] Note: All assessment participants released rights to privacy for their test scores as a TPA submission requirement. When TPA test scores are reported, the N=58. For all other aspects of the study N=52.

[22] Graduate candidates are placed for 24 weeks, elementary undergraduates are placed for 32 weeks and secondary undergraduate are placed for 12 weeks.

Candidates in this study are not as diverse as the population at large but represent demographics at Sterner. Diversity in recruiting for teacher preparation is a current concern among university educators (AACTE, 2011; Morrell, 2010; NCATE, 2010). Participants were 18-49 years of age and the majority female. All participants were from the US. All were fluent in English. Most candidates declared their racial background as Caucasian with a minority few from Asian, Latino or other racial backgrounds.

Table 3.5

Programmatic Differences in Age

n=46/52 (89%)	**Undergraduate**	**Graduate**	**Totals**
18-20	1/26 (4%)	0%	2%
21-29	23/26 (92%)	16/26 (76%)	85%
30-39	1/26 (4%)	3/26 (14%)	9%
40-49	0%	2/26 (10%)	4%
Total	26	26	

Table 3.6

Programmatic Differences in Sex

n=47/52 (90%)	**Undergraduate**	**Graduate**	**Total**
Males	4/26 (16%)	8/26 (36%)	26%
Females	21/26 (84%)	14/26 (64%)	74%
Total	26	26	

Table 3.7

Programmatic Differences in Racial Background

n=47/52 (90%)	**Undergraduate**	**Graduate**	**Total**
Caucasian	22/26 (92%)	21/26 (96%)	94%
Black or African-American	0%	0%	0%
American Indian or Alaskan Native	0%	0%	0%
Asian	1/26 (4%)	1/26 (4%)	4%
Native Hawaiian or other Pacific Islander	0%	0%	0%
From multiple races	1/26 (4%)	0%	2%
Hispanic/ Latino	1/26 (4%)	0%	2%
Total	26	26	

In order to determine motivation and background, candidates were surveyed prior to the start of ST. The survey asked candidates to indicate areas of strength and readiness for teaching and areas of weakness, or concern. Candidates were also asked to identify their motivation for completing the TPA. In Table 3.8 data collected from candidates' self-reported initial motivations are listed. Note that the most important reason, also the highest scoring reason overall, was that it was a requirement for licensure. This is particularly significant because, during the state-wide field test, completion of the TPA was not a requirement of licensure. The second most important motivation identified by candidates was that the TPA was required coursework for their program. In both programs, candidates received a pass/fail based on TPA submission, not TPA performance. The least important reason selected was to be motivated by the TPA as a professional development opportunity. This is an interesting result since twenty-four (55%) of the candidates also listed that determining their teaching strengths and weaknesses was one of their top three motivations. Similarly, as discussed in Chapter two, one of the primary benefits of using performance assessment, versus other types of assessment such as item-based multiple choice, is that it offers formative growth for the test-taker (Chung, 2005).

Table 3.8

Candidates Initial Motivation for Completing the TPA.

I will complete the TPA because:			**Response Rate: 44/52 (85%)**		
	(1) most important reason	**(2)**	**(3)**	**(4)**	**(5) least important reason**
It is a professional development opportunity	4/44 (9%)	2/44 (5%)	13/44 (30%)	11/44 (25%)	14/44 (32%)
It is required for certification	26/44 (59%)	11/44 (25%)	1/44 (2%)	5/44 (11%)	1/44 (2%)
I want to challenge myself as a professional	2/44 (5%)	5/44 (11%)	5/44 (11%)	18/44 (41%)	14/44 (32%)
I will use the TPA to determine my strengths and areas for improvement	5/44 (11%)	4/44 (9%)	24/44 (55%)	5/44 (11%)	6/44 (14%)
It is required coursework in the SOE	7/44 (16%)	22/44 (50%)	1/44 (2%)	5/44 (11%)	9/44 (21%)

Mentor Teachers

Mentors were partnered with candidates based on voluntary participation and permission from their school administrator and district office. Some mentors selected candidates based on prior experience with the candidate, while others were randomly assigned based on district and building requirements and needs. All mentors were licensed in their grade and discipline areas. All mentors had at least three years of teaching experience. Mentors were invited to participate in the study through a letter describing the study and a conversation with the supervisor who served as the liaison between the university and the placement site. Seventy-five mentors were assigned to candidates in the program. Mentor participation varied over the study period (33%-82%).

Table 3.9

Mentor Differences in Sex

n=58/75 (77%)	**Responses**
Male	13/58 (22%)
Female	45/58 (78%)

Table 3.10

Mentor Differences in Level of and Recent Experience

n=58/75 (77%)	**Responses**
First year	14/58 (24%)
2-5 years	24/58 (41%)
5-7 years	5/58 (9%)
More than 7 years	15/58 (26%)
Recent Experiences	
Hosted last year	34/58 (59%)
Hosted two to three years ago	14/58 (24%)
Hosted three to five years ago	4/58 (7%)
Hosted more than five years ago	6/58 (10%)

Because the TPA expectations and format has been compared to the National Board Certification requirements for practicing teachers, mentors were asked to report whether they had earned NB certification. Of those that responded, few (12%) had achieved NB.

Table 3.11

Mentors with National Board Certification

n=58/75 (77%)	Responses
Yes	7/58 (12%)
No	51/58 (88%)
Currently in the process	0%

University Supervisors

Supervisors are primarily retired teachers and administrators and are contracted employees of Sterner. Participation in this study was a contractual requirement. Sixteen supervisors participated in the study and participation rates were high (88%-100%) though participation varied by task and over the term.

Table 3.12

Supervisor Differences in Sex

n=16/16 (100%)	Responses
Male	7/16 (44%)
Female	9/16 (56%)

Table 3.13

Supervisor Differences in Level of Experience

n=16/16 (100%)	Responses
First year	3/16 (19%)
2-5 years	8/16 (50%)
5-10 years	4/16 (25%)
More than 10 years	1/16 (6%)

Table 3.14

Supervisors with National Board Certification

n=16/16 (100%)	Responses
Yes	1/16 (6%)
No	15/16 (94%)
Currently in the process	0%

University Faculty

Faculty in the SOE at Sterner were asked to voluntarily participate in the study through a letter describing the study and a conversation with the researcher. Most core faculty involved in preparing candidates for student teaching (methodology courses, content-pedagogy courses) participated in the study.

Study Design

Given the complex nature of ABV inquiry, a pre-research pilot was used to refine the surveys and other study instruments and to provide interview practice for the investigator in order to help reduce research bias and for the purpose of best addressing the research questions.

The data collection in the four phases of the study was aimed at validating the TPA as a measure of readiness. The researcher collected four candidate surveys: pre-TPA experience, during-TPA experience, post-TPA submission experience, and post-ST experience. In addition, the researcher and a second interviewer conducted case-studies with six candidates during the pilot. Case study candidates were interviewed following the same phase schedule for data collection–pre, during, post-submission and post-ST. Mentors, supervisors and faculty were also surveyed and interviewed. Finally, West E and TPA raw data were scored by an external scoring agent, Pearson.

1. Pre-TPA: The primary purpose of data collection prior to experience with the TPA was to establish a base-line for both attitudes about ST and the TPA and candidates' self-reports regarding their current abilities and weaknesses.

2. During-TPA: Data collected during the taught segment and TPA writing up process reveals candidate thinking about the instrument during completion. In particular, candidates are asked to provide information about areas of struggle and mastery.
3. Post-Submission: This data provides the initial reaction to the instrument after submission. Questions emphasized the formative and summative learning experience provided.
4. Post-ST: Data collected at the culmination of the term asked participants to review the TPA with respect to the entire experience of ST.

A quantitative component included the collection of data using ST evaluations from mentor(s) and supervisors which also included observation ratings on the entire sample of candidates, score reports for the fifteen sub-traits of the TPA, four surveys administered to candidates (one in each phase), three surveys administered to mentors and supervisors (beginning, post-TPA, post-ST), and university summative instruments. Where possible, scores on the fifteen sub-traits of the TPA were compared with the corresponding variables in ST evaluations and observation ratings to see whether there were any systematic differences between the two rankings of candidate abilities. In addition, the responses of candidates were analyzed to examine the factors that impacted performance on the TPA. The sample sizes for each of the quantitative instruments are found in Table 3.3.

In analyzing responses from the surveys and case study interviews, differences between the undergraduate and graduate programs were found. These differences may be a result of program components or they could be attributable to age and experience, length of placement, and comfort with placement setting or other factors. While all candidates had similar coursework, there were some significant variances in programs, most importantly the length of time spent in the placement. Most elementary candidates in the undergraduate cohort and candidates in the graduate program had two full semesters (32 weeks) at their placement site working with the same mentor and classroom culture. This placement started out part-time and moved to full-time in the second semester. Secondary candidates in the undergraduate cohort had only twelve weeks of ST in one semester (starting with one day a week and developing to full-time for ten weeks). Most importantly, there were differences in the districts and school contexts for ST placements. Some candidates were at high-need, high-risk schools and others were placed at schools from backgrounds of socio-economic advantage. Variables in cohort experiences will be discussed below and in more detail in the analysis chapter (Chapter four) and were taken into account when selecting participants for the case studies.

The qualitative component of the study included case study interviews of six candidates. Qualitative methods are useful for identifying patterns and casual linkages that are not apparent in the quantitative data. Three case study candidates participated from the graduate program and three from the undergraduate program. Cross-case studies from the same institution allowed for an examination of both the meaning candidates made of the TPA (motivation, view of consequences of scores) and the placement contexts and preparation factors that better prepared or prevented candidate learning during the teaching event. The primary purpose of the case studies was to closely examine the experience candidates underwent as they completed the portfolio assessment and whether they believed the assessment accurately measured their readiness and preparation to teach (for licensure). The second purpose of the case studies was to identify the contextual factors that impacted the experience, and ultimately, the outcomes. If placement and context of ST proved to be a significant factor in TPA success, as measured by scores on sub-traits, but those factors were perceived beyond the control of the candidate, it would raise questions about the consequential validity of the operationalized construct of the TPA. Finally, case studies offered the IHE opportunities to examine the types and levels of support needed for future candidates.

Because the TPA and the length of ST were theorized to have an influence on teacher learning (Chung, 2005), case study subjects were selected based on literal replication logic (cases are expected to produce similar results) as well as theoretical replication logic (cases are expected to produce contrasting results but for predictable reasons) (Yin, 1994). Based on program membership, placement levels, and TPA subject, these six candidates represent multiple levels of placement diversity. For each program, one elementary, one secondary and one middle school candidate was selected in order to evaluate long and short-term placement length and program differences on TPA experiences and performance. Table 3.15 lists the case study subject, their program, and placement.

Table 3.15

Case Study Subjects and their Placements

	Elementary	**Middle School**	**Secondary**
Undergraduate Year Placement	Jason (4th grade)		
Undergraduate Semester Placement		Jane (7th grade Math)	Jill (9-12 English / Theatre Arts)
Graduate Year Placement	Jackie (1st grade)	Jamie (5th grade)	Jennifer (10th grade Social Studies / Science)

Differences in programs and placements will be further examined in Chapter four. Case study participants were intentionally selected based on these criteria (program and placement). In addition, placements spanned three different school districts and also varied by the number of mentors assigned to each candidate. Differences between districts and schools are discussed below.

School District #1

District #1 is the third largest district in WA with over 31,000 students. This urban district includes seven high schools, six middle schools, and thirty-four elementary schools. The levels of poverty are higher in this district (56.7%) than the state average (45.5%) and significantly higher than the other two districts represented in the study. Student test scores on state exams to measure achievement are collected annually starting at the third grade (8-9 year olds). Test scores rank this district slightly below the state average with the exception of eighth grade and high school mathematics and science.

Table 3.16

School District #1 End of Course State Test Scores

Grade Level	Reading	Math	Writing	Science
3rd Grade	66.6%	65.3%		
4th Grade	69.7%	63.8%	57.6%	
5th Grade	68.0%	66.3%		69.6%
6th Grade	69.3%	70.6%		
7th Grade	68.3%	62.3%	67.8%	
8th Grade	62.8%	53.9%		69.2%
10th Grade	77.7%	See EOC[23] below	81.8%	See EOC below

Grade Level	EOC Math Year 1	EOC Math Year 2
All Grades	75.2%	82.8%

Grade Level	EOC Biology
All Grades	66.3%

Jason and Jackie were placed in grade K-6 (elementary) schools in this school district. Jackie's placement was in one of the poorest schools in WA with a free and reduced lunch rate (measurement of poverty) at 83.1% of the population. Jason's school ranks slightly less poor with a free and reduced lunch rate of 78.2% of the school population. Jackie's school reports 68% of the population are Caucasian, 8.5% are Hispanic and 11.5% are bi-racial. Jason's school reports that 71.3% of the school population are Caucasian, 9.1% are Hispanic, and 11.2% are bi-racial.

Jason was assigned one mentor and one supervisor to form a traditional ST triad. Jackie, however, was partnered with two mentors who participated in a work-share to split the week. Jackie worked with one mentor Monday through Wednesday morning and the other mentor Wednesday afternoon through Friday. The student population and classroom did not change.

[23] EOC indicates end of course.

Student test scores for these two schools show some significant differences in the rates of achievement, as measured by the end of course exams, in comparison to state-wide averages. This is especially true of Jackie's placement.[24]

Table 3.17

Jason's School End of Course State Test Scores

Grade Level	Reading	Math	Writing	Science
3rd Grade	66.7%	62.5%		
4th Grade	77.0%	64.9%	64.9%	
5th Grade	56.5%	58.8%		68.2%
6th Grade	66.2%	67.6%		

Table 3.18

Jackie's School End of Course State Test Scores

Grade Level	Reading	Math	Writing	Science
3rd Grade	59.5%	58.7%		
4th Grade	62.5%	50.7%	39.4%	
5th Grade	60.0%	53.3%		43.3%
6th Grade	57.4%	61.8%		

Jill was also placed in this district at one of its larger high schools. Both Jason's and Jackie's school sites are feeder schools for Jill's high school placement site. Jill's placement was also in a school with higher than average poverty (54.1%), as measured by free and reduced lunch rates. Studies have suggested that these rates for middle and high school are difficult to obtain because many students and their families no longer submit the required application (http://www.fns.usda.gov/). Jill's school reports 77.3% of the population are Caucasian, 6.2% are Hispanic and 8.2% are bi-racial.

[24] Scores are derived from students Jason worked with in his 4th grade placement. Because Jackie worked with 1st grade students, and 1st grade is not a level for which the state tests achievement, scores for her students are not available.

Jill's diverse endorsement areas meant that she was assigned one mentor for her English classes and another for her Theatre Arts classes. She taught two sections of each subject for a total of four preparation periods. Student test scores show some differences in the rates of achievement, as measured by the end of course exams, in comparison to the state-wide averages.

Table 3.19

Jill's School End of Course State Test Scores

Grade Level	Reading	Math	Writing	Science
10th Grade	79.0%	See EOC below	82.9%	See EOC below

Grade Level	EOC Math Year 1	EOC Math Year 2
All Grades	68.7%	89.7%

Grade Level	EOC Biology
All Grades	53.3%

School District #2

School District #2 is a small, suburban district with 9,000 students and fourteen schools. The levels of poverty are significantly lower in this district (31.4%) than the state average (45.5%). Test scores rank this district well above the state average in every testing area.

Table 3.20

School District #2 End of Course State Test Scores

Grade Level	Reading	Math	Writing	Science
3rd Grade	80.2%	82.4%		
4th Grade	87.2%	78.7%	87.6%	
5th Grade	80.8%	76.5%		82.8%
6th Grade	81.1%	74.6%		
7th Grade	82.9%	66.3%	86.8%	
8th Grade	78.2%	67.2%		86.1%
10th Grade	88.9%	See EOC below	94.1%	See EOC below

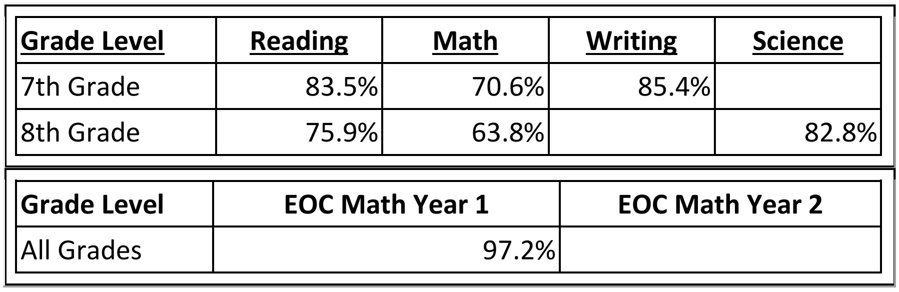
Grade Level	EOC Math Year 1	EOC Math Year 2
All Grades	78.8%	88.6%

Jane and Jennifer were both placed in this school district. Jane was placed at a middle school for seventh and eighth year students. Jennifer was at a high school for ninth through twelfth year students. Jane's placement was in one of the more affluent schools in the state with a free and reduced lunch rate of 24.7% of the population. Jennifer's school free and reduced lunch rate is 27% of the school population. Jane's school reports 82.7% of the population are Caucasian, 5.6% are Hispanic and 7.5% are bi-racial. Jennifer's school reports that 88.5% of the school population are Caucasian, 4.6% are Hispanic, and 6.1% are bi-racial.

Jane and Jennifer were both assigned two different mentors. Jane taught all but one section with the same mentor, who served as her "primary" mentor. Jane taught three sections of pre-algebra and one section of advanced pre-algebra (an honors course) for a total of four classes with one and one-half preparations. Jennifer taught two sections of social studies with one mentor and two sections of biology with another mentor for a total of four classes and three preparations. Student test scores for these two schools show some significant differences in the rates of achievement, as measured by the end of course exams, in comparison to the state-wide averages.

Table 3.21

Jane's School End of Course State Test Scores



Grade Level	**Reading**	**Math**	**Writing**	**Science**
7th Grade	83.5%	70.6%	85.4%	
8th Grade	75.9%	63.8%		82.8%

Grade Level	**EOC Math Year 1**	**EOC Math Year 2**
All Grades	97.2%	

Table 3.22

Jennifer's School End of Course State Test Scores

Grade Level	Reading	Math	Writing	Science
10th Grade	88.7%	See EOC below	93.2%	See EOC below

Grade Level	EOC Math Year 1	EOC Math Year 2
All Grades	77.0%	89.5%

Grade Level	EOC Biology
All Grades	78.3%

School District #3

School District #3 is a medium sized, suburban school district with twenty-two schools and 12,600 students. Like School District #2, the levels of poverty are lower in this district (39.6%) than the state average (45.5%). Though the differences are not as significant as in District #2, the test scores for District #3 rank this district well above the state average in every testing area.

Table 3.23

School District #3 End of Course State Test Scores

Grade Level	Reading	Math	Writing	Science
3rd Grade	74.2%	72.7%		
4th Grade	75.1%	71.5%	68.2%	
5th Grade	73.2%	68.6%		87.2%
6th Grade	75.7%	66.2%		
7th Grade	79.4%	72.8%	74.1%	
8th Grade	72.5%	67.6%		76.3%
10th Grade	87.6%	See EOC below	90.5%	See EOC below

Grade Level	EOC Math Year 1	EOC Math Year 2
All Grades	80.0%	85.1%

Grade Level	EOC Biology
All Grades	78.0%

Jamie was placed in this school district at an elementary school where 40% of the population qualified for a free and reduced lunch. Jamie's school reports 84.5% of the population are Caucasian, 8.9% are Hispanic and 4.2% are bi-racial. Jamie was partnered with two mentors who were co-teachers in a fifth grade classroom. These teachers divide responsibility for each of the curriculum subjects but share the planning, assessment, and teaching for two classes and groups of students. Student test scores for these two schools show some significant differences in the rates of achievement, as measured by the state end of course exams, in comparison to the state-wide averages.

Table 3.24

Jamie's School End of Course State Test Scores

Grade Level	Reading	Math	Writing	Science
3rd Grade	79.8%	73.4%		
4th Grade	77.9%	78.8%	75.0%	
5th Grade	84.9%	75.5%		96.2%

All of these candidates are students at the same university completing ST during the same term, so are comparable to each other. In addition, because three of the candidates came from each program and two candidates were assigned each placement level, and candidates spanned three different school districts, the experiences and testimonies from these candidates are believed to be comparable to a more general population.

Quantitative Data Collection

The quantitative evidence included candidate surveys administered at four discrete phases throughout ST, evaluations collected at the end of the term from both the mentor(s) and supervisor assigned to each candidate, an observation rating tool created for the purpose of this study and used by supervisors throughout the term, a survey conducted at the end of the term for university faculty, the professional growth plan developed by candidates at the end of ST, and from the fifteen sub-constructs of the TPA. In addition, undergraduate candidates were asked to complete additional surveys during the writing process and at the conclusion of their planning task along with a submitted draft of Task 1. Also during this phase, undergraduate candidates were excused from ST for three half days during one week and brought to campus to write their TPA. While in this "writing up" phase they were observed by the researcher and also surveyed at the conclusion of each of the three half days.

Data from each of the surveys was analyzed separately to determine both the process and impact of the TPA on the ST term and then as a whole to find patterns to address the universe of generalizability as well as the perceptions of purpose and fairness of the testing system. Data from the four surveys was merged (based on student IDs) and matched by candidate, which allowed for a comparison to TPA scores and ST evaluations. However, this reduced the sample size considerably, since not all student teachers completed all four surveys. The table below indicates the sample size for each survey. Uneven response rates and low response rates decreases the reliability and validity of the findings drawn from the correlational data.

Table 3.25

Teacher Candidate Samples and Response Rates

Respondents/ Sample:	Pre	During	Submission	Post ST	Task 1	Writing Day 1	Writing Day 2	Writing Day 3
Under-graduate	25/26 (96%)	25/26 (96%)	26/26 (100%)	25/26 (96%)	23/26 (89%)	20/26 (77%)	21/26 (81%)	20/26 (77%)
Graduate	22/26 (85%)	18/26 (69%)	23/26 (89%)	14/26 (54%)	0	0	0	0
Total	47/52 (90%)	43/52 (83%)	46/52 (89%)	39/52 (75%)	23/52 (44%)	20/52 (39%)	21/52 (40%)	20/52 (39%)

In addition to surveys for the candidates, supervisors and mentors were surveyed to solicit expert feedback on the construct representation, evaluation, generalizability, fairness and the impact of this new assessment on ST. Response rates for supervisors were very high but mentor response rates dropped significantly between the first and last phase. Uneven response rates and low response rates decreases the reliability and validity of the findings drawn from the correlational data (see **Appendix E** for surveys).

Table 3.26

Mentor and Supervisor Sample and Response Rates

Respondents/Sample	Phase 1	Phase 3	Phase 4
Supervisors	16/16 (100%)	14/16 (88%)	n/a
Mentors	59/75 (79%)	24/75 (32%)	28/75 (37%)
Total	75/91 (82%)	38/91 (42%)	28/91 (31%)

Student teaching evaluations. The ST evaluation instrument asked supervisors, mentors, and candidates to rate the candidate's performance using teaching standards aligned to PESB expectations and Sterner's Conceptual Framework.[25] The ST triad (supervisor, mentor(s), teaching candidate) was asked to rate the candidate's performance at both the midterm point and end of term (undergraduate candidates) or only at the end of the term (graduate candidates). For undergraduate candidates, the midterm point corresponded to the timing of the TPA submission. The instrument was shared in a three-way conference at the midterm point as a formative evaluation to assist the candidate's professional growth in ST. Supervisors received reliability training in sessions conducted by the researcher before the midterm conferences and final conferences. Supervisors submit seven to ten formal observations using a tool aligned to Sterner Conceptual Framework, PESB standards, and the TPA, to support their evaluations on this instrument. Because many of the performance expectations on this evaluation tool align to those on the TPA, correlational data on the generalizability and the reliability of the TPA scores is possible.

Teacher candidate placement evaluation. Upon term completion, undergraduate candidates are asked to complete an evaluation of their placement site to determine whether the placement met the needs of ST. This survey asked respondents to:

[25] SOE develop a conceptual framework to capture the essence of the university mission, knowledge base, and goals for candidates. This document is typically required and used for accreditation and because it documents the unique teacher traits sought by the SOE for teacher preparation, university faculty align their coursework to this document, in addition to PESB requirements.

- Rate their overall experience using a four point Likert scale
- Describe positive and challenging aspects of the placement site
- Describe the quality of the mentorship received
- Describe the relationship with the school community, including administrative support
- Describe whether and how the TPA impacted the placement

Table 3.27

Sample and Response Rates for Student Teaching Evaluation Survey

Respondent/Sample:	Student Teaching Evaluation
Undergraduate	26/26 (100%)
Total	26/52 (50%)

Supervisor observation rating tool. An observation rating tool was developed for the purpose of this study and used by supervisors throughout the term to provide formal feedback to candidates on their lesson performance. This tool was adopted by Sterner and supervisors received training on this tool. Undergraduate supervisors conducted observations throughout the term and submitted these observations. This tool allows for the computation of “gain scores” between the start and the end of the term and between the submission of the TPA and the end of ST. This allows for a reliability study to determine whether TPA scores correspond to the end of term scores for candidates.

Faculty survey. University faculty were surveyed to determine the impact of the TPA on their courses, their level of knowledge and understanding of the TPA, and the overall beliefs about the TPA, as a measure of candidate readiness. This survey was used to understand whether the components were well understood and embedded within courses preparing candidates for ST. Additionally, the extent to which the TPA required faculty development and course adaptation was of interest to the researcher to determine how the requirements the TPA differed from the performance assessment it replaced. All SOE faculty were invited to participate in the survey.

Response rates for faculty teaching methods courses were high. This is likely due to the fact that this is the population most impacted by the TPA and who share the highest responsibility for preparing candidates to successfully complete TPA tasks.

Table 3.28

Faculty Sample and Response rates

Respondents/Sample:	Participant Numbers
Core methods	12/16 (74%)
Non-Methods	10/24 (42%)
Total	**22/40 (55%)**

Teacher candidate professional growth plan. At the midterm point, candidates met with their supervisor and mentor(s) to discuss strengths and weaknesses in their teaching. This midterm conference was conducted within two weeks of TPA submission. Candidates used the feedback from this conference, their own reflective evaluation, and experiences with the TPA to create a Professional Growth Plan (PGP). The PGP lists candidate strengths aligned to Sterner's conceptual framework and outlines three goals for professional improvement for the first year of teaching. While self-reports, perceptions, and candidate choice certainly influence the validity of this instrument, reflective practice and goal setting have long been considered a valuable tool for the professional growth of teachers. Because the TPA claims to be an authentic assessment of candidate abilities, responses on this PGP were compared to TPA scores and candidate survey responses to determine the benefit of the TPA in helping candidates to better understand professional practice and to provide another set of data to verify the reliability of TPA scores.

Table 3.29

Sample and Response rates for PGP

Respondent/Sample:	PGP
Undergraduate	26 (100%)
Graduate	0
Total	**26/52 (50%)**

TPA scores. The final data source for the quantitative component of the study included the scores from the fifteen sub-constructs of the TPA for each candidate during the field test.

Table 3.30

Sample Response Rates for TPA Scores

Respondent/Sample:	TPA
Undergraduate	26/26 (100%)
Graduate	26/26 (100%)
Total	52/52 (100%)

Qualitative Data Collection

To address the inferences in the VA, qualitative data was also collected including case study interviews, drafts of candidate work on the TPA, observation of candidates during the writing of the TPA, large group candidate interviews, supervisor interviews and university faculty interviews at the conclusion of the term, and document analysis.

Teacher candidate focus group interview. To better understand the patterns and experiences of the cohort, candidates (not case study participants) were invited to a group interview in the last three weeks of ST. Volunteers for these group interviews exceeded the researcher's expectation and were selected based on diversity of placement and program. Sixteen candidates participated in the interview.

Supervisor focus group interview. Supervisors serve as the liaison between the university setting and the placement sites. When the experience is going smoothly, candidates and mentors collaborate with the supervisor, not university administration. Supervisors visit the school, conduct observations of the candidate's teaching, and observe the relationship between the mentor and the candidate. Supervisor reports of candidate ability are the basis of much of the university performance data on a candidates' practical readiness. For this reason, supervisors were solicited to share their experiences in a structured group interview at the end of the term. This interview was conducted on the last day of the term for supervisors, immediately before they submitted final ratings. All sixteen supervisors participated in the interview.

Core methods faculty focus group interview. Methods instructors in the SOE were asked to participate in a structured interview at the end of the term. Seven faculty members from the undergraduate program and the three full-time faculty from the graduate program participated.

Document analysis. Lastly, a qualitative document content analysis was applied to analyze the PESB requirements for teacher licensure using the *Teacher Benchmarks* and the *Standard V* documents. Teacher preparation programs use these documents to design their programs so as to provide opportunities to learn the knowledge, skills, and dispositions expected and assessed for WA licensure of pre-service teachers.

Case Study Data Collection.

Data collection for the case study component included the following:

1. Each of the six case study subjects participated in four structured interviews (approximately one hour each), with interviews separated by study phase (see **Appendix F** for interview protocols). During these interviews, candidates were asked about their educational backgrounds, motivations for entering the teaching profession, experiences completing the TPA, and experiences during ST. To better understand each candidate and their learning experiences, candidates were asked about the value of their program, coursework and prior experiences in and with education and teaching. In order to appreciate their contexts for first immersion into the teaching profession, candidates were asked about the district, school, and classrooms of their placements (this is also

the TPA context), any limitations placed upon their experience, and their relationship with their mentor(s), supervisor, and students. Finally, candidates were asked to specify sources of learning they considered significant during their program of study and/or ST term.

2. Each case study subject was asked to share experiences with the assessment, specific assistance or support provided to complete the assessment, and the overall impact of the assessment, and as a formative learning experience. In order to better understand their needs, advice was requested for programmatic improvement.
3. These interviews also provided candidates an opportunity to share their perceptions of teaching strengths and areas of focus for teaching before, during, and after the TPA. Candidates were asked to articulate their philosophy of education, or view of good teaching, and whether they felt that they were able to explore and apply this philosophy on the TPA.
4. Visits were made to each of the subjects' school placement sites to better understand the school environment, available resources, and the student body at each site. Candidates also participated in the observation of one of their taught lessons in the placement context. Candidates were asked to respond to reflection lesson debriefing questions after the researchers' observation. In a post-observation conference, candidates were asked to explain their plan, how well they executed that lesson plan, and what they would have done differently. Case study subjects provided their submitted TPA for review. Using the assigned TPA scores for each submission, these samples can help to provide evidence to document the reliability of scoring judgment.
5. Finally, to triangulate and explain the survey responses, candidates were asked to elaborate on questions asked in electronic surveys.

Analysis of Data

The core question of this project is whether the TPA is a strongly credible measure for determining teacher readiness. This core question is further divided into the nine sub-questions of the VA. To validate this argument, TPA and investigation data were collected to address these sub-questions and the evidence gathered was both independently and holistically analyzed. These data include: quantitative data (surveys, evaluations, test scores), qualitative data (focus group interviews and document analysis), and case study data (interviews, lesson observations, TPA samples).

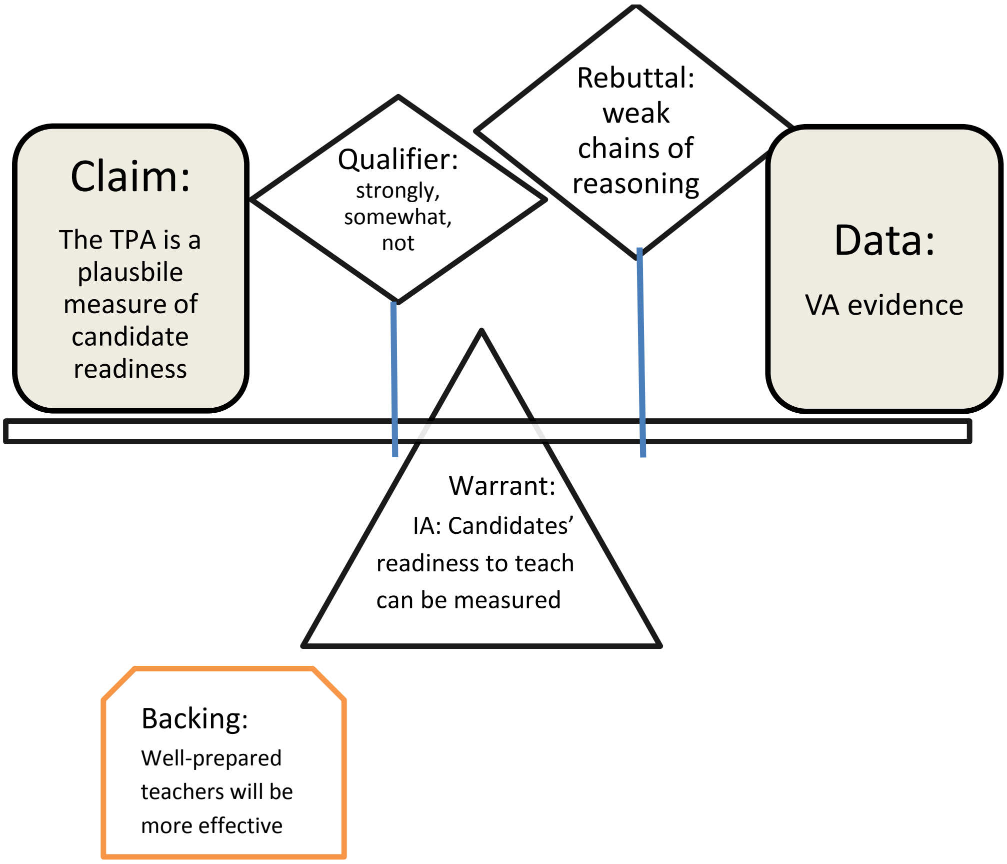
The design of this study is based on Kane's ABV where assumptions about, and consequences of, the assessment are used to determine whether it is a plausible measure of the assessed construct. The analysis of the data collected is particularly multifaceted and complicated because the validation argument includes a complex set of data sources collected, applying seven distinct kinds of data and twenty-seven separate instruments. To add further complexity, framing the core research question in terms of an argument-based interpretation led to nine sub-questions, each of which was separately addressed, in order to respond to the central question.

The methodology of ABV was selected because it provides a framework for a more robust discussion of validation, in comparison to other frameworks in which validation might take place. In the case of this study, the strength of the ABV also made it difficult because no one piece of evidence alone can (should) answer the research question (or sub-questions). It is the interaction of the data that allows for a full-bodied conversation and evaluation of the RQ. For this reason, the data collected must be analyzed both independently and in combination to address the questions of the study. Thus, the study is multiple methods because that is the methodology that the IA demanded to respond to this particular set of research questions.

Kane's framework applies Stephen Toulmin's (1958) model of argumentation to ABV inquiries. The Toulmin model is based on textual dissection of an argument by breaking down the argument into different parts. The parts of an argument are: (1) Warrant, (2) Claim, (3) Data, and (4) Backing. Once the argument is dissected, it can be better understood and a judgment can be made about its strength and value.[26] Applying Toulmin's argumentative model is helpful in validation efforts since test-developers and test-users make claims about the instrument that need to be evaluated. The argumentative claim that the TPA is a valid measure for determining teacher readiness needs to be supported. The IA serves as the warrant. Figure 1 applies Toulmin's model to this study.

Figure 3.1

Toulmin's Argumentative Model Applied to the TPA


Claim:
The TPA is a plausbile measure of candidate readiness
Qualifier:
strongly, somewhat, not
Rebuttal:
weak chains of reasoning
Data:
VA evidence
Warrant:
IA: Candidates' readiness to teach can be measured
Backing:
Well-prepared teachers will be more effective

Toulmin's model will be applied in the argumentative analysis of the data collected in this validity investigation using the shorthand of "support for validity" and "threats to validity" (Shaw & Crisp,

[26] See **Appendix G** for Toulmin's model.

2012, p. 165). As discussed in Chapter two, the IA is the statement of inferences and assumptions that support the consequences and uses of test score decisions.

Types of Evidence

Because each of the sub-questions in the chain of evidence required multiple sources of evidence, different types of evidence analysis were applied. The IA demonstrates that the claims for the TPA, as high-stakes for both candidate licensure and as a larger accountability measure, are quite ambitious. Therefore, “more evidence and more kinds of evidence” were required (Shaw & Crisp, 2012). Table 3.31 summarizes the types of evidence applied and the analysis method utilized.

Table 3.31

Types of Evidence and Data Analysis Method

		Types of Evidence	Analysis Methods	Validation Instrument #
Qualitative	1	Interviews	Coding	1, 5, 10, 14, 19, 20, 25, 26
	2	Documents	Document Analysis	27
	3	Observations	Coding / Factor Analysis	8, 18
Both	4	Surveys	Coding / Transformation	2, 4, 6, 7, 9, 11, 13, 15, 17, 24
Quantitative	5	Course Evaluations	Descriptive/Factor/Bivariate Analysis T-Tests Interaction Graphs Inter-rater Agreement MTMM	12, 16, 21
	6	Candidate Self-Reports		12, 16, 22
	7	Test Scores		3, 23

Qualitative data. Table 3.31 indicates that there are seven different types of evidence used in this study. Three of these types were collected as qualitative evidence. The first type of qualitative evidence was interviews. Interviews were conducted with six case study candidates once in each of the four phases and with candidates, supervisors and faculty in the final phase of research. Interviews were transcribed and transcriptions were used for data analysis.

Coding. To analyze the interviews, the researcher blocked, coded, and labeled data using the central themes developed from the IA: 1. Teaching Readiness; 2. Construct Representation; 3. Scoring/Evaluation; 4. Generalizability/Fairness; and 5. Decision-Making. Within these three labels, further sub-themes developed (see **Appendix G** for codes and descriptors).

This process of grouping evidence and coding based on labels allowed larger and broader themes and patterns to immerge from the data and thus to make meaning of the "chunks" of words, phrases, sentences or larger statements and whether they were connected to specific settings or experiences (Miles & Huberman, 1994, p. 56). The themes and patterns were then related, compared, and synthesized based on the research questions.

After analyzing the qualitative data for codes and themes, and when appropriate for each research question, these qualitative data procedures were also used:

- Counting the number of instances of each of the codes, themes, and patterns.
- Entering those numbers into SPSS to generate a visual portrayal of the results using graphs, charts, and tables.

Document analysis. The second type of qualitative data was document analysis. Deliberate primary sources were sought such as the TPA Handbook, including the instructions and the rubrics, the standards documents for both WA licensure and national certification, and the description of the consequences and uses described for the assessment by SCALE, AACTE and PESB were analyzed. While most of these documents are available for research, the TPA handbook is a proprietary document owned by SCALE. In addition, scoring guidelines, documents and procedures are copyrighted by Pearson. Access was sought for this study and, ultimately, granted by SCALE. For publication purposes, description of the handbook is limited to:

- Pages 1-2, regarding the purpose of the TPA as described to candidates and to clarify the differences between the national TPA and the WA TPA.
- Page 12, to provide general rubric descriptions and to establish rubric levels that met standard.
- Page 19, to provide a description of how a candidate might score an "Automatic 1."
- Rubric numbers, names and guiding questions will be used to frame the test construct and to explain candidate scores, where needed.[27]

The approach to the documents included a careful reading for both witting and unwitting evidence. The evidence and information the author of the document *wanted* to share is witting evidence. However, in some cases, there is much more that can be inferred from the document, for instance about the biases and underlying assumptions around the domain and uses of the assessment scores. This is unwitting evidence (Bell, 2010, p. 127). Applying critical analysis of the documents called Internal Criticism, the content of each document was analyzed to seek answers to these questions:

1. What does the document *say*? What does the document *mean*? In the case of the TPA handbook, what does it *ask* of candidates? When specialized language is used, was it used in the same way as the researcher? Look for witting and unwitting evidence.
2. Who produced the document? What is known about them? What is the author's (or organization's) role, background, past experience and aims, especially with regard to political views around accountability for teachers and teacher-trainers? What is their role in the profession and decision-making process around the TPA?

[27] Copyright © 2013 Board of Trustees of the Leland Stanford Junior University. All rights reserved.

3. What is the purpose of the document? Who, when and for what reason was it produced? Is it a typical example of a document with this type of purpose?
4. Is the document a revision or edited copy of an original? If so, in what way has it been changed?
5. Is this document reliable? (Bell, 2010)

Given the number of TPA handbooks[28] and standards documents, a sampling strategy was devised for a balanced but manageable selection. In standards documents, critical analysis focused on the knowledge, skills, dispositions and any performance rubrics used to describe and explain the construct.

Then a checklist was developed by modifying Bell (2010):

1. Decide whether and how to use the evidence from this document.
2. Compare the document with the other sources used. Is there bias?
3. Code the document. (pp. 138-39)

Observation. Finally, case study participants were observed during one taught lesson and all candidates in the undergraduate program were observed during the three days writing on campus. Researchers used the same university observation tool employed for all formal observations during the lesson observation in order to collect comparable data on candidate performance. Candidates were interviewed immediately after their lesson and asked to provide written reflections of their practice. During the writing up, candidates were digitally recorded and their questions, comments, and concerns coded using the method discussed above.

[28] For the TPA, the Secondary Social Studies (SS) handbook was carefully reviewed and used to represent the TPA as a whole. Other TPA were compared to SS to highlight their unique criteria and differences.

Mixed Data.

		Types of Evidence	Analysis Methods	Validation Instrument #
Both	4	Surveys	Coding / Transformation	2, 4, 6, 7, 9, 11, 13, 15, 17, 24

Some of the instruments used in this study collected both qualitative and quantitative data. Surveys were administered to candidates in each of the four phases of the study, to mentors and supervisors at the beginning, mid-point, and end of the study, and to faculty at the end of the study. Surveys asked both qualitative and quantitative questions. For instance, on one survey a respondent may have answered both open-ended questions requesting a unique, personal response, nominal questions categorizing their response (for instance age range), or ordinal questions, such as responding using a provided Likert scale (Likert, 1932). Open questions were coded using the method discussed in the previous section. Other survey questions provided qualitative data that could be transformed in order to compare qualitative and quantitative data collected. For quantitative analysis, SPSS was used.

As is the case with any multiple method data research study that involves the coding of qualitative research, care was taken to be as accurate as possible in coding the data to develop themes and patterns, but some judgment is always used in determining how to code a datum. To minimize the potential threat to the validity of the data collection and analysis process, the following was practiced:

- Quantitative and qualitative samples were drawn from the same population.
- When contradictory results were found, the data was reexamined and, in the case study interviews, whenever possible, followed up.
- As much as possible and appropriate, data collection procedures were straightforward and standardized.
- The codes and labels used to transform the information were kept straightforward and as simple as possible.

- The same research questions were used for both qualitative and quantitative data.

Quantitative Data.

		Types of Evidence	Analysis Methods	Validation Instrument #
Quantitative	5	Course Evaluations	Descriptive/Factor/Bivariate Analysis T-Tests Interaction Graphs Inter-rater agreement MTMM	12, 16, 21
	6	Candidate Self-Reports		12, 16, 22
	7	Test Scores		3, 23

Researchers often try to quantify things that are difficult to directly measure in the social sciences. Teaching readiness is one of those difficult to measure constructs. Several data sources collected in this study are quantitative (course evaluations, candidate self-reports, TPA test scores) and were analyzed using SPSS. In particular, comparison and contrast and correlation analysis was used to provide a robust answer to RQs.

Descriptive/Factor/Bivariate analysis. Using SPSS, data was analyzed by both looking at the quantitative and transformed data graphically, finding patterns, and also fitting statistical models to the data. Exploratory factor analysis is used for identifying groups and clusters of variables. Frequency charts, interaction graphs, t-tests, and plot tests were used to assess and report the data (i.e., the distribution of scores, central tendencies, variances, and error). Differences between two variables and relationships between two variables are reported.

MTMM. Several instruments asked participants and scorers to provide data on a candidate's performance traits (TPA scores, mentor course evaluation, supervisor course evaluation). Because these traits, referred to as standards (PESB, INTASC, NB, Conceptual Framework) or content constructs (TPA), are aligned across instruments, correlations were made that helped to respond to validity strengths and challenges for the TPA construct. As discussed in

Chapter two, the overarching construct for the TPA is teaching readiness but that is further defined in the fifteen separate rubrics, each of which can be considered its own construct. A Multi-Trait, Multi-Method (MTMM) (Campbell & Fisk, 1959) analysis was applied to find the similar and dissimilar traits and the differential effects of these various collection methods. In particular, MTMM evaluates convergent and discriminant validity. Convergent validity identifies the degree to which related theoretical concepts are interrelated in reality. Discriminant validity addresses the degree to which purposely unrelated theoretical concepts are also unrelated in reality. To address the research questions that require construct validity, MTMM can demonstrate both convergence and discriminant validity. To argue that an assessment has strong construct validity, a plausible claim must be made in both convergence and discriminant validity. Of course, it is not possible to measure all traits with all methods and so only those that can be cross-referenced are reported with MTMM.

Summary of Data Analysis.

To summarize, the following table provides an overview of the validity methods used in the study and the validity evidence collected to address each research question. An asterisk (*) indicates the use of TPA data (See **Appendix I** for IA and Evidence).

Table 3.32

Summary of Validity Argument Question, Evidence, and Method

Validity Question	Validity Method	Validity Evidence
	Inference 1: Construct Representation	
1. Do the tasks elicit performances that actually reflect the intended construct (teacher readiness)?	1. Analysis of documents to correlate teaching standards and performance criteria. 2. Analysis of performance data using statistical methods (descriptive, factor, bivariate analysis) from the sample of candidates to explore the relationships between items and constructs. 3. Analysis of examiner/expert responses and opinions about the types of cognitive demands on candidates, construct irrelevant variance, and construct validity. 4. Analysis of performance data and corresponding scores for insights as to how the questions were answered by candidates and scored by examiners.	Documents* Test scores & submissions* Interviews Course evaluations Surveys Factor analysis* Item analysis* MTMM*

	5. For misfitting items, analysis of candidate survey and case study responses to gather insights into sources of construct irrelevant variance.	
Inference 2: Scoring/Evaluation		
2. Are the scoring procedures sound and reliable? 3. Are the rubric score levels achieved by the candidate actually representative of what that candidate performed on the TPA? (Does candidate performance correlate to the assigned test score)?	1. Review of SCALE/Pearson documents on marking and scoring procedures. 2. Analysis of the reliability of scores. 3. Statistical analysis of candidate exam results. 4. Statistical analysis of candidate exam results in relationship to results of similar traits collected from different instruments. 5. Analysis of candidate responses on surveys and in case study interviews.	Documents* Test scores* Interviews Course evaluations Self-evaluations Surveys MTMM*
Inference 3: Generalization		
4. Are the score levels achieved on the rubrics a true representation of a candidate's performance? In other words, are the scores a candidate earned consistent and generalizable with other samples of that candidate's teaching performance? 5. Does poor performance on the TPA imply a lack of adequate mastery of the construct? 6. How generalizable are the criteria, rubrics, procedures, and scores derived from the TPA across different candidates and handbooks? 7. Does TPA proficiency depend upon factors beyond the candidate's control? How generalizable are the criteria, rubrics, procedures, and scores derived from the TPA across testing sites, placements and placement length and programs?	1. Analysis of examiner/expert opinions on TPA construct and rubric constructs and the tasks that evaluate candidate KSJ within each construct. 2. Analysis of mentor/ supervisor opinions of the TPA construct and rubric constructs and the tasks that evaluate candidate KSJ within each construct. 3. Analysis of documents to compare and contrast teaching standards to performance criteria on tasks. 4. Analysis of candidate TPA scores. 5. Correlation of candidate TPA scores in relationship to student teaching success. 6. Analysis of candidate responses on assessment procedures from observations, case studies and surveys. 7. Analysis of candidate TPA scores correlated to supervisor/mentor course evaluations and lesson observations. 8. Analysis of candidate responses on assessment procedures from observations, case studies and surveys. 9. Analysis of candidate TPA scores correlated to supervisor/mentor course evaluations and lesson observations. 10. Analysis of candidate TPA scores across programs and disciplines. 11. Analysis of reliability of scores.	Documents* TPA Scores* Surveys Interviews Course Evaluations Observations Self-Reports MTMM* Case Study Interviews Document Review* Composite Reliability Analysis*

Inference 4: Extrapolation		
8. Do TPA test scores provide reliable indicators of a readiness to teach? Are the TPA scores, as a whole, a true measurement of teaching ability?	1. Correlation of candidate TPA scores in relationship to student teaching success. 2. Analysis of candidate responses on assessment procedures from observations, case studies and surveys. 3. Analysis of candidate TPA scores correlated to supervisor/mentor course evaluations and lesson observations. 4. Analysis of candidate TPA scores across programs and disciplines.	TPA Scores* Course Evaluations Observations Interviews Surveys MTMM*
Inference 5: Decision Making		
9. Guidance is in place so that all stakeholders know what scores mean and how the outcomes will be used?	1. Analysis of stakeholder surveys and interviews gathering views on what they know about the TPA procedures, scoring, meaning and uses. 2. Review of guidance documents, the TPA handbook, and materials relating to score meaning and use.	Documents* Surveys Interviews MTMM*

The ABV framework and the IA developed for this study mandated that the VA include multiple qualitative and quantitative data for each of the RQ. This section outlined the data analysis methods used, based on types of evidence collected.

Limitations

As is often the case with ABV, one of the limitations of this study is that the breadth of data collected was extensive. Though the TPA is one assessment, its complexity, and the numerous types of data required to validate each inference, meant that this was a broad study, rather than an in-depth review of a single construct (i.e., one set of rubrics or one task). Compared to other validation studies, complexity in the number of types and instruments limited a detailed analysis and discussion of any *one* part of the instrument, or any *one* sub-group of participants. However, some data is more significant in addressing the research questions and the IA of this project. Those data include the TPA scores, course evaluations, and the case study interviews. These will be analyzed in more depth because they helped to better explain the consequences and uses of the TPA. They are, simply put, more significant sources of evidence. Recommendations and suggestions for future studies will be discussed in Chapter five.

Another limitation in this study is that the participants came from different programs that did not always collect the same sources of information on candidates' performance during ST. For this reason, it was often the case that the data analyzed applied to only one of the programs, approximately half of the participants. Where similar instruments were collected, they were not always identical. Efforts to simplify the data by coding same standards and traits made it possible to compare these evidence types. When comparison was not possible, or the population size is affected, it has been indicated.

Finally, the principal investigator was charged with assisting candidates in their understanding and completion of the TPA during the field test. Because her role was to help candidates perform *well* on the TPA, the struggles, concerns, and advice revealed in the early phases of data collection were immediately applied for programmatic improvement. As such, the data is irrevocably shaped by the investigators influence. For instance, TPA scores were changed by these programmatic improvements, in ways that cannot be fully measured. Whenever possible, the investigator noted when a programmatic change resulted from research collected *in process* from this study and will be discussed as the evidence is analyzed and presented in Chapter four.

Conclusion

ABV guided a set of comprehensive procedures for the development of the research design, including the IA, based on the assumptions and claims of the uses of the TPA scores. A series of assumptions determined the types of data to be collected, which were analyzed using both qualitative (coding) and quantitative (statistical) methods. The methodologies used in the analysis of the data are varied and diverse because the types of data are similarly diverse, as is the nature of a multiple method study, and requisite in order to fully answer the research questions. Using ABV, this chapter proposed five core inferences and nine sub-questions that determine the research questions of this study. In Chapter four, the data results from the validation study and the findings for each of the research questions will be shared.

Chapter Four

Validity Argument

The primary research question guiding this study is whether the intended interpretations and uses of TPA test scores are an appropriate and valid measure for determining teacher readiness. In order to answer this question, an ABV methodology was adopted. The methodology first required the development of an interpretive argument (IA) identifying the assumptions underpinning the assessment. The IA was further analyzed into five inferences, each with its own question(s) and assumptions. The final step, the validity argument (VA), involved collecting data to address each of the inferences in order to justify the IA. This chapter presents the findings of the VA using the Cambridge framework developed by Shaw, Crisp and Johnson (2012). Evidence is organized by inference, including the corresponding research question(s), analysis method, and findings, followed by a discussion of both supports and threats for validity. Finally, the chapter concludes with a synopsis of brief study evidence.

Inference 1: Construct Representation

Research Question

1. Do the four TPA tasks represent the six categories (operationalized construct) relevant to the theoretical construct (teacher readiness)?

Validity Evidence 1: Expert Opinions

To determine whether the operationalized construct of readiness in the TPA are aligned with the theoretical construct of readiness, experts were interviewed and surveyed. Specifically, mentors, supervisors and university faculty were asked to provide the characteristics that exhibit candidate readiness to teach (Phase 1) and their judgment of the TPA, as a measure of readiness (Phase 4).

Findings. Like most performance assessments, the results of expert opinion indicate that there is strong evidence to suggest that the construct is measured by the TPA. All measured TPA traits were selected as “essential,” “important,” or “somewhat important.” Classroom management was the one trait not aligned to the TPA construct but ranked as “essential” by experts (.76 of mentors, .57 of supervisors) (Pre-TPA Experience Survey or Mentors, 2/20/12; Pre-TPA Experience Survey for Supervisors, 2/21/12). These findings suggest that experts support the choice of traits (Tasks and Categories) selected for performance in the TPA (see **Appendix J**).

However, after working to complete the TPA, these same experts are less positive that the TPA reflects the traits essential to teaching readiness. Table 4.1 reports expert responses in Phase 4. Experts were asked, “Based on your experience mentoring/supervising/teaching during the TPA, which of the following statements best reflects your view of the TPA?” The majority responded that it was “a good instrument but may not indicate candidate readiness to teach.”

Table 4.1

Expert views of TPA as measure of candidate readiness (Phase 4)

“The TPA is…”	Mentors n=24	Supervisors n=14	Faculty n=12
A good instrument measuring candidate readiness to teach	4%	21%	17%
A good instrument but may not indicate candidate readiness to teach	50%	57%	50%
Not a good instrument	17%	7%	8%
I am unsure	29%	0%	8%

Several respondents offered an explanation for their ranking. Many described the concern that the TPA must be part of a larger framework of evaluation. For instance one expert wrote, “At first review the TPA looks to be a good instrument. However, it cannot stand alone. Evaluation and support of the mentor and supervisor are essential for a more complete profile of the candidate” (4/29/12). Faculty offered similar comments. One professor penned, “It is an instrument that concerns me if used as the only, or the definitive, determination for certification” (4/27/12). Another supervisor explained:

> I think the TPA is an instrument through which teacher candidates can look at their teaching practices and reflect deeply about them. . . . [But candidates] are learning about so many aspects of being a good teacher, and I have noticed this has been a distraction. The quality

> of their teaching actually dropped when this assessment was added to their workload. (4/29/12)

Much like the supervisor comment above, concerns about the timing of the TPA during student teaching (ST) and the impact of the assessment on the developmental process of a teacher dominated faculty comments. Namely, the TPA, "Is so high-stakes that it eclipses everything else about student teaching" (4/27/12).

To find parallels between the experienced theoretical construct and the operationalized construct of the TPA, experts were asked to clarify traits that measured candidate readiness. Responses ranked direct observation of candidate teaching as the most highly valued measure of readiness (.42), yet direct observation only constitutes 13% of the candidate score on the TPA. Similarly, candidate ability to collaborate and develop rapport with students and their families are listed as necessary by 28% of respondents, but these traits are not measured by the TPA.

Interviews. In Phase 4, Supervisors participated in a small group interview and were asked to expand on their survey responses regarding the theoretical construct of teacher readiness. In addition to missing traits, experts reported that the TPA measured something unintended, or measured something not relevant. The three missing traits include dispositions (.57), collaboration (.57), and classroom management (.28) (see **Appendix K** for missing traits). Supervisors also reported that TPA performance required mastery of construct irrelevant skills including a positive partnership with a willing mentor (.86), teaching with provided materials (.71), ability to use technology (.64), and reading and writing ability (.57) (see **Appendix L** for irrelevant traits).

In the interview (5/10/12), supervisors offered feedback on their experiences working with candidates during TPA submission. The following statements help to explain why expert opinion shifted between Phase 1, in which they supported the TPA theoretical construct, and Phase 4, when they argue that the TPA may not be a good operationalization of teaching readiness:

- "[The candidate] said that on the commentaries you could write anything. You can write anything and they wouldn't even know if it was true. You could get by on writing a wonderful commentary on everything that didn't happen and the scorer wouldn't even know."
- "And I have students calling me and crying because they cannot figure out what the words in the glossary even mean. They say, 'What do they mean by academic language? What does this mean? What do I do? What is this thing?' It is like reading a legal document. It is not a friendly document."
- "If you don't have management and relationships, you are not a positive influence on kids' lives and it [the TPA] has nothing that reflects that."
- "After looking at all the work the student teachers did, all the planning, those video clips, editing those minutes, the commentaries they had to do, in my opinion, it was an enormous waste of their time. It is such a tiny little segment of what they are doing and you [scorers] can't, you can't isolate a situation like that and get anything valid out of it. . . . It is a lot of stress and effort for no real learning."
- "It is too small of a slice of the pie. You take this many minutes, that you cannot edit, and you can only use those minutes to score. Wonderful things might be happening right outside of that, but you cannot use it. . . . And it happens way too early in the program [term] and it is scored by somebody who doesn't know the student teacher, the students, the teacher, the school, what happened at recess. It is unreliable."
- "When the mentor is not onboard with the TPA, there is additional pressure and external forces that the student can do nothing about. So while the student is trying to make their way through, they have the mentor expressing a lack of support."

- "One of the difficulties that one student teacher had is that everything is scripted and laid out and we have not taught them how to teach that way. She was disappointed because she wanted to use her creativity but was told that 'today is page 50.'"
- "When you have the scripted programs there is not a lot about the candidate that is reflected in the TPA. They are at a true deficit to show their creativity. They need an instrument that goes with the district programs. I had one in [school district A] and one in [school district B] and it was a really different TPA experience for each of them."
- "I must say the mentors with my kids were all willing to do the camera work and work it out but still what you get from them, those little eight minute segments, or whatever they are, are just disingenuous. This is not teaching. It becomes about filming and editing and scoring and not at all about teaching."

Concerns reflected in these comments reoccur throughout the study, from mentors, faculty, candidates and case study participants and will be further explored in Inference 3 and 4.

In addition to supervisor feedback, faculty were interviewed in Phase 4 about how well the TPA represented the construct of teaching readiness. The majority of faculty (.83) responded that important components of readiness were omitted from the TPA including:

- character and dispositions,
- creativity,
- fostering an enjoyment of teaching,
- classroom management,
- collegiality and leadership skills in the school setting,
- the centrality of relationships with both students and parents,
- and cultural competence.

Support for Validity. Experts agree that the theoretical constructs of the TPA are the characteristics that exhibit candidate readiness to teach. Some experts agree that the TPA measures teacher readiness.

Challenges to Validity. The majority of experts believe the TPA is a good instrument, but may not operationalize teacher readiness because there are some aspects of teacher readiness that are not measured in the TPA. These include what is often described as "dispositional" traits, such as collaboration, interest in students, and rapport. In addition, experts indicated that a potential weakness in the operationalized construct was the lack of a classroom management rubric. Finally, a concern was shared by many experts that success on the TPA required skills and abilities that were construct irrelevant, such as a candidate's ability to write and use forms of technology needed to complete the TPA, like video equipment. These concerns are further examined in Inferences 3 and 4.

Omission of traits identified by experts is not necessarily problematic, so long as the university has the ability to supplement the data collected on candidate readiness during ST. However, claims that TPA performances are impacted by construct irrelevant traits are more serious. In these cases, high stakes decisions about candidate performance may be questionable. If the university is able to endorse (or not) a candidate for licensure, despite TPA scores, based on these missing traits (for instance, observation of candidate performance outside the TPA lessons, or failure of a candidate to demonstrate rapport with students, or to collaborate in teaching), such concerns can be addressed. However, if university judgment of a candidate's readiness is to be solely (or primarily) determined by TPA scores, as proposed by the state, it appears likely that, (1) candidates who perform well on the TPA but not well in ST could be licensed; or, (2) candidates that perform well in ST but not on the TPA could fail to be licensed; or, (3) traits that are significantly important to teacher readiness, but not measured on the TPA, could fail to be evaluated during ST while traits irrelevant to readiness, but measured by the TPA, will be evaluated instead.

Validity Evidence 2: Standards Review and Document Analysis

The theoretical construct described in the PESB Standards for Initial Licensure (PESB, 2013) were compared to the operationalized construct as described in the TPA rubrics (SCALE, 2012). An alignment was published by PESB. While a preliminary document, it demonstrates how the TPA collects evidence in support of the thirteen criteria (organized into three standards) of Standard V, using the Elementary Mathematics handbook. The document was analyzed to determine alignment between the theoretical and operationalized constructs of readiness.

Findings. Of the three PESB standards aligned to TPA rubrics, "professional development" traits were 100% aligned and "effective teaching" was 80% aligned. The standard "teaching as a profession" is "not applicable" (see **Appendix M**). PESB notes that the strength of the alignment between components varies and the alignment may be "ambiguous" (PESB, 2013).

Support for Validity. The PESB document describing Standard V, as the theoretical construct identified for teacher readiness in WA, is well aligned to the TPA.

Threats to Validity. None identified.

Inference 2: Scoring and Evaluation

Research Questions

2. Are the scoring procedures sound and reliable?
3. Are the rubric score levels achieved by the candidate actually representative of what that candidate performed on the TPA? (Does candidate performance correlate to the assigned test score)?

Validity Evidence 3: Internal Consistency with Cronbach's Alpha and Factor Analysis

Test reliability is a validity concern primarily aimed at the accuracy of measures. Reliability can be determined in two ways. First, reliability is based on internal consistency, the precision and regularity of test scores on one testing event. Second, reliability is concerned with the stability of the test scores over time; and stability is often measured through a test-retest. The reliability of scoring, or inter-rater agreement, is critical for determining internal consistency. As a new, high-stakes field test, data to determine the reliability of test scores over time was not available. For this reason, internal consistency and inter-rater agreement of TPA scores was the primary objective of the traditional analysis of test score data.

Exploratory factor analysis was conducted in SPSS on item level score data to explore the traits that may underlie the test scores for the fifty-eight participating candidates. For example, are the rubric items within a task testing the same construct (trait), or are some rubrics testing unrelated skills such as writing or technological ability? Factor analysis can provide information to explore relationships between scores on different rubrics and provide insights about how, together, the rubrics contribute to the measurement of a single construct, how they might measure different constructs, or unintended traits (construct validity). If the latter occurs, it raises a question about the relevancy of the trait being assessed and the meaning for the construct, as a whole. Factor analysis is a well-established methodology for determining construct validity. Cronbach and Meehl (1955) argued that factors can be considered synonymous with constructs on an assessment.

Findings. To determine reliability across all TPA subject areas, Cronbach's alpha (α) was calculated for each of the first three tasks and the two embedded categories. Cronbach's α is a common indicator of the internal consistency of a testing method and is used to estimate the reliability of scores across a sample of test-takers. Task 4 is measured by one rubric and could not be calculated using Cronbach's α. The three subscales of the TPA all had high reliabilities, with Cronbach's α= <.731 (see **Table 4.2**).

Table 4.2

Cronbach's α by TPA Task and Category

	Cronbach's α
Task 1	.839
Task 2	.779
Task 3	.805
Academic Language	.792
Student Voice	.731

A principal component analysis (PCA) was conducted on the sixteen rubrics (rubric 2 reported two sets of scores) with orthogonal rotation (varimax). The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO=.835 ("great" according to Field, 2009), and all KMO values for individual items were above the acceptable limit of .5 (Field, 2009). Bartlett's test of sphericity x^2 (120) =465.068, p < .001, indicated that correlations between items were sufficiently large for PCA. In factor analysis, eigenvalues help researchers determine how many variables are significant. An initial analysis obtained eigenvalues for each component.

Table 4.3 shows an abridged factor analysis output table displaying the total variance explained. The "total" column indicates the eigenvalues corresponding to the three factors of interest. The "% of Variance" column demonstrates how much variance can be explained by each of the three individual factors. The "Cumulative %" column shows each consecutive factor added together to indicate the amount of variance. Three components had eigenvalues greater than one. An eigenvalue of one indicates that the factor variability explains as much as a single original variable might (Shaw, Crisp, & Johnson, 2012). The total variances explained by the first three components were the most influential.

Table 4.3

Abridged Factor Analysis

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
Task 3: Assessment	7.095	44.343	44.343
Task 1: Planning	1.697	10.607	54.950
Task 2: Instruction	1.286	8.037	62.987

Three components had eigenvalues over Kaiser's criterion and in combination explained 62.99% of the variance. Figure 4.1 shows the scree plot for analysis. The scree plot displayed inflexions that would justify retaining components one through four. Given the sample size, and the convergence of the scree plot and Kaiser's criterion on three components, three is the number of components that were retained in the final analysis.

Figure 4.1

Scree Plot Analysis

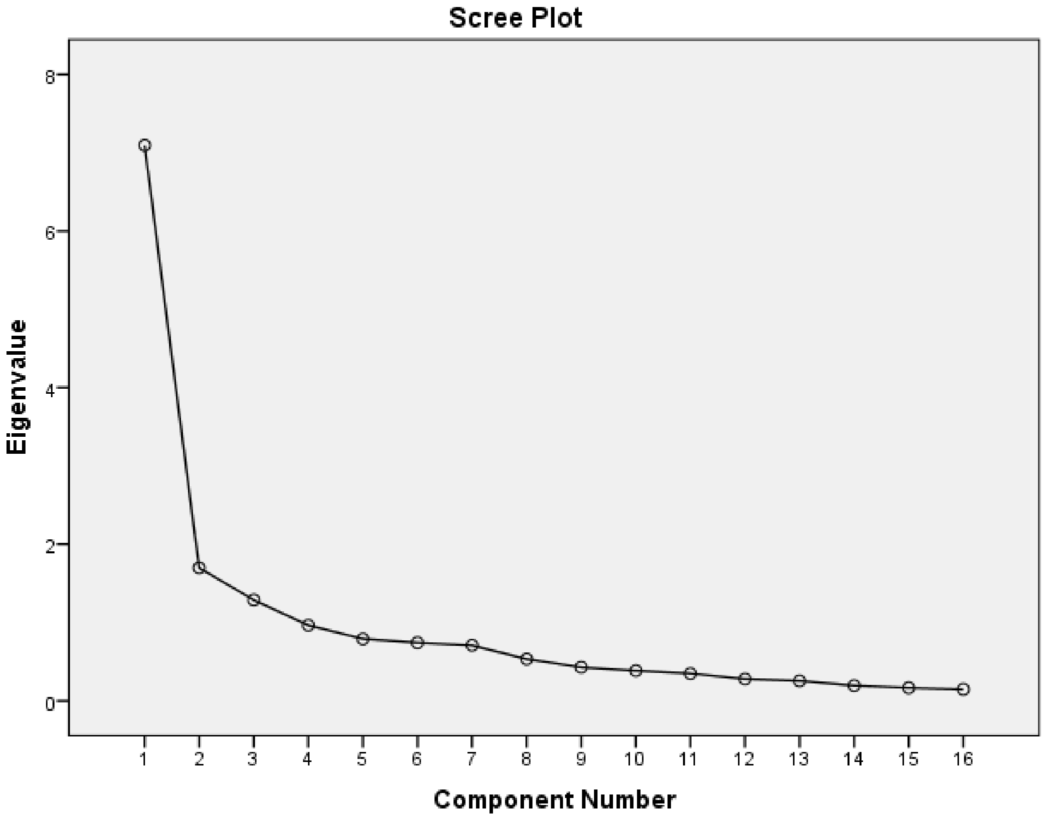


Table 4.4 shows the related component matrix for factor loadings after rotation. The rotation allowed questions relating to each factor to be considered for commonalities. The items that cluster on the same components suggest the following inferences about the meaning of the factors:

- Factor 1 rubrics appear to measure teaching readiness around *assessment* and correlate with *assessment of student voice* and *Task 4: professional reflection*.

- Factor 2 rubrics appear to measure teaching readiness to *plan the taught segment* and correlate with *planning for academic language*.
- Factor 3 rubrics appear to measure teaching readiness around *observed practice* from the instructional video, including *observed use of student voice and academic language*.

Table 4.4

Rotated Component Matrix

Rotated Component Matrix[a]

	Component		
	Assessment	Planning	Instruction
Student Voice: Supporting Student Use of Resources to Learn and Monitor their own Progress	.769		
Assessment: Using Assessment to Inform Instruction	.723		
Assessment: Using Feedback to Guide Further Learning	.719		
Student Voice: Reflecting on Student-Voice Evidence to Improve Instruction	.665		.438
Assessment: Analyzing Student Work	.630		
Planning: Planning Assessments to Monitor and Support Student Learning	.519	.486	
Analyzing Teaching Effectiveness	.509		
Planning: Planning for [Content Specific] Understandings		.841	
Planning: Using Knowledge of Students to Inform Teaching and Learning, A		.821	
Planning: Using Knowledge of Students to Inform Teaching and Learning, B		.817	
Academic Language: Understanding Students' Language Development and Associated Language Demands	.436	.621	
Academic Language: Developing Students' Academic Language and Deepening Content Learning		.509	.415
Instruction: Engaging Students in Learning			.853
Instruction: Deepening Student Learning			.774
Student Voice: Eliciting Student Understanding of Learning Targets	.472		.534
Academic Language: Scaffolding Students' Academic Language and Deepening Content Learning			.521

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.[a]

a. Rotation converged in 5 iterations.

Support for Validity. The factor analyses indicated generally good coherence of the factors assessed by Task 1: Planning, Task 2: Instruction, Task 3: Assessment, and the embedded categories of academic language (AL) and student voice (SV).

Threats to Validity. None identified.

Validity Evidence 4: Internal Consistency with Rubric Scores

Rubric scores were reported for all participants. Mean scores were calculated in SPSS. Using the cut score published by the PESB in November 2013 (PESB, 2013), pass rates were determined.[29]

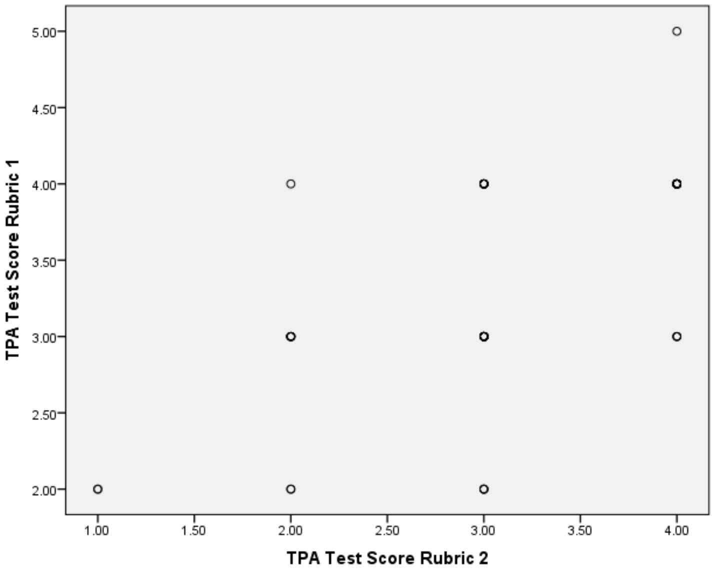
Findings. Four (7%) of the candidates in this sample would have "failed." These candidates were in the graduate program. Two males (14%) and two females (5%) did not meet cut scores. Three of the four candidates submitted Elementary Literacy TPA and the fourth submitted a Secondary Mathematics TPA. Cut scores indicate that candidates can score an average of 2.2 on each rubric and pass (see **Appendix N** for raw scores).

Scatterplots were conducted comparing the scores from the rubrics. Figure 4.2 compares the scores from Rubric 1 and Rubric 2 and is a sample that illustrates the overall trend for all scatterplots. The pattern indicates that scores occurred across the spectrum of rubric levels.

[29] In this field test, only raw scores were reported. Score meanings (pass/fail) were not known.

Figure 4.2

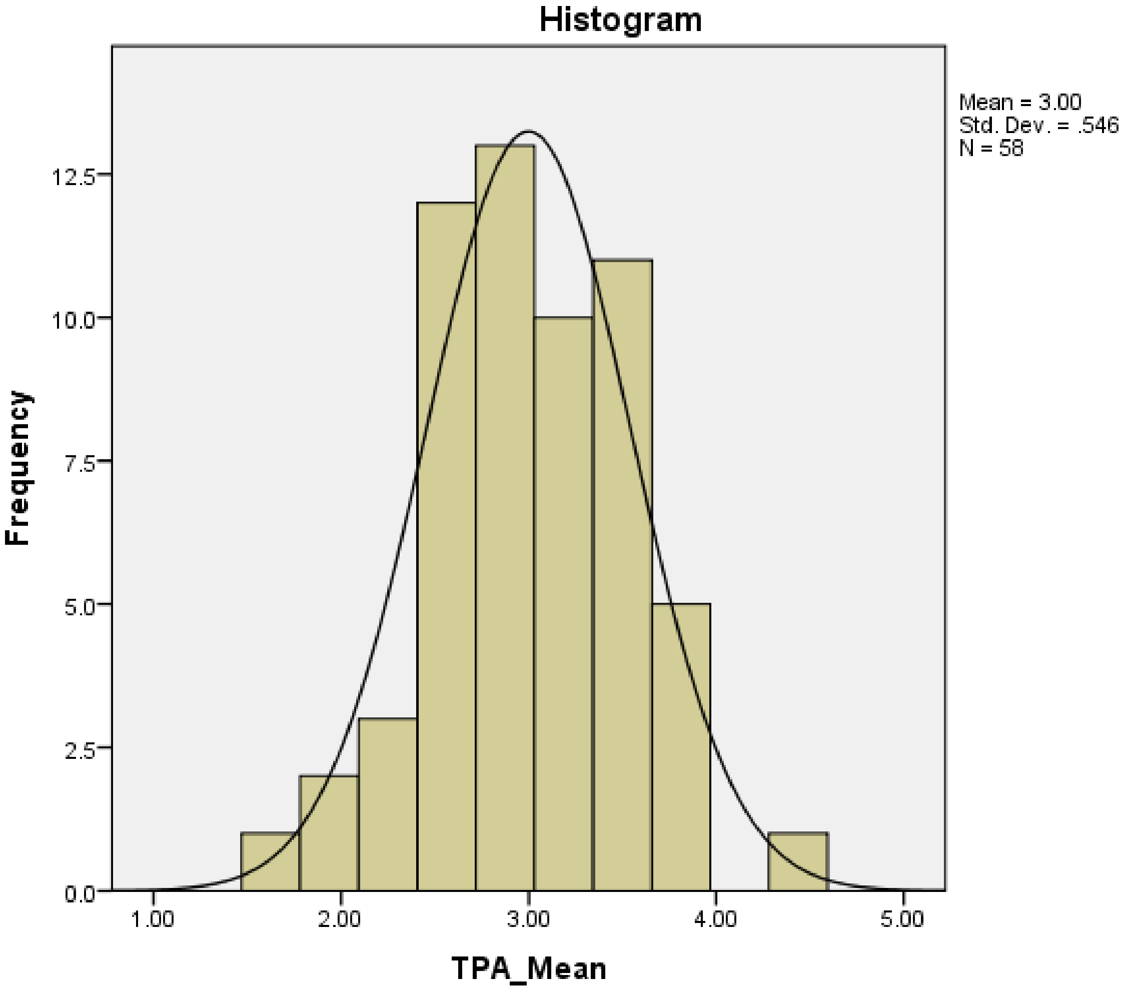
Scatterplot of Rubric 1 and Rubric 2 Scores



Using the score means, a histogram visually indicates the frequency and standard deviation of the scores from the sample. Figure 4.3 is a histogram that demonstrates a fairly normal distribution curve.

Figure 4.3

Histogram of Score Means



The mean provides the average of the scores. However, passing is established through a cut score rather than an average. A cut score is the minimum sum of rubric scores that represents a passing rate.[30] Table 4.5 displays the distribution of cut scores using the range recommended by SCALE (37 to 42) and adopted by PESB (minimum 35) in November 2013. This table demonstrates that the scores are unevenly weighted toward passing.

[30] SCALE describes the process for standard setting and cut score determinations in the Summary Report (SCALE, 2013, pp. 25-28).

Table 4.5

Cut Scores and Pass Rates

Teacher candidate pass rate n=58	
Cut score	Overall pass rate
35	93%
36	-
37	91%
38	90%
39	-
40	88%
41	83%
42	78%
43+	71%

Support for validity. The rubric scores, histogram, scatterplot and descriptive statistics for TPA means provide evidence of internal consistency with rubric scores.

Threats to validity. Score means reported by the histogram confirm SCALE findings from other 2013 pilots (SCALE, 2013, p. 17). However, the same scores reported using the cut score range with a minimum passing level of 35 reveal high passing rates**.** Given that this was the field test for the TPA, and that some components of the assessment were unfamiliar (namely AL but also ways of using assessment data), the high pass rate (.93) is unexpected. Evidence on pass rates (versus means) confirms this data. Passing was defined by cut scores, rather than means. The means for all of the candidates who "failed" the TPA were <2.0 out of 5.0. Candidates can earn a mean of >2.3 and pass the exam. A "Level 2 represents the knowledge and skills of a candidate who is *possibly* ready to teach" (TPA Handbook, p. 12, emphasis added).

Validity Evidence 5: Use of Rubrics with Inter-Rater Agreement

As mentioned above, one way in which internal consistency can be determined is through inter-rater reliability and agreement. High levels of inter-rater reliability refer to the consistency "between evaluators in the ordering or relative standing of performance ratings, regardless of the absolute value of each evaluator's rating" (Graham, Milanowski, & Miller, 2012, p. 5). Inter-rater agreement indicates the degree to which multiple evaluators use the same rubric scale to give the same score to identical evidence (p.5). For this validation study, inter-rater agreement is of more importance than inter-rater reliability because score consequences are based on cut scores and agreement on absolute levels of performance.

Pearson is contracted to facilitate scoring. One set of raw scores per rubric, per candidate, were reported to Sterner. Pearson did not report scores to candidates. In addition to Pearson scores, supervisors were asked to score candidate TPA. Using these two sets of rubric scores, rubric score differences and similarities were analyzed. Two common indices for measuring inter-rater agreement were calculated, including the percentage of absolute and adjacent agreement and Cohen's Kappa.

The percentage of absolute of agreement refers to how often raters agree on the exact level or score given to each rubric. This percentage is simple to calculate in this study because the number of raters is small. However, results are difficult to interpret because there is no calculation of the level to which "chance" or "random" may explain agreement and, because the TPA rubric has five levels and scores may fall across different levels, it does not distinguish between these different levels of disagreement. For this reason, adjacent agreement percentages, or the frequency with which scores occurred within +1/-1, +2/-2, +3/-3, will be shared. Cohen's Kappa reports how well scorers agreed. Kappa addresses the "chance" or "random" factors that may influence scores. Kappa is considered a better estimate of agreement when raters are reporting for different groups of candidates. Cohen's Kappa is, however, more difficult to interpret when the rubrics have many levels (five or more) and can be "misleadingly low if a large majority of ratings are at the highest or lowest levels" (Graham, Milanowski, & Miller, p. 8).

Findings. In total, 920 rubrics were scored by Pearson and the supervisor, for thirty candidates. Of these, 271 were different (.29). The majority (.79) of this difference occurred within one rubric level (+1/-1). Most (.52) of that difference happened when supervisors scored the candidate as +1, though some of the difference (.27) occurred when supervisors scored the candidates -1 from Pearson scorers. The total difference in rubric scores within two rubric levels (+2/-2) was 18% and within three rubric levels (+3/-3) was 3%.

The three rubrics with the greatest number of scorer difference include Rubric 3 (Planning), Rubric 11 (AL), and Rubric 13 (SV). Other rubrics that scored >30% scorer difference include Rubric 1 (Planning), Rubric 2b (Planning), Rubric 8 (Assessment), and Rubric 10 (AL). These also correlate with those rubrics with the most variance in score range (+3/-3) (see **Appendix O** for agreement indexes).

One question sometimes asked about inter-rater agreement is whether that agreement should be assessed at the standard (rubric) level, construct level (task, category, domain), or by overall scores. Construct agreement varied (.68 to .74). Most agreement occurred in Task 4 scores (.74). Task 4 has only one rubric. Least agreement occurred for Task 1 (.68), which has the highest number of rubrics (see **Appendix P** for agreement percentages).

Support for validity. Inter-rater agreement levels for absolute agreement between 75% and 90% are considered acceptable (Graham, Milanowski, & Miller, 2012). Two rubrics fall into this category (R12 and R14). Four rubrics miss the cut by one percentage point. The majority of the difference in rubric scores (.79) fall within one rubric scoring level (+1/-1). A review of the literature from teacher evaluation studies suggests most report a lower average in most studies of evaluations of teaching, around seventy percent (p.10). If using this average, the percentage of Absolute Agreement for TPA scores falls into an acceptable range (.71). In addition, half of the rubrics, and the rubric averages for Task 2, Task 3, Task 4, AL and SV, meet this benchmark level. Applying Cohen's Kappa and taking into account the "chance" that agreement would occur between the two scorers, the inter-rater agreement levels is .65, which meets the minimum level for consequential use. This calculation is well above the average reported in the literature (.54), suggesting that inter-rater agreement is stronger than the average reported in the studies on assessment practice in teaching observation evaluations (Graham, Milanowski, & Miller, 2012).

Threats to validity. The more consequential the examination results, the greater the burden for high reliability levels. Given that one would *not* expect to see scorer agreement outside of the one rubric level difference. Variance derived from scores that differ by +2/-2 or more levels is .21. Given that the scores will be used to determine candidate licensure, an absolute agreement in the mid-to-high .8 range would be more acceptable. Similarly, though well above the average from published studies, the Kappa should be closer to .8 than .6, which indicates that there may not be high levels of agreement in scorer rankings.

It is important to note that the data suggest that we could estimate that .29 of those rated could have received a different score had their TPA been scored by a different rater. This study indicates that scores could have differed for seventeen of the fifty-eight students (n=17/58). This is a fairly substantial number and it is difficult to evaluate whether high-stakes decisions should be based on these findings.

Inconsistent scorer interpretations of rubrics and TPA language may have produced inconsistent scores. Therefore, this field test may not consistently represent how the candidate actually performed. Many factors can affect inter-rater agreement. In fact, 100% agreement is not really preferable because of the high cost and time commitment that such agreement would require. However, some professional agreement is necessary, especially for high-stakes decisions made from scores. The following are some factors that can affect agreement: rater training, rater selection, accountability for accuracy in ratings, adequacy of rater compensation, rubric designs, rubric scales, pilot programs and redesigns, and technology use (Graham, Milanowski, & Miller, pp. 15-22). Validity Evidence 4 through 6 will examine some of these factors.

Validity Evidence 6: Scorer Quality, Training, and Recruitment

Before scorers can be trained, they need to be recruited based on professional qualifications that promote scoring reliability and the internal consistency of the assessment. The importance of rater-expertise cannot be underestimated because scoring requires knowledge of the construct within a specific subject area but also experience working with novice teachers. Scorers with this level of expertise can be considered highly reliable for training. However, scorers that lack experience in any one aspect required for scoring, such as working with novice teachers, evaluating other teachers, or teaching in the subject area, could impact the reliability of the scores. To determine rater-selection and rater-training, interview and survey data and document analysis of the Field Test Summary Report (SCALE, 2013) was examined for evidence of scorer quality.

Findings. Pearson solicited scorer interest through their website, emails, and a network of universities (WACTE, 2012). Some faculty across WA completed the application and interview process but were not selected as scorers (WACTE, 2012). At Sterner one faculty application was accepted. At some WA universities, no faculty were selected. Stanford's report indicates that Pearson hired 650 scorers to review 12,000 submissions during the field test (pp. 1-2). Half of the scorers came from academia and half were P-12 educators. Of the educator group, about half were NBCT (p. 15).

To solicit scorer interest from Sterner, a survey asked mentors, supervisors, and faculty whether they would score TPA. Scorer interest was <.22. Mentors (.27) report the highest interest, supervisors (.18) and faculty (.17) were least interested in scoring TPA. The number of interested faculty increased (.42) if TPA scoring was part of a contractual work load (rather than as compensated by Pearson). Faculty expressed concern that scorers were required to become Pearson employees (Faculty Interview, 5/10/12). Contractually, faculty are often obligated to get permission for outside employment such as this which could create a (modest) barrier for scorer recruitment.

Pearson Training. Pearson scorers complete a twenty hour online and interactive group training process (SCALE, 2013, p. 15) and then qualifying TPAs to be considered proficient (p. 15-16). Scorers were considered qualified if they scored nine out of fifteen rubrics correctly, with no scores that differed more than +1/-1. Each TPA requires between two and four hours to score. Scorers are asked to complete at least four TPA and are compensated 75 US dollars for each. Given the time intensive nature of the scoring process, some scorers opted not to participate. This was true of the Sterner rater who trained on "personal time" with the intent to better understand the TPA. Scorer interest dropped between spring 2012 and fall 2013. Pearson continues to work to recruit scorers via email, robo-calls, advertising, and the university network (personal communication, September 2013; December 2013; January 2014).

Sterner training. Sterner asked supervisors to score two TPA and provided a training stipend and meals during a two day group training session facilitated by a trained TPA scorer. Supervisors scored one practice TPA, using samples from a pilot in 2011. Scoring TPA, and then debriefing with candidates, was a contractual obligation.

Support for validity. The *Standards* advise that scorers clearly understand the domain(s) and subjects assessed (AERA, APA, NCME, 1999). The more subjective the judgment, the more training is needed. According to Graham, Milanowski, and Miller (2012), training sessions that are longer than five hours are more effective (p. 16). For high-stakes assessments, there is a benefit to training twenty-five or more hours (p. 16). Graham, Milanowski, and Miller also recommended that scorers prove a minimum level of agreement before scoring live assessments and confirm that scorers are participating in conversations around a “gold standard” judgment (pp. 15-16). Pearson’s training platform was developed to support scorer agreement and inter-rater reliability. The twenty hour training process, though on-line, has an interactive component that helps raters to compare their scores on a sample TPA to the gold standard and to qualify raters must meet a scorer threshold. The surveys of Sterner faculty, mentors, and supervisors suggests that there are fair numbers of interested scorers, at least for the short term, so long as compensation is reasonable or included in faculty contracts by replacing something else in their workload.

Threats to validity. Scorer thresholds qualified raters even if the score they gave six rubrics (.40) were scored incorrectly. This may explain the low rater-agreement levels (see VE 5). For the field test, candidate work was not necessarily double-scored. Candidates and programs were provided one set of scores and were not told whether the scores were from a single rater, a composite of multiple raters, or from the highest/lowest rating. The time intensive nature of the TPA scoring process, combined with the relatively low compensation offered and the requirement that scorers become Pearson employees, may limit scorer quality by reducing the interest of more experienced raters.

Validity Evidence 7: Issues that impact scores

In Phases 2 and 3 candidates were surveyed about challenges using the Handbook (instructions, prompts, rubrics). In addition, during the writing process, undergraduate candidates completed a survey after each writing day.[31] Finally, undergraduate candidates completed a survey and a draft version of their Task 1 segment.[32] Responses were coded and analyzed for patterns. In addition to the surveys, and to better understand the issues that may have impacted scores, the six case study participants were interviewed and those interviews were coded.

Findings. Candidates report that their TPA performance was impacted by construct irrelevant performance expectations (.59), or what is not typically understood as evidence of teaching readiness (for instance videography), the timing of the TPA (.31), the intense workload integrating the TPA into ST (.28), and trying to interpret and understand the handbook (.27). Feeling stressed, frustrated, and concerned about the workload requirements dominated candidate concerns during Phase 2. Issues with TPA timing, using the handbook, and new or unfamiliar performance expectations were the primary concerns during Phase 3. One candidate wrote that the program should, “Not emphasize the TPA so much! I wish there was more of an emphasis on student teaching” (TPA Completer Survey, 3/22/12).

Phase 2 and 3 Surveys. After submitting the TPA, candidates responded to an open-ended question, “Is there anything else you would like [Sterner] to know about your experience?” Sixty-one percent of the candidates reported a TPA related issue, even though TPA were complete. Several

[31] Undergraduate candidates were pulled from their placement sites for three half days of writing in campus labs.

[32] Recorded observations during the undergraduate writing days informed questions for case study interviews in Phase 3 and 4.

candidates demonstrate their ability to contextualize the importance of the assessment purpose (and its intent to measure readiness) while also identifying issues that may impact scores assessed to their performance, either on the TPA or on ST, more generally. Survey excerpts appear below.

- “The TPA, in my opinion, is a good formative assessment for teaching because it shows you where you can improve. However, it should never be the final say in a teacher's readiness. The teacher could turn it on for five lessons and then suck for the rest of the time. Be wary of this talk that the TPA will determine whether or not you become a teacher.”
- “It is hard to do the TPA when you are just starting your student teaching. At least for me, being in a classroom where I had very little say or control of what is going on, I had hard time being able to implement all the requirements of the TPA in a manner that I would like to. Getting student voice from students who are not use[d] to or understand what student voice is, was not very effective. If it was my classroom from the beginning I believe my TPA would indicate a little more of my abilities.”
- “The TPA handbook was confusing even after we had discussed it multiple times. When all of us were introduced to it together, it was overwhelming and even when we were broken out into our content areas it was still a little blurry.”
- “The TPA was a very frustrating process through the time of submission. The TPA did not help me develop as a teacher. My special education experience did... Not the TPA. The TPA also was not appropriate for kindergarten. It asked for significantly more metacognition than my students were able to do.”
- “The TPA has had a negative impact on my student teaching experience. Not only did I miss several days because of it, but also I am just now reaching full time teaching status because my TPA segment had to be taught late and my teachers did not want me to take over until after I had the TPA out of the way.”
- “It was miserable. I think I cried every night. Now I'm swamped in lesson planning [that] I couldn't do last week because I was too busy writing.”

- “It was one of the worst things I have had to do as far as school of education requirements. I don't think it is an accurate portrayal of our abilities as teachers and I changed so much of how I teach just to fit into the TPA mold (as far as assessments and academic language, etc.).”
- “I think the format required for Task 2 was unreasonable. Asking us to demonstrate our communicating the learning targets to the students, effective, engaging instruction, periodically referring to the learning targets, students having an opportunity to engage with each other, and students having an opportunity to reflect on their learning all in TWO clips of no more than fifteen minutes total is ridiculous. It's hard to fit all that into multiple parts of a 50 minute lesson, let alone two clips of 15 minutes.”
- “The TPA is designed to make teachers really think about their students-- what each student needs, to take backgrounds into account, to accommodate different learning styles. The TPA does the exact opposite of this. In addition the video segment was really difficult to capture-- the camera made my kids weird, and me weird. Getting what totals 15 minutes in only two continuous clips caused me to leave out a lot of great teaching moments that I captured. Also, the process of writing it-- of having so many different documents, that I couldn't touch and feel and organize for submission was stressful for me and I know that it made me do a worse job-- that is part of the way I learn. The TPA doesn't help you be a better teacher-- it just weeds out the people who don't want to be a teacher badly enough to jump through those ridiculous hoops.”

Writing Surveys. During the three writing days, undergraduates participated in surveys asking about their experiences completing the TPA. The majority of their questions were for Task 2: Instruction (Day 1: .45; Day 2: .29; Day 3: .08) and Task 3: Assessment (Day 1: .25; Day 2: .14; Day 3: .33). Candidates were asked to provide descriptions of their questions. Responses indicate that they struggled with construct-irrelevant performance related issues. On day one, a candidate wrote, “Why isn't my movie file loading” (Writing Day 1 Survey, 3/21/12)? Another wrote, “Oh boy, mainly about the size of the videos and formatting based on the criteria they gave us" (Writing Day 1, 3/21/12). Other responses demonstrate candidate difficulty in interpreting and understanding requirements from the handbook and the rubrics. On Day One, a candidate explained, "I had a question about one of the prompts in the task 3 simply because a lot of them were very similar and I wasn't sure if I was answering them the way I should have been" (Writing Day 1, 3/21/12). Another said, "I was confused by the way Task 2 describes the Learning Targets. Specifically, the relationship between task 2 and task 4" (Writing Day 1, 3/21/12). Other examples include, "I had trouble with prompt 4. I did not know how I was supposed to answer these questions" and "I was confused about Academic Language forms, the submission process and PDF format/word format" (Writing Day 1, 3/21/12).

Many of the candidate questions, although mostly about Tasks 2 and 3, involved the embedded categories of AL and SV. For instance, on Day Two one candidate wrote, "Figuring out Forms and Functions" as the question in Task 3 (3/22/12). Forms and Functions are one of the basic concepts of AL. Another candidate wrote, "I was wondering mostly about integrating the targets and what exactly to do if the video did not meet the exact TPA requirements" and another simply stated the difficulty was "'understanding the 'learning target' jargon" (3/22/12). Again, communicating learning targets is a central component of SV. These questions, and their timing in the writing phase of the TPA, indicate that the centralized concepts of both AL and SV were not sufficiently understood prior to the TPA nor uniformly practiced in classrooms where candidates taught and that candidates may not have had sufficient opportunity to learn these before completing the TPA.

Candidates were also asked about errors in handbooks. Candidates reported six handbook errors like science criteria in the social studies handbook, differences in the number of allowed video clips and submission requirements, and typos. Confusion about handbooks was an issue for some candidates even without these errors, but the errors further complicated candidate understanding of the requirements.

Task 1 Draft and Survey. Undergraduate candidate concerns after submitting the Task 1 draft included timing issues (.31), frustration that the TPA is overly complicated (.25), and time management (.19). Candidates in secondary placements appeared most disadvantaged by the due date of the TPA in the term. One candidate explains:

> It seems to just come far too soon in the semester, which I know probably can't be helped. Particularly in the secondary setting, I felt as if I had barely got my feet wet, barely learned names of the students, and then I was expected to have this huge and articulate unit designed for them. (3/2/12)

Frustration about complexity was often aimed at Task 1. For instance, one candidate wrote, “I am understanding why we need to do the TPA but am still unclear as to why it is SO EXTENSIVE. It's a little unnecessary” (3/2/12). Another candidate explained:

> The TPA, although [it] is helpful, has preoccupied time I could be concentrating on my student teaching and forced me to spend hours, after a long day of school, writing. It has also forced me to go out of my way to get video release forms and put a video camera in my students’ faces. This is taking precious time away from my day and adding more stress to my load. (3/2/12)

The Task 1 Survey asked candidates what Sterner could provide to help support work on the TPA. One candidate wrote, “Interpret the commentary questions! Some of them I have no idea what they are asking and it is really frustrating” (Task 1 Survey, 3/2/12). Despite program intervention, the same questions and concerns persist from Phase 2 through Phase 4 of the study.

Case Study Interviews. Coded case study interviews reveal confusion about expectations and performance requirements, about handbook repetition, and handbook issues that impact performance. The following case study excerpts further explain survey results.

Understanding the Expectations. Many candidate comments exposed confusion interpreting handbook prompts and rubrics. For instance,

- "It is really hard for me to know that I am writing the commentary and certain parts in a way that answers the question, the question being asked, because I am not completely sure what the question is that is being asked" (Jamie, Phase 3).
- " [If I meet standard] is one of my confusions because I see the rubrics and I know what I am submitting but when they actually take the rubric and do the grading, are they looking at my commentary, at my lesson plans, are they looking at it as a whole. . . . I have my lesson plans filled out [and] all of the requirements will be there but if they just go off of my commentary, I am afraid that they might misunderstand something that I am saying or that I might miss something that they are looking for, especially since their prompts are like the longest, most drawn out sentences in the world" [Jennifer, Phase 2].
- "I am still confused about what we were supposed to do" [Jason, Phase 3].
- "I wouldn't say 100% for any section that I know I am going to meet standard. I think I did everything right. . . . But reading those rubrics, oh, my goodness, I just, I just don't know, I mean even just trying to figure out, like, okay, so I am writing this question and it says I will be graded with these four rubrics and having to go back [in the handbook] and like find it [the rubrics] and then sometimes what it asks you to do in the overall outline of the TPA [the prompts] is different than what you find in the rubric. Like there is something they just slipped in there, in the rubric, that wasn't in the large write [prompts] so when I find those things it makes me nervous because I am like, oh, wow, I hope I caught all of those things" [Jennifer, Phase 3].
- "I think it would have been helpful to have really understood the commentaries before teaching my lessons. I don't feel like I understood what I was doing fully when I was teaching my lessons and when I was collecting work and I think I just would have liked to have had

everything laid out for me in a way that I understood so that I could go into it feeling more confident about it" [Jamie, Phase 3].

- "I think I struggled just with the way that some of the questions were worded. I just really struggled with trying to break down, break apart what they wanted from me. . . . There were times throughout writing the TPA tasks or doing my videoing where I was like, 'I wish I would have understood this better while I was doing my lessons.' So that was challenging for me feeling like I maybe didn't do as well as I could have because I didn't understand part of what I was being assessed on. I think I would say that was probably the biggest challenge just for me not fully understanding the expectations for me before going into my lessons" [Jamie, Phase 4].
- "But a lot of the candidates including myself, we were pretty comfortable with this whole idea of the TPA and all of the components. . . . But then when we opened up the TPA packets and really started to dive into them we realized in the fine print the TPA was looking for us to demonstrate very specific skills for our lessons. . . . So it is one of those things where we all know the components but then when we actually saw the exact requirements that they were looking for, for the individual subject area, it came as a little of a shock" [Jennifer, Lesson Debrief].

Repetition and confusion. Candidates articulated that they felt the prompts across the tasks to be repetitive. This led to more confusion about expectations:

- "I just feel like I am answering the same questions over and over and over and I feel like they want really specific answers but their wording of the questions are four times longer than they need to be. Very convoluted" [Jennifer, Lesson Debrief].
- "There were times that I was repeating the exact thing I said before and that was just kind of like 'okay, I have done this why do I have to say it again?'" [Jamie, Phase 3].

Issues with the Handbook and Rubrics. Some misperceptions were due to poorly written rubrics, instructions that were not "user-friendly," or rubric requirements that did not always align to prompts. In addition, there were discrepancies between the handbook directions and the on-line scoring submission requirements:

- "I guess I would say sometimes in the rubric I felt like a [level] four listed very clearly what it wanted, it wanted x, y, z but then when you read a [level] three, it is more vague. It is almost like they were mirroring what they thought a three would look like versus what a four would look like in their writing style. It was really bizarre so it wasn't even that I was like, 'Oh, I am going to try to go for a four and be super cool.' It was more like, I understand what a four wants me to do and I am not exactly sure what a three is asking me to do so. I am going to go right to the four. . . . It was very clear what a [level] two was. I can see what they are missing, that is why you would get a two. I can see why you would get a four because you add these three things, very clear, but a three was like, nah, kind of wishy-washy. I couldn't write to that. So I went just for the fours" [Jennifer, Phase 3].
- "Some of the prompts I had to go to [my professor] because I did not understand what they were asking. [Like for] supporting student learning question D, "describe common developmental approximations and misunderstandings within your literacy content," I just was a little bit confused. So maybe just I mean just maybe more clarification on some of the questions and what they mean. Be able to put them in my terms" [Jamie, Phase 3].
- "When I started to upload Task 3 and I was like, 'Oh, the students had to reflect on these three things?' They gave you like three prompts that the kids have to also reflect upon. That took me completely by surprise. So, I had to go back in and pull those three kids aside and be like, 'I am really sorry, can you do this extra work for me. I will give you bonus points.' I would not have been able to do that if I would have waited until spring break [to do the write up]. I would just be out – and I don't know what they would entail if I was just missing those reflections" [Jennifer, Phase 3].

- "There was a lot of translation between the criteria on the rubric and the prompts. Since I just read the prompts [when planning] and then when I went to go write my commentary I went to the rubrics [and] I was like 'Oh, well, I have to go back and re-splice my video' … I didn't know that until I looked at the rubrics. It wasn't in the prompt!" [Jennifer, Phase 3].
- "You have to kind of look and dig a little bit to find out [which rubric goes with the prompt]. I know that that is our responsibility to read but I think that even maybe just telling us read and find out . . .what rubrics will be used to score that so you know what you have to get" [Jane, Phase 3].
- The handbook indicates that, "We wouldn't have to submit a separate feedback, if the feedback was written on the work, but when we uploaded it you have to submit a page for feedback. So I just submitted the document that said that, you know, like, that they didn't need feedback written on it. But just little things like that were really frustrating because you know in the instructions it says you don't need to submit this but then it won't let me submit the TPA without that document. …And I know a lot of my cohort had problems with the size limits and uploading process and when it is something so high stakes that it makes a difference for whether you get your certificate you are really relying on technology" [Jackie, Phase 4].

Support for validity. Many candidate responses indicate an understanding of the purpose of the TPA and the need to measure candidate performance to determine readiness to teach. Not all candidates reported concerns or issues that they felt would impact scores.

Threats to validity. The ABV method predicts that scoring and the operationalization of a performance assessment is likely to be a weaker link in the chain of inferences. Survey and interview data exposes several issues that impact score performance. Analyzing candidate perceptions and experiences indicates that there were common concerns throughout the four phases of the study. These include workload issues that put strains on ST, the timing of the TPA in the term, unfamiliar requirements, and confusion about TPA expectations. Candidates suggest that the handbook lacks student-friendly language appropriate for novice teachers, creating confusion about performance expectations. The problem of repetitive prompts, errors, and construct irrelevant requirements (especially in Tasks 2 and 3) likely led to scores that may not have been accurate about candidates' readiness to teach.

Validity Evidence 8: Use of Rubrics with Handbook Analysis

Rubrics that require a great deal of scorer professional judgment lead to poor inter-rater agreement and scoring confusion, especially confusion about the expectations of the assessment. In order to determine whether scoring rubrics were confusing in the scoring process, or required extensive professional judgment, TPA rubrics were analyzed using the Secondary History/Social Studies Handbook. Once analyzed, the results were compared to both Elementary Mathematics and Secondary Science, to confirm that the scoring issues appeared across all scoring rubrics. In particular, the focus was on the difference between not meeting expectation (Level 2) and meeting expectation (Level 3). Rubrics were analyzed for generic terms that may be difficult to define or quantify in practice or for expectations that were unclear when applied to teaching, and novice teaching specifically.

Findings. Nine of the fifteen (.60) rubrics contain potentially unclear vocabulary. For instance, rubric 1 scores determine how well the planned lessons are aligned to standards. Level 2, or not meeting standard, is described by "loose" alignment. The terms "loosely" and "inconsistently" are difficult to quantify and require professional judgment. The extent to which lesson standards must be aligned is unclear. It is possible, then, that one rater may determine that a candidate's submission met standard while another may not, based on different notions of "inconsistently." Raters are told to look at the "preponderance of evidence" to make a judgment. However, this, too, requires professional judgment that may differ for the same candidate's work. Similar language judgment issues occur in rubrics 3, 4, 6, 7, 8 and 11 and apply terms like "deepen," "in a general way," or "attempts to" as the qualifiers between meeting and not meeting standard. These terms are difficult to measure and require professional judgment.

Other rubric concerns include combining differing skills and confusing penalties. For instance, rubric 1 requires candidates not just to align their lessons with standards but with a central focus that addresses the assigned subject area performance (i.e., document analysis in social studies, data analysis in science). It is unclear how to derive a score if a candidate aligns the standards, objectives, and tasks well and consistently throughout the learning segment (earning a level 3), but the learning segment does not address the performance focus required by the subject endorsement (earning a level 2). Is the skill assessed in this rubric *alignment* or content *pedagogy*? Similar combinations of skills in one rubric also occurs in rubric 2. Finally, rubric 14 penalizes a candidate who does not have the ability (permission) to create teaching materials. Meeting expectations (level 3) examines whether the candidate "creates one or more tools." This language penalizes students for adopting a tool or strategy when they may have selected a better instrument than one they can create. Similarly, some candidates do not have the freedom to create tools (see **Appendix Q** for language and judgment issues).

Several researchers have studied the professional growth of teachers in the profession (Darling-Hammond, Chung Wei, Althea, Richardson, & Orphanos, 2009; Walling & Lewis, 2000; Schön, 1983; Grenfell, 1998; Greenberg, Pomerance, & Walsh, 2011). Most agree that there are predictable patterns of professional growth between novice and experienced practioners. In addition to verbiage and professional judgment issues, six rubrics (.40) include requirements that can be questioned based on expectations that may not be appropriate for novice teachers (Stotsky, 2006; Smith & Lev Ari, 2005)(see **Appendix R**).

Support for validity. The more general the criteria, the more judgment involved, both for the candidate to interpret those criteria and for the rater to use those criteria to score. Specific rubrics and better trained and experienced raters reduce bias and subjectivity, but Kane (1992) notes that "these problems cannot be completely eliminated" (p. 172). Allowing for the level of variation in the contexts with which two otherwise equal candidates will complete, the TPA requires that the rubric criteria used to evaluate each candidate be general enough to account for those variations. The very need for this level of standardization and generalizability will create "substantial errors of measurement" (p. 172).

Six rubrics (.40) are free of issues that may require scorer professional judgment. Of the remaining rubrics, several can be corrected by a clearer description of the performance expectations.

Threats to validity. Several rubrics contain issues that require scorer professional judgment which could lead to poor inter-rater agreement, scoring confusion, or, as seen above, candidate confusion about expectations. These could lead to confusion and questionable decisions made from scores. Nine rubrics (.60) contain these issues. In addition, six (.40) rubrics include performance expectations that could be considered questionable for novice teachers, based on typical developmental issues during ST. Some of these reflect the classic developmental issues for novice teachers. Many of these are the result of lack of experience. Careful, intentional, strong preparation of teachers places them in gradually progressing teaching experiences which culminate in a full-time co-teaching internship (ST). This is the type of program Sterner has developed. ST is meant to provide an initial full-time teaching experience and to offer candidates an opportunity to practice teaching in a carefully managed and mentored setting. It is questionable to hold novice teachers, in their first teaching setting, to the standards and expectations of more experienced professionals, as the analysis of rubrics suggests. One candidate wrote in her survey, "[Sterner] did a great job preparing for the TPA. It is just tough because I am not a "good" teacher yet. 5 years would be a good time to take this assessment" (TPA Completer, 3/22/12).

Inference 3: Generalization

Research Questions

4. Are the score levels achieved on the rubrics a true representation of a candidate's performance? In other words, are the scores a candidate earned consistent and generalizable with other samples of that candidate's teaching performance?
5. Does poor performance on the TPA imply a lack of adequate mastery of the construct?

6. How generalizable are the criteria, rubrics, procedures, and scores derived from the TPA across different candidates and handbooks?
7. Does TPA proficiency depend upon factors beyond the candidate's control? How generalizable are the criteria, rubrics, procedures, and scores derived from the TPA across testing sites, placements and placement length and programs?

Validity Evidence 9: Multitrait-Multimethod Matrix (MTMM): TPA Scores, Mentor and Supervisor Evaluations

One way in which generalizability can be determined is through consistency between multiple methods that score the same traits for the same candidates. Two methods were compared: The Teacher Candidate Final Progress Report[33] (PR) and TPA Scores. PRs were completed by supervisors and mentors as a summative evaluation and recommendation of candidate readiness. For this method, raters scored candidates on twenty-seven separate traits aligned to Standard V. Twenty-one of the twenty-seven PR traits aligned to TPA rubrics. Inter-rater agreement was calculated to determine the degree to which multiple evaluators gave the same score to identical evidence (Graham, Milanowski, & Miller, 2012, p. 5). Because eight candidates had two mentors, differences in rater agreement were considered separately. These scores were compared to the set of scores provided by Pearson. Following a similar process described above (see VE 5), the common indexes for measuring inter-rater agreement were calculated including the percentage of absolute and adjacent agreement and Cohen's Kappa. In addition to providing the agreement correlations, a three-way Multi-Trait, Multi-Method (MTMM) was developed to compare scores. Finally, a summary of candidate score differences was calculated, focusing on where high-stakes results differed.

[33] Sterner revised the PR to align with PESB Standard V criteria in 2008. The PR was aligned to the TPA in 2012, but not significantly modified from its 2008 version.

Findings. Thirty teacher candidate samples were scored by both methods. In total, 1512 traits were evaluated on the PR using a 5-point Likert scale. Of these, 247 differed (.16). Nearly all (.88) of this difference was within one rubric level (+1/-1). Some (.21) difference is explained by supervisors scoring +1. Most (.67) difference occurred when mentors scored the candidates -1. The total difference in rubric scores within two rubric levels (+2/-2) was 10% and within three rubric levels (+3/-3) was 1%.[34]

Scorer differences were greatest for Standard 3: Knowledge of Teaching and Instructional skills.[35] No standards had more than >28% agreement difference (see **Appendix S** for PR indexes and **Appendix T** for agreement percentages). Based on this data, the PR is a consistent scoring tool with higher levels of scorer agreement than TPA scores (see VE 5).

However, score scatterplots indicate that the PR results demonstrate a ceiling effect with a bunching of scores at the upper level. Studies have indicated that evaluations of teachers can be subjective[36] and the ceiling effect on the PR may be a result of such subjectivity. Results with a ceiling effect may also be the result of inherent flaws in the instrument design, for instance the Likert scale may not sufficiently distinguish between upper levels of the scale. It is important to note

[34] Much of these differences are accounted for by one outlier in the data. When the outlier data is removed, there are no +3/-3 differences, and +2/-2 differences fall to 2%.

[35] Specifically, Standards 3.5, 3.7, 3.8 and 3.9. In addition, Standard 2.1 “Accurately assesses student needs (emotional/academic) and provides feedback” and Standard 2.6 “Communicates clearly and professionally with families” ranked higher, although 2.6 was ultimately not used because it did not align to the TPA requirements.

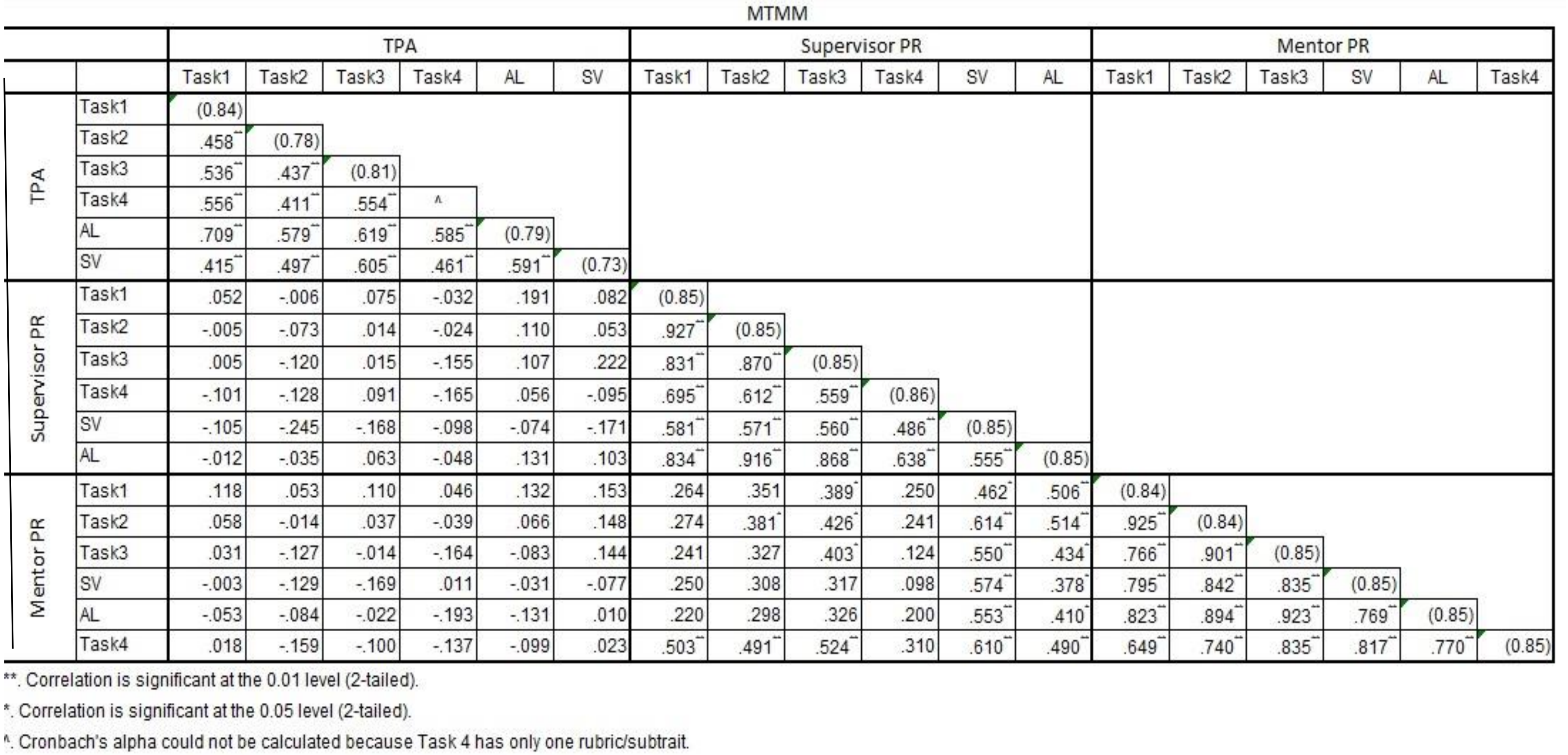
[36] In fact, researchers have identified a “good subjective” and “bad subjective” in the evaluation of practicing teachers (Rockoff & Speroni, 2011). Others have found that principal evaluation distinguishes between poor teacher quality and excellent teacher quality, but rarely distinguish between teachers who fall in the middle of the distribution (Jacob & Lefgren, 2005).

that only candidates that have passed university benchmarks are allowed to student teach. For these reasons, higher scores of teacher readiness are expected because candidates not ready to teach should have been removed from the sample prior to student teaching. While still useful as a threshold test (for licensure decisions, for instance), PR scores do not allow for a ranking of top performers. Like the TPA results, the purpose of the PR is to provide the university with data to support a recommendation for licensure (a yes or no question). Therefore, its purpose as a threshold test is supported. However, the ceiling effect on the PR limits the variability gathered from any one trait and may reduce the power of statistics on correlations between any two variables.

MTMM. An item-correlation matrix provides the correlations between aligned traits in the PR and TPA. An MTMM was produced using these correlations and Cronbach's α. MTMM analysis indicates that reliability for each method is strong (.73-.85) with good internal consistency for each measure. In addition, heterotrait-heteromethod correlations are low demonstrating that the traits do not correlate highly with each other between methods. However, the heterotrait-monomethod correlations are high, suggesting that the traits evaluated by both methods are highly correlated. This could influence Cronbach's α and method reliability. In other words, high reliability indicators could be the result of item similarity rather than internal consistency. The purpose of an MTMM is to demonstrate the degree to which two different methods, intended to measure the same traits, actually do so. Significantly, for the purposes of this study, the MTMM indicates that these measures do not correlate with each other. Results clearly indicate a method-effect. Surprisingly, there appears to be no discernable pattern between summative PR scores of teacher readiness and summative scores of teacher readiness on the TPA.

Table 4.6

Multi-Trait, Multi-Method Analysis of TPA Scores and University Progress Reports



MTMM

		TPA						Supervisor PR						Mentor PR					
		Task1	Task2	Task3	Task4	AL	SV	Task1	Task2	Task3	Task4	SV	AL	Task1	Task2	Task3	SV	AL	Task4
TPA	Task1	(0.84)																	
	Task2	.458**	(0.78)																
	Task3	.536**	.437**	(0.81)															
	Task4	.556**	.411**	.554**	^														
	AL	.709**	.579**	.619**	.585**	(0.79)													
	SV	.415**	.497**	.605**	.461**	.591**	(0.73)												
Supervisor PR	Task1	.052	-.006	.075	-.032	.191	.082	(0.85)											
	Task2	-.005	-.073	.014	-.024	.110	.053	.927**	(0.85)										
	Task3	.005	-.120	.015	-.155	.107	.222	.831**	.870**	(0.85)									
	Task4	-.101	-.128	.091	-.165	.056	-.095	.695**	.612**	.559**	(0.86)								
	SV	-.105	-.245	-.168	-.098	-.074	-.171	.581**	.571**	.560**	.486**	(0.85)							
	AL	-.012	-.035	.063	-.048	.131	.103	.834**	.916**	.868**	.638**	.555**	(0.85)						
Mentor PR	Task1	.118	.053	.110	.046	.132	.153	.264	.351	.389*	.250	.462*	.506**	(0.84)					
	Task2	.058	-.014	.037	-.039	.066	.148	.274	.381*	.426*	.241	.614**	.514**	.925**	(0.84)				
	Task3	.031	-.127	-.014	-.164	-.083	.144	.241	.327	.403*	.124	.550**	.434*	.766**	.901**	(0.85)			
	SV	-.003	-.129	-.169	.011	-.031	-.077	.250	.308	.317	.098	.574**	.378*	.795**	.842**	.835**	(0.85)		
	AL	-.053	-.084	-.022	-.193	-.131	.010	.220	.298	.326	.200	.553**	.410*	.823**	.894**	.923**	.769**	(0.85)	
	Task4	.018	-.159	-.100	-.137	-.099	.023	.503**	.491**	.524**	.310	.610**	.490**	.649**	.740**	.835**	.817**	.770**	(0.85)

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

^. Cronbach's alpha could not be calculated because Task 4 has only one rubric/subtrait.

Pass Rate Comparisons. Finally, using TPA cut scores establishing pass rates (35 points), TPA pass rates were compared to PR pass rates. Results indicate that four students who passed PR would have failed the TPA in this field test. Three of those candidates were recommended to pass ST. Table 4.7 lists these four candidates' TPA subject, TPA score, PR score, and recommendation to pass ST.

Table 4.7

TPA Failures Compared to PR/Student Teaching Recommendations

	TPA Subject Area	TPA Score	Score on PR	Pass Student Teaching?
1	Elementary literacy	29	89%	Yes
2	Elementary literacy	26	39%	No
3	Secondary mathematics	31	69%	Yes
4	Elementary literacy	34	97%	Yes

The three highest TPA scores were compared to PR scores, TPA subject area, and recommendations to pass ST in Table 4.8. A perfect score on the TPA is 80 points. The results indicate that one candidate passed the TPA with high scores but did not receive a recommendation to pass ST.

Table 4.8

TPA Top Scores Compared to PR/Student Teaching Recommendations

	TPA Subject Area	Score on TPA	Score on PR	Pass Student Teaching?
1	Secondary social studies	72	79%	Yes
2	Secondary science	61	61%	No
3	Elementary mathematics	61	100%	Yes

Finally, Table 4.9 identifies TPA and PR scores for candidates who struggled to meet expectations during ST or required intervention from the supervisor or university faculty during the term.[37] Of these nine candidates, two earned the top three highest scores on the TPA and one failed the TPA.

Table 4.9[38] [39]

TPA Scores Compared to Student Teaching Support Intervention Needed

	Score on TPA	Score on PR	Pass Student Teaching?
1	49	66%	Yes
2	72	78%	Yes
3	41	76%	Yes
4	57	93%	Yes
5	56	98%	Yes
6	40	98%	Yes
7	61	61%	No
8	47	73%	Yes
9	26	39%	No

Supports for Validity. Errors of Measurement (EoM) are not necessarily errors in the assessment instrument but they influence the data and how it might be interpreted. The data from

[37] Intervention was determined through a midterm progress report, or by interviews with supervisors or faculty. Intervention can occur in many forms, from a single discussion reminding mentors and candidates of university expectations, to a formal contract of required behaviors. Not all interventions are due to candidate performance. Occasionally, a traumatic life-event, conflict with course requirements, or differences in expectations of mentor support can prompt an intervention. The study investigator facilitated interventions for undergraduate candidates during data collection.

[38] Note that candidates 1 and 2 in Table 4.8 also appear in Table 4.9 as candidates 2 and 7.

[39] Note that candidate 2 on Table 4.7 is candidate 9 in Table 4.9.

the MTMM may reflect random errors based on statistical fluctuations that exist when measured values are inconsistent and these cannot be predicted. One way EoM are minimized is through standardization (see Chapter two). However, greater standardization in performance assessment are problematic because it requires increased interpretation by candidates and scorers.

Threats to Validity. Where overall candidate performance is concerned, PR and TPA scores agreed in only one in four cases (.25) for failures, two in three cases (.67) for mastery, and two in nine (.22) marginal cases. In addition, a significant, systematic method-effect was found when analyzing the data using MTMM. Method-effect indicates that score differences are the result of the method of measurement, rather than participant performance. One would expect to see higher levels of correlation between two summative instruments of teacher readiness given to the same population in the same term. Because the analysis of PR data suggests a ceiling effect, and this was a field test, or pilot, for the TPA, it is difficult to be certain whether the method effect belongs to one or both measures.

Generally, score levels on the TPA represent candidate performance. However, data indicates that TPA scores could have led to one candidate (.02) receiving licensure who had not yet established readiness and three candidates (.05) to be denied licensure when their ST performance indicated a readiness to teach. In addition, for the nine candidates who required intervention, TPA scores should have reflected candidate difficulty to meet expectations; however, only one of nine of these candidates' TPA scores suggests potential readiness issues. It is not clear, therefore, from the MTMM and pass rate comparisons that score levels consistently represent candidate performance or whether poor performance on the TPA is a reliable indicator of lack of teaching readiness.

Validity Evidence 10: Mentor Survey on Candidate Performance at TPA Submission

Mentors were surveyed after candidates submitted TPA to correlate performance on the TPA to current performance in ST. Responses were coded and analyzed for patterns.

Findings. Mentors reported that candidates were meeting expectation at the midterm point in the ST term (when TPA were submitted). Twenty-four mentors (.32) responded. No (.00) mentors report their candidate "doing harm" or "struggling" in their classroom practice, four (.17) indicated candidate performance was emergent, two (.08) described their candidate as meeting expectations, fifteen (.63) of the mentors surveyed labeled candidate performance as successful, and three (.13) indicated the candidate was surpassing expectations.

Mentors also described experiences with the TPA. Many (.41) mentors felt negatively about their experience with the TPA, others (.37) were neutral, and some (.25) mentors responded favorably about the formative experience they observed for candidates. Almost one in four (.22) responded that TPA requirements were a challenge. One mentor indicated that the TPA was disruptive, "While [candidate] is phenomenal, having the video camera in the room for five days was a little distracting" (3/30/12). Similarly, this mentor questioned the value of the TPA requirements writing, "It has kept her focused, but not necessarily on what we need to do in class, or what she needs to do to practice teaching. Kind of the tail wagging the dog" (3/30/12). Like candidates, mentors found the requirements of the TPA to be something "outside of" the ST experience, stating, "It took [time] out of our regular schedule to complete the requirements of the TPA" (3/30/12).

When asked about candidate professional weaknesses at the midterm point, half (.50) of the mentors responded that the most important growth needs remained classroom management and additional planning, including a better understanding of the pacing of lessons, scope and sequence of grade level instruction, and the need to plan using units (rather than stand-alone lessons) (.45). Other mentors (.16) indicated that candidates needed to focus on communication and rapport, mostly with parents and student families. A few mentors (.12) wanted the candidate to relax. One mentor wrote that the issue was, "experience, the solution is simply experience" (3/30/12).

Every mentor (100%) responded that candidates required support to complete the TPA. Mentors were asked to indicate what support was needed. Most mentors (.92) indicated that candidates need support understanding the classroom context and student needs, a fundamental prerequisite for success on the TPA. The majority (.83) indicated that mentors need to review lesson plans and provide logistical support; or, be willing to adjust their classroom routines or style to meet TPA requirements (.79). Finally, many mentors (.67) said that they had to adjust or change the curriculum (see **Appendix U**).

Supports to Validity. Mentor reports of candidate ability at the midterm point correspond with the broader pattern of TPA scores.

Threats to Validity. One of the questions of this study asked whether TPA success depended upon factors beyond the candidate's control, which would violate assumptions of systematic and procedural fairness. Most mentors needed to alter classrooms routines, styles, and curriculum in order for candidates to complete the TPA requirements. Candidate placement is often determined by district and principal willingness (see Chapter two) and candidate endorsement requirements for placements (specifically in secondary education). Placement issues are a perennial concern in teacher training (Greenberg, Pomerance, & Walsh, 2011). Mentor willingness to change is outside of the control of the candidate and university. However, performances that meet TPA standards appear to depend upon mentor willingness to support candidates' TPA work. Survey results indicate that most mentors question the generalizability of the TPA data.

Validity Evidence 11: Case Study: Candidate Reflection at TPA Submission

Case study candidates were questioned about criteria, rubrics, procedures, and potential scores derived from the TPA, across different candidates and handbooks. Candidates ranked their confidence in having met rubric standards and to identify areas in which they struggled or felt concerned about their performance. Finally, candidates described their placement site, teaching conditions, and unique factors that they felt may have helped or restricted their performance on the TPA. Interviews were coded to find patterns.

Findings. ST placements differed. Some candidates had multiple mentors and others had only one. The teaching setting differed dramatically between urban and rural, high and low poverty, and numbers of students with special needs (see Chapter three). Interviews reveal that candidates perceived the following factors to have influenced candidate TPA performance: relationship with mentor(s) (100%), ST context and environment (.83), access and comfort with videography (.83), and the complexity of the assessment requirements (.83).

Relationship with Mentor(s). Each of the six case study candidates (100%) indicated that their relationship with their mentor affected their experience with the TPA. Jennifer explained that her relationship with her mentors had a direct impact on her TPA options. She describes her mentors as "very different." For instance, "[Mentor A] is a younger teacher, very global and an excellent mentor." However, Mentor B "is nearing retirement. Kind of burnt out. Very old school in his approach. Lots of like worksheets and that is about it. … He is not big into collaborating." Jennifer found that Mentor B "decides what he is going to do when he walks in the door which makes it kind of hard for things like the TPA." Specifically, she was concerned about working with Mentor B to plan Task 1, "When he doesn't know what his plan is for tomorrow, how in the world can he accommodate like any sort of planning on your part that is a month away?" (Phase 1). Jennifer discusses how the difference in teaching styles influenced her decision to complete a social studies TPA (with Mentor A) instead of biology, her first choice:

> I was going to do it in biology. When I originally asked [Mentor B] the question, he told me that my unit, which is evolution, would be starting like a week ago but because he doesn't plan or look at the calendar or anything I found out last minute that I will not be teaching in there until March which is past the TPA due date. (Phase 1)

Jennifer felt that Mentor A gave her, "free reign to do whatever I want to do which is super cool. I was basically told, 'six weeks do World War II and do your TPA whenever you want to'" (Phase 1).

Like Jennifer, Jill stated of her Mentor B, "I don't even think he is aware of what I am doing for the TPA" (Phase 2). Jill described her introduction to the co-teaching experience as, "I was alone in the classroom and I was not expecting that" (Phase 2). She reported not knowing what to expect: "every time that I came in to teach a lesson he had done something differently. Like the second time he had like taken half of the students out of the classroom and the third time he had rearranged the desks" (Phase 2). She was constantly "trying to adjust my lesson plans and adjust what I was teaching. One time he told the students before I had gotten there that I was only going to teach for half [the lesson] …but he didn't tell me" (Phase 2). Jill felt she could not complete her TPA in this setting and switched TPA subjects from English to drama right before the taught segment. Unlike Jennifer, however, Jill felt her second mentor was unsupportive and a critic of the TPA. Jill described this in her final interview, "the biggest thing has just been the impact that it [TPA] has had on my perception of student teaching and my mentors' perception of my student teaching." Jill reported that:

> It has been a huge stress issue with my mentor teachers; they have also seen it as a huge burden on the student teaching experience. They have hated having to deal with it because of the modifying we have to do to their lessons and their classroom to make it fit into the TPA. (Phase 4)

This may have influenced the level of support her mentor was willing to provide. In the planning phase she explains that "I sent," my lessons to Mentor A, "but he never actually got a chance to sit down and talk about it. And he doesn't read my script" (Phase 2).

Conversely, Jason and Jane both indicated that their relationship with their mentors was a key reason the TPA was a positive experience. Jason described his relationship with his mentor as, "We really get along. We're really compatible, and co-teaching is just going to be a blast" (Phase 1). Unlike Jennifer who developed her own units, or Jill who was not sure if her mentor knew she was completing a TPA, Jason and his mentor met "on a weekend and [we] had all of the math curriculum out and we are kind of figuring out 'Okay here is where we are in math right now'." Together they "went through the curriculum, picked three or four things, like important chunks for that unit, sat down, [and] typed up all of the lessons that day" (Phase 2).

Jennifer and Jill created their own plans for the taught segment with varying degrees of mentor input. Jason's TPA content, however, was based on "a math curriculum in the class that we follow to a 'T'" and that all math is taught "after lunch, which is awful scheduling, but we do math stuff in the afternoon from about 1:00 to 3:00. So it will be at the end of the day. And all days are the same" (Phase 1). Jason and his teacher were model examples of the co-teaching strategy advocated by Sterner (Picanco, Darragh, Tully, & Henning, 2011). Jason mentioned several times throughout the term that this made a significant difference in his feeling of confidence and ability to understand the complexities of teaching. For Jason, one of the best experiences of ST was,

> Collaboration time with my mentor teacher. We did that a lot. I mean really every time we had a free moment we were talking about the last lesson, or what is coming up next, where our students are at, a funny moment that happened that day. . . . Like we both, I think, both grew off of each other. (Phase 4)

Like Jason, Jane had a positive co-teaching experience. Jane was partnered with a teacher who also taught at Sterner. She described her mentor as, "I've known her for about a year now and I had a placement in her classroom in the fall and she's amazing and just very supportive and she has a ton of resources" (Phase 1). When planning her lessons, like Jason, she had a lot of "hands on support" from her mentor. She recognized that she was at an advantage because her mentor, "is an adjunct for [Sterner], because she does methods [courses], and so she has seen the TPA and I think she's probably more familiar with it than some of the other teachers would be" (Phase 1).

Jackie and Jamie were in elementary classrooms but had placements with multiple mentors. For instance, Jackie's mentors shared a first grade classroom and each worked a partial week. Jackie described the influence of her mentors on her TPA in this way, Mentor A "went to the mentor teacher orientation and that was so helpful because she . . . was really able to kind of calm me down. . . . She is the one who suggested the segment that I ended up doing." Mentor B "is the one who helped me with the videotaping, and pulling the kids out. . . . They were really helpful" (Phase 2). Jamie, however, had some difficulties with the differing styles of her two mentors. She noted that while writing up her TPA her “concern so far is just the situation with the two mentor teachers” (Phase 3). She described the problem in this way “[Mentor A] is a little bit different than my other mentor teacher in that she has had student teachers before, so she was very, very willing to hand over everything and say ‘here it is yours, you plan, you do this’.” However, “my other mentor teacher, who I am with right now, has been a little bit more hesitant in the fact that she is very protective of her students and has a hard time letting things go” (Phase 3).

Differences in mentors caused two of the six (.33) candidates to change TPA subjects, three of the six (.50) to feel supported in their work with the TPA and one (.17) to feel like she was not allowed to assume solo teaching responsibility. In fact, several students in both cohorts expressed concern that the TPA caused their mentors to wait to allow them to begin the solo teaching experience, several weeks beyond the university recommendation. The TPA caused some candidates to postpone full-time teaching until after the TPA was submitted (at least nine weeks into the term). Conversely, those candidates that did take on a full-time teaching load, who experienced a TPA advantage because they better understood their students and learning context, were also at greater risk for not leaving enough time to complete the writing. If placed too late, candidates did not have enough time to complete a revision or resubmission during ST.

If the TPA due date was too early in the term candidates did not have had enough time to adjust to their role as teacher and to understand their students. Sterner asked candidates to submit their TPA at exactly the mid-point of the term. This caused students to see the TPA as the primary reason they were prevented from practicing teaching because many supervisors and mentors opted to have candidates wait before taking over a solo teaching role until the TPA was submitted. The TPA was meant to be a snapshot of teaching performance integrated within ST. However, because of its significance as a high-stakes test,[40] it was seen as disconnected and a “hoop” to jump.

In the spring term, lack of solo teaching opportunities were made more problematic because mentors were preparing their students for annual mandatory state tests. Student test scores are published and used by districts to evaluate teacher performance and to secure federal grants. These tests occurred just weeks after the TPA due date and mentors were less likely to let candidates assume solo teaching responsibility just as students were preparing for tests. As a result, many candidates expressed frustration that they were less prepared, overall, and wished they had more than two or three weeks of solo teaching. In her final interview, Jill talks about this:

> I need more practice. I think that the Ed[ucation] department here has done a really good job preparing us, but there just wasn’t enough time. . . . It has been kind of rushed. . . . I think that I had the skills to student teach but I don’t think I have the skills to be a full time teacher, which is one of the reasons that I am pursing substitute teaching and masters programs and one of the reasons that I am not looking at being a full-time teacher. Because I don’t think that I am ready.” (Phase 4)

Jill passed both ST and the TPA. In fact her TPA score was well above the TPA cut for passing (she earned a 56, 35 is required to pass). However, like others in her cohort, she felt unprepared for the reality of teaching. This result is not surprising because feeling ready is not the same as being ready nor something test proficiency scores can provide. You wouldn’t necessarily expect them to always correlate. It is also not surprising that mentor support, guidance, and involvement during ST impacted the experience of all six case study candidates. Similarly, it appears to have a direct impact on TPA performance.

[40] The TPA was not consequential during the field test. Survey and interview data suggest that this made little difference to candidates in Phases 1 through 3. Some data suggests that when candidates reflected back on ST at the end of the term, their views changed (see VE 13).

Classroom Context. A second factor that five of the six (.83) candidates identified were significant differences in the classroom environment and student populations. Jason described his classroom context as, “We've got a few IEPs and 504's and an Asperger’s child. And we have a gifted student in the TESSERA program, and a couple other students who should be in the TESSERA Program” (Phase 1). Jason’s placement was in a school with a high population of students in academic risk, some of whom change schools several times in a year. Jason explained “right before I was about to teach it [TPA] we got a brand new student.” And that, “I don’t think I tried as hard as I could have to address” the new student’s needs. “He had a lot of baggage and so I wasn’t prepared to accommodate for him right away and so he was in my lesson and the poor kid, he was just clueless . . . just on another planet” (Phase 3). Jane’s teaching context was quite different from Jason’s. She states, “I don’t think that there are any IEP’s or 504 plans or anything. I’m pretty sure that none of them are English Language Learners. But they are typical seventh graders” (Phase 1). Jamie teaches in a classroom with “one student on an IEP and he was in, he was gone during my lessons, so I don’t have any work from him” (Phase 3).

Classroom context was predictably different because candidates were placed across three districts in schools quite diverse from each other (see Chapter three). Classroom diversity is also a reality for practicing teachers. However, candidates are placed by districts and schools based on willingness and availability of the district. Candidates have almost no voice in the placements they are given. For this reason, classroom demographics are beyond candidate control. Candidates whose students better meet the expectations of a “typical classroom,” as defined by the TPA, are likely to find the TPA to be an easier task than students whose classroom settings are populated by students who do not fit that mold.

Videography: The Hidden Curriculum of TPA. Another factor that was perceived to influence a candidate’s performance was access to and comfort with technology, specifically videography. Five of the six (.83) candidates expressed frustration about video collection and submission. One candidate hired a videographer who set up $2000 worth of equipment, including microphones and lights, to capture her video segment (Jennifer, Phase 4). She reported that her students were fascinated with the camera and "they all participated in the lesson, more than they usually do. . . . I heard from people who I never hear from. And, they all behaved on camera." She reflected that, "They were angels. It was so funny, they were perfect. . . . And then the next lesson was just like craziness. I was like, 'oh, bummer'" (Phase 4). Jennifer's comment illustrates how artificial the video clip collection felt for both the candidate and students, and the lengths candidates would go in order to provide a useable clip.

Jill asked her mentor to record her lesson and found that it, "went really well because [students] were really conscious of the video camera and so they were kind of nervous about who was being filmed and what was happening and [that Mentor] was doing the filming." Jill's mentor, "would walk up to them and film them and they would give a quick response to him and get back on task" (Phase 3). But when her mentor was not filming, students reacted differently:

> Today [Mentor] wasn't in the room and so he wasn't filming and they were just, I mean they were off task, some of them were off task more than usual and they weren't offering up suggestions. Every time I asked a question, typically I had to call on someone, except for a couple in the front row who I had talked to before and I kind of told them about the TPA and that I was nervous and so they were being nice and trying to help me out. (Phase 3)

Jill described some of the difficulties capturing an authentic lesson in her Lesson Debrief Interview. Below is an excerpt that highlights how a good lesson video can be hindered by things beyond the candidate's control:

> Researcher: "Did you learn anything about teaching from this lesson?"
>
> Jill: "There are always things going on that are not lesson-related. Today, I am not sure what he was doing but [Mentor] was in and out of the room and he has this guy coming in. I am kind of peeved about it because [Mentor] didn't tell him that I was filming and he kept

walking back and forth in front of the camera and so I don't think that I am going to be able to use any of the footage that I shot today."

Researcher: "Actually you can. Just address that this random person came. As long as you mention what it is. That is not your fault."

Jill: "It was interesting having him in the classroom. I've met him briefly before school started. He was in there doing technical stuff. So occasionally a sound would go off on [Mentor's] computer or he would test the volume level of the computer. So lots of, lots of distracting activities happened in the lesson that I was not counting on."

Researcher: "That actually is the next question. If you want to add anything you can. Identify any unexpected challenges or situations that impacted your TPA lesson."

Jill: "Challenges? The camera! My goodness, what a riot. It turns off after three minutes. It worked really well having [mentor] film. That was shot beautifully. But I didn't have a videographer [today] and so I had to set up the camera by myself and students would talk and I would try to bring the camera to them and try to manage the classroom at the same time and I just couldn't do it so I ended up leaving the camera set up and that was a mess. And then, yah, the technical stuff that was going on in the classroom, and [meanwhile the mentor] has this stage group that is coming in and out" (Jill, Lesson Debrief 1).

In addition to distractions and the lack of a videographer, other situations were beyond candidate control that disadvantaged candidates on Task 2. For instance, candidates must secure legal permission from guardians to record students' for their lessons. Faces of students without permission cannot be seen in the videotape. Jackie explained that she had to plan two lessons, one for the students she had permission to record and another for those students whose permission slips were never submitted. In some cases, candidates indicated that the lessons taught for the TPA were not authentic. Jackie's perception that the video component of the assessment was artificial is clear from the exchange below:

> Jackie: "I think it is difficult because only half of my kids turned back the permissions with a 'yes' and so I had my lessons taped with fifteen kids which was amazing but it was, it just happened to be all of my really, really well behaved kids and I didn't even plan it like that but, I think, especially with the videotape it is not always, it is kind of like a trick kind of thing. It is like this little moment that you have taped that might not be how you really teach."
>
> Researcher: "I want to make sure I understand what you said. It can feel like the video component can be artificial because you can kind of, for lack of a better term, "fake it" through whatever that requirement could be?"
>
> Jackie: "Yes." (Phase 2)

Jennifer related a similar experience where she felt like she could not teach the way she would normally in order to meet the expectations of Task 2. "Sometimes the kids were asking me these questions that typically I would have really enjoyed." And normally, "I would have taken the time to [say] 'let's discuss that' or 'let's diverge a little bit and I will bring you back around'" (Phase 3). However, in her TPA recordings, "because I felt like, wow, I have to make sure that I get this, this, and this in a ten minute video clip [I told students] I don't have time to talk to you about this right now" (Phase 3). This had a direct impact on student learning. Jennifer explained that she "actually had to tell my kids at the very beginning of the lecture, 'I know that you are going to have some really interesting questions but I need you to wait until the camera is off so that I can be certain that I get exactly what I need in ten minutes" (Phase 3).

Video requirements allowed for only one cut in the lesson and this changed the way candidates planned their lessons and interacted with students during the taught segment. "If you have a clip here, a clip here, a clip here, a clip here for each requirement" Jennifer said, "it doesn't work because they want it [the video clip] unedited" (Jennifer, Phase 3). Jennifer stated that "just knowing that if you stray at all from your lesson plans that you have to go back and explain exactly why that happened really kind of made we want to work in extra time." This was not her usual approach to lesson planning, "Usually, I like to just go right up until the very last moment," however, "for these lessons I made sure that there was like a ten to fifteen minute buffer because, if it spilled over into the next day, I would have had to bring another camera and we would have had to splice video clips" (Phase 3).

Jackie handled the video requirements by recording her first solo lesson before officially starting the TPA. "I know that [Sterner] suggests that you get two or three observations in first but I did it before my first observation. . . . I just felt a lot more comfortable getting it done" (Phase 1). Jackie described conversations she had had with other candidates which confirmed this choice for her. "I have heard a lot of my cohort talking about planning on filming and the fire alarm went off, or they were planning on filming and someone starts screaming, you know, and they are just freaking out. . . . Let's say something goes wrong, a lot of people kept forgetting to turn on the tape. Batteries were running out" (Phase 3). However, in the writing phase she expressed some regrets about this. She advised that candidates should do "the commentary quickly after the film because right now I am kind of forgetting why I did things and what I did" (Phase 3).

Based on the pass score, Jackie did not pass the TPA, earning only five points in Task 2. Jennifer's Task 2 scores earned six points and Jill earned seven points, indicating that video quality may not have been a factor in scoring. However, the perception from candidates was that the videography, more than content, mattered in the evaluation of their performance. Videography and editing are not taught in the teacher-preparation program and are construct irrelevant traits. In addition, the process of collecting the video relied on factors beyond candidates' direct control (mentor willingness and ability to videotape, working batteries, access to equipment that did not turn off in three minutes, or lesson interruptions). In addition, interviews suggest that video segments do not reflect authentic student behavior or authentic candidate performance. Candidate experiences show that lesson video clips were not always authentic to actual classroom and student behavior and may not portray actual teacher readiness.

TPA Handbook/Subject Area. One question of this project was whether the differences in the handbooks (based on subject area endorsements) were generalizable across the entire population. In other words, are all handbooks created equal? Five of the six case study candidates indicated that the specific TPA subject was a factor in their success. Discussion of the tasks and traits that were *universal* among all TPA handbooks were the focus of TPA trainings sessions. However, when candidates began their TPA, they discovered that there were some very specific, subject requirements. Jennifer describes how this felt:

> [We] were pretty comfortable with this whole idea of the TPA and all of the components. Like we know what student voice is and we know how to identify resources. . . . Then when we opened up the TPA packets and really started to dive into them we realized in the fine print the TPA was looking for us to demonstrate very specific skills. . . . A lot of people designed their history lessons without a whole bunch of primary sources and analyzing. Well, there might have been primary sources but not necessarily drawing these conclusions and deep analyzation processes. Another one of my friends in science did not realize that she would have to plan a hard core data analysis for biology and so the unit that she is supposed to be doing her TPA on doesn't lend itself to data analysis in a lab and so it is one of those things where we all know the components but then when we actually saw the exact requirements that they were looking for, for the individual subject area, it came as a little of a shock" (Phase 3).

Jennifer originally planned to complete her TPA in biology and then switched to social studies based on her relationship with her mentor. In the excerpt below, Jennifer compares the requirements for the social studies and biology TPA:

> Researcher: "So you would suggest a math/science TPA over a humanities TPA?"

> Jennifer: “Oh, definitely. Because you know what the crucial vocab is. It is so easy to say, ‘these are the crucial vocab words’, ‘this is the content’. And it fits perfectly with the biology curriculum. Humanities, not so much, especially when [school’s] history department doesn’t even have a standard …curriculum that they are all on board with. So it was much more up in the air for humanities.”
>
> Researcher: “Would you say that you feel that those performance objectives around primary sources are a fair expectation for social studies teachers or would you say that you feel like it is stifling you in your planning?”
>
> Jennifer: “You know, it should come naturally, but it is just not in history because I want to be more global and Socratic and they want very specific things for the TPA, which is what threw me. It’s like when I looked at it I was like ‘Oh, well the first day that I tape has to be for primary sources and the second day has to be drawing conclusions from those primary sources’. I think it is an appropriate expectation. It is very important that they know we can teach primary sources, so I understand why they put it in there. But it caught me by surprise, especially since I was planning to do it in biology. But even in biology I was caught by surprise when I looked up the two and I saw that in biology they had to observe me teaching an integrated lab. Well, I was planning to teach evolution. It is not really a lab based unit.” (Phase 1)

Jennifer may be reflecting the view that “the grass is always greener on the other side,” but her concern is reasonable. The TPA only recognizes one specific pedagogical strategy that must be used to teach the subject (lab based analysis in biology or primary source analysis in social studies). The perception is that this limited the types of units and lessons a candidate can plan in order to meet the requirements. Secondarily, it also led to the perception that the performance objectives were not equally difficult across subjects. In some ways, this perceptional data can seem like conjecture. The problem is that we cannot know how candidates would have performed if they had submitted TPA in another subject. A larger sample of secondary candidates would help provide data to support or refute these claims. If some curriculum can be more easily adapted to the TPA requirements than another, those candidates had an advantage beyond a candidates' control which unfairly advantages some over others.

Support to validity. Differences in candidate experiences are unavoidable in a performance assessment of this nature. Efforts to standardize candidate experiences are likely to create systematic errors in construct representation and extrapolation (see Chapter two). The need for high levels of standardization in the assessment procedure and flexibility to address the unique contexts of ST makes it more difficult to generalize across a set of scores and compare diverse experiences. In fact, efforts to standardize a performance assessment may reduce its benefits as an authentic instrument. As observations of performance become more standardized, they may be perceived as less representative or accurate. Procedures to improve standardization are important for reliability, efficiency, scoring, and, especially fairness, but may contradict the benefits of authentic, context-specific and unique characteristic of performance-assessment.

Threat to validity. Accountability comparisons across programs are likely to be influenced by how well programs prepared candidates to take the test and trained mentors (especially around TPA requirements), as separate from preparing candidates to teach. Candidate concerns about videography and mentor relationships reflect such differences. When significant differences occur in test preparation and procedures, the "resulting systematic errors can be especially serious for the kinds of absolute interpretations (e.g., in terms of achievement levels) typically employed in accountability systems" (Kane, 2011, p. 26). Candidate interviews suggest that several factors beyond their control impacted their performance on the TPA. These factors include relationships with the mentor, classroom and school setting, and handbook (or endorsement) area. In addition, differing performance expectation across handbooks may introduce errors of measurement. Success may also be attributed to a supportive mentor-candidate relationship, or to a mentor's knowledge of the instrument or prior experience with National Boards.

Inference 4: Extrapolation

Research Question

8. Do TPA test scores provide reliable indicators of a readiness to teach? Are the TPA scores, as a whole, a true measurement of teaching ability?

Validity Evidence 12: Professional Growth Plans: Do Identified TPA Strengths and Weaknesses correspond to Mentor/Supervisor/Candidate Evaluations?

At the culmination of ST, undergraduate candidates, supervisors, and mentors rated the candidate's teaching performance. From these ratings and their own personal reflections on the experience completing the TPA, candidates crafted a Professional Growth Plan (PGP) identifying three strengths and three goals for their first year of teaching. Each plan was authenticated by the mentor and supervisor. TPA scores were averaged by task and categories (AL and SV). The lowest task average was identified as a candidate weakness and the two highest task averages were recorded as strengths. The results of candidate PGP were then compared to TPA rankings to determine if TPA scores provided reliable indicators, for each of the measured constructs, and whether TPA scores, as a whole, provide a similar measurement of teaching ability.

Findings. Strengths and weaknesses identified in the PGP were correlated with TPA means in 31 out of 90 (.34) cases. When comparing candidate strengths on these two measures, the rate of agreement is lower, just 17 in 60 cases (.28) agreed. TPA scores and PGP self-reports for candidate weaknesses in teacher readiness agree in 14 out of 30 cases (.47). An examination of the reasons for disagreement between PGP and TPA scores may be due to the identification of "classroom management" as a strength (8/90) or weakness (20/90) in the PGP. In addition, rapport with students and their families, and the development of professional relationships was rated as a strength or weakness in 39 of 90 (.43) responses. However, classroom management and rapport are not directly measured in the TPA. When removing rapport and classroom management selections, rater agreement rises (.59).

Support to validity. TPA were submitted at the mid-point in the ST term. It is therefore expected that candidate teaching performance would improve between TPA submission and the PGP. In addition, mentors, supervisors, and candidates had not yet received TPA feedback scores when the PGP was developed. However, weaknesses identified in the TPA corresponded to weaknesses identified by mentors, supervisors, and candidates at the end of the term in half of the cases (fifty-nine percent, without classroom management and rapport). This finding suggests that there is some evidence that the TPA may be reliable in measuring weaknesses in teacher readiness and may be plausible as a threshold test.

Threat to validity. One of the purposes of the TPA is to provide an on-going measure of teachers' abilities throughout a career. To meet this purpose, TPA scores should be reliable measures of teaching strengths and weaknesses. Identification of the strengths of a novice teacher is an important part of that process. Rater-agreement regarding teaching readiness was low, between the TPA and PGP, and especially for teaching strengths. This finding suggests that raters may not agree on candidate strengths. While TPA scores may be more predictive of candidate weaknesses and areas where they are not yet ready, this data suggests that it is problematic to hold teachers accountable for weakness and strengths identified by TPA scores, beyond licensure decisions (see Chapter two).

Validity Evidence 13: Case Study: TPA Scores as Measurement of Teaching Ability

Case Study responses have been reported throughout this chapter. In this section, the whole of the TPA and ST experience was emphasized. Case study candidates were asked whether scores derived from the TPA reflected their perception of their teaching readiness and whether they found the TPA to be a true measure of their teaching ability. One difficulty with data collection and analysis was that TPA scores were not reported to the university nor to candidates prior to the term end. Therefore, candidates were asked to provide views without knowing scores or pass rates.

Findings. Interviews reveal that candidates perceived the TPA to be disconnected from teaching and a “hoop” in achieving licensure. Conversely, candidates also responded that the process of completing the TPA was formative in their understanding of the practice of teaching.

The TPA is a “hoop” toward licensure. Every candidate interviewed (100%) expressed a view that the TPA was not an authentic assessment of their teaching. Reports of the formative value of the experience were mixed but that “it was just a hoop that I needed to jump through” was universal (Jennifer, Phase 4). Below are excerpts from interviews:

- “In the end, for me personally, I don’t feel like the TPA was something that made me a better teacher. It was just a hoop that I needed to jump through. I am not saying that is necessarily wrong [be]cause there needs to be standards but I put more stock by far into how my professors think I am doing; how my supervisor and mentors think I am doing” (Jennifer, Phase 4).
- “It doesn’t really evaluate how I will teach. It is just a small hoop. I guess the balance between learning and accountability. … I guess for me it felt a lot like accountability. I really did. I felt like ‘Alright I just need to get these lesson plans done and they need to look good,’ on paper. And then this one video clip has to practically be staged. I mean it has to *look* good. I’ve got to get my smart students talking and saying it with academic language and if I don’t I am toast and so there was a lot of planning into how can I make this perfect. But really it is like, there is no way, there is no way that I can teach like this. So, yah, definitely a lot more about accountability” (Jason, Phase 4).
- “It’s a little bit like jumping through hoops, but at the same time we’re responsible for students’ futures and their education so I think that it’s fair to expect that we can do all these things” (Jane, Phase 1).
- “It’s a little difficult because I feel like I am not as focused on my student teaching as I could be. I feel like so much of my energy is directed at writing these extensive lesson plans and writing up for the commentaries that I don’t get to spend as much time in the

classroom. Even now, like the fact that [Sterner] has pulled us out of our classes to write this just shows how much our energy is more focused on the TPA than it is on our classes, which is unfortunate, because I really miss my kids and I want to get back to my classes but I have, you know, four more pages to write so that is not a possibility at the moment" (Jill, Phase 3).

- "I just found it frustrating that while I was teaching the TPA I really felt like I wasn't focused on my students at all. I was just focused on, 'Is the technology going to work? Do I have both of the cameras recording in case something goes wrong? I can have both of them and what happens if they malfunction? And am I collecting all the evidence that I need rather than focusing on my students actually understanding this?' So it was really hard for me professionally, as well" (Jane, Phase 4).

In her interview, Jackie reiterated comments in Phase 2 that success could be "faked." Her attitude about the process demonstrates that candidates felt the requirements of the TPA were disconnected from the actual practice of teaching. Below is an excerpt from that interview:

> Jackie: "I think [the focus is] definitely the licensure part – completing the tasks by jumping through the hoops was much more important than learning how to be a good teacher."
>
> Researcher: "Okay, and why would you say that?"
>
> Jackie: "Well because you can fake the teaching part, kind of. I mean, you are writing a lesson and that is something that you, I mean, obviously the parts you are not going to miss are the parts that they require. In the lesson plan and the video, you can tape as many times as you want with or without however many kids, so you are not learning how to do anything. You are just, giving [scorers what they want to see or] going to start over and over."

Jennifer describes the process as a game, "my kids were great. They were so angelic and just played the game and they did it well" (Phase 3). Of her own performance, Jennifer reflects that the TPA in ST, "Not only seemed redundant but at times it seemed to come in conflict [with] expectations of my mentors and the schools … if we are talking in terms of relationships and being creative and being there for your students. I felt like it pulled me away" (Phase 4). She goes on to explain that:

> The two or three weeks that I was probably least effective at my job was the weeks where I was planning and writing and doing the TPA because I was so concerned about all these nitpicky little things the TPA is going to want and trying to jump through the TPA hoop that I lost connections with my kids. (Phase 4)

The feeling that the time spent on the TPA took away from time spent teaching increased candidate perception that the TPA was less about teaching performance and more about writing to the prompts or capturing the right kind of evidence on the video. All of the candidates interviewed, to varying degrees, expressed the view that the TPA, rather than an embedded and authentic performance of assessment of their teaching, was a detached and sometimes arbitrary evaluation. The most significant factor that caused candidate experience to vary (and candidate attitude toward the TPA to be most negative) was the relationship between the candidate and the mentor (see VE 11).

Teacher formation. When asked about the formative experience of the TPA, each (100%) of the candidates indicated that they *did* grow or learn something valuable about teaching in the TPA process. Jason expressed mixed views of the value of the TPA in ST. In the first three phases, he was critical of the TPA but in his summative reflection he states:

> I didn't like it, but there were good things. I retract my last statement [from last interview] that I didn't care for it. It is not a bad assessment piece but tell someone that they need to do this for every lesson [and it] is completely and utterly unrealistic, and you should be fired from your job, because obviously you have never taught in a classroom before. (Phase 4)

Jason's change of perspective is, at least in part, a result of discussions he had with his mentor about the TPA experience. He goes on to say:

> Me and my mentor teacher actually talked about this, too, and I think the one good thing …is the self-reflection piece. ... And I mean we do that a lot now, more so, we think about, 'okay how could this lesson have gone better? What do we need to do to catch these kids up? Or, where can we move now or do we need to go back and do something else?'… Another thing was academic language, I guess. We just call it vocab. …We are doing a lot of vocab-oriented steps, especially with math because math is just huge on vocabulary…. Then the final piece was a little bit [of] differentiation. My mentor teacher didn't think about it that much but now that I told him that we have to have this in here now, we are thinking about it more. 'How can we catch so-and-so up? What can we do to help him out? Should we talk with his specialist or resource room teacher and collaborate with them?' …. So we are thinking about that and trying to do better. (Phase 4)

Support to validity. Candidates found that the process of completing the TPA helped them to identify strengths and weaknesses in their teaching practice. Specifically, they felt they grew in their understanding of AL, SV, using assessment to inform instruction, and reflecting on their practice. This process was significantly more beneficial when undertaken *along* with the mentor in a co-teaching setting.

Threat to validity. Participants uniformly agreed that they did not view the TPA as an authentic measure of their teaching practice (see also VE 7). The evidence shows that candidates feel like the TPA overwhelms the ST experience. Rather than an integrated component or, as was the intention expressed in the handbook (see Chapter two) and a source of formative feedback for growth, it took away from learning. One candidate explains, "The TPA should be during the semester before student teaching. That way it wouldn't *interfere with student teaching*" (TPA Completer, 3/22/12, emphasis added). Each case study candidate described the TPA as a "hoop" during the course of their interviews in the term. Many candidates used the term "hoop" when describing the TPA requirements in surveys. One illustration from was the candidate response, "This just felt like it was a useless hoop we had to jump through that didn't test how well we could teach, just tested how well we could vomit up teacher-words and follow a rubric, which doesn't seem to have much to do with real teaching" (TPA Completer, 3/22/12).

Several candidates indicated that the TPA performance could be "faked." In various ways, case study candidates argued that the TPA was not a reliable indicator of readiness, though it did prove helpful in the formative development of their teaching. In her final interview, Jill was asked if there were any correlations between the strengths and weaknesses she identified in her PGP and her expectations for TPA scores. She responded that, "Kind of, in the sense that it is opposite. I think the TPA shows me as a very stilted, rehearsed, not very calming presence but the kids are really well behaved and the classroom seems very well managed. So it is almost kind of the opposite" (Phase 4).

Inference 5: Decision Making

Research Question

9. Is guidance in place so that all stakeholders know what scores mean and how the outcomes will be used?

Validity Evidence 14: Document Analysis

PESB documents, published on the website, *www.pesb.wa.gov,* were analyzed to understand the consequences of test score use in WA. These documents include a timeline of policy decisions (see **Appendix V**). In addition, the document *edTPA Field Test: Summary Report* published by SCALE in 2013 was analyzed for recommendations of cut scores and test score uses.

Findings. Data was collected during a field test, the first state-wide pilot of the TPA, in WA. It was understood that decisions about consequential scores and dates would be made after the field test, set to begin in 2013. In September 2013, PESB set the consequential date to January 2014. After this point, only candidates that have completed a TPA may be recommended for licensure. In November 2013, following the recommended cut scores published by SCALE, PESB adopted the cut score of 35. In addition, regarding the WA specific SV requirement, PESB decided that:

> The student voice rubrics will not be consequential for candidates but that candidates will continue to submit portfolios that address the student voice prompts, vendor will continue to

> score student voice rubrics, staff will return annually with student voice data, and members will consider taking action on student voice rubrics in three years. (PESB, 2013)

Universities were introduced to the TPA in September 2011 and had one year to prepare for the field test and another year to make programmatic adjustments before the TPA became consequential.

Support to validity. None. Guidance was not in place before the field test. Some guidance was offered during the field test through an on-line TPA network and statewide meetings for SOE administrators.

Threat to validity. Score meaning was provided two months before the consequential start date. Prior to this, universities had one year to prepare for the field test and another year to make programmatic adjustments before the TPA became consequential. Given that candidates enter a three to four year program, this did not provide enough time to alter course catalogs, adjust curriculum, evaluate faculty, and make necessary modifications to support candidates. Because only raw scores were reported, candidates and programs were not provided enough time to understand score meanings for accountability before the TPA became consequential.

Validity Evidence 15: Faculty Interview

In order to determine whether university faculty understood the meaning of scores and how those scores would be used, Sterner faculty were interviewed in Phase 4. The interview was coded for patterns.

Findings. Eight (73%) faculty interviewed responded that they needed more guidance to know what scores meant and how the outcomes would be used. Specifically, they were uncertain about how scores would be reported and shared with constituents and what kinds of supports could be provided by universities for candidates completing TPA. The faculty member charged with TPA coordination at Sterner described the dilemma with Pearson’s score reporting for the field test:

> We will not be sharing score numbers [with faculty or with candidates] because we were not provided with any explanation or interpretation as to what the scores mean or what the

> cut score is. Nothing yet has been shared with faculty, but we will be trying to understand the scores in order to best evaluate our programs. (5/14/12)

Without clear understandings of score meaning connected to actual candidate performance, universities were left to "interpret" what scores meant for program evaluation and candidate pass rates. Pearson only allowed faculty hired to score to have access to scoring materials that might facilitate accurate interpretations, like the scoring document "The Thinking Behind the Rubrics." Faculty-scorers had to sign a legal agreement not to release any scoring materials or information about scorer training. This further hindered the ability of programs to understand score meanings. Programs were left to wait until the following academic year to receive scores that might include explanations for performance rankings.

In addition, faculty expressed concern about the consequences of the operationalized construct of TPA for the profession. Some concerns echoed supervisor interviews and candidate comments about the ability of candidates to submit work that was artificial. One faculty member stated, "Students will do whatever is needed to pass the TPA. I can see it coming down the road where some students will be 'buying' answers to the TPA online, just like they can buy a term paper" (5/14/12). Especially troubling is the prediction that the TPA is likely to negatively influence diversity in teacher recruiting, particularly for teachers whose first language is not English. Reading and writing ability as a construct irrelevant trait measured by the TPA will cause a, "depreciation in the number of English Language Learners who will be completing Teacher Education programs. The technical writing skill levels required to pass the TPA are not compatible with the language skills of ELLs" (5/14/12). In addition, the guidance and support for students with disabilities remained unclear. A professor of special education stated that, "there have been no accommodations made for students [candidates] with disabilities. This has been brought up many times with the PESB with no response" (5/14/12).

Not surprisingly, faculty expressed concerns about how resources would be found to train and support all candidates. The undergraduate department Chair stated that:

> The TPA is daunting and draining on resources (time and budgets). It is stressful for faculty. Some of our new classes will be picking up the training and instruction lost from the seminar courses [that have had to add TPA preparation], but it has been a big project to get those courses off the ground. We have had challenges training the adjuncts who don't 'work for us' but for other departments at the university. These resource demands filter to those other departments as well.

These concerns have been expressed by faculty at other universities, as well. Alan Singer (2013) writes,

> [We] received from Pearson numbered ratings on each portion of the submission for each of the students. However, the evaluations did not include any comments on their strengths and weaknesses and there was no notice about what is considered a passing grade. . . . I could only guess at how they were rated. . . . University faculty who are supposed to prepare new teachers can only guess why the students who participated in the field test received the scores they received and how to help students improve in the future. (p. 1)

Support to validity. None.

Threat to validity. It is important to qualify these findings because the data collected to address Inference 5 occurred during a field test of the TPA. One of the purposes of the field test was to determine base line data for use for passing criteria and to learn how to "build the plane while flying it" (WACTE, April 2012). Some faculty concerns were, therefore, unavoidable. In addition, one complication involves "all the cooks in the TPA decision-making kitchen." Decisions about the operationalized construct are controlled by the TPA developer, SCALE, part of a private university. Decisions around the scoring procedures, candidate submissions, and scorer training are contracted through Pearson, the largest for-profit educational company in the world. The ultimate consequences and score meanings for *candidates* are determined by each state government and, in WA, the PESB. The consequences for accountability for *university programs* are determined by AACTE, CAEP (nationally) and PESB (WA).

However, by limiting the scorer training to only those who are hired by Pearson, important teacher preparation faculty do not receive knowledge of the scoring procedures essential to understanding the operationalized construct. One of the stated purposes of the exam was to improve teacher preparation, but un-trained faculty are less able to interpret what scores may indicate about readiness or to provide necessary support to prepare candidates and those who are trained cannot share what they know. Because raw scores were reported in the spring 2012 scoring group, and those scores did not provide explanations, faculty interpretation of score meanings was limited. Without score interpretations, it proved difficult to make some necessary programmatic changes before the assessment became consequential. Faculty interviewed were concerned about score meanings and consequences, TPA process and procedures, and the impact of potential decisions about TPA scores on the field and profession. Faculty work within systems that require national and state accreditation. These agencies have already endorsed the TPA (edTPA) for licensure decisions. Singer (2013) expresses feelings of futility about the lack of time to fully prepare candidates and how the failure of guidance will have no impact on the consequential timeline in his state. “I have no idea what State Education Departments in these states will decide constitutes an acceptable score for certification or if they will ever agree, but the new tests are scheduled to go into operation in May 2014 anyway” (p. 1).

Validity Evidence 16: Case Study

Case study candidates were asked how well they understood the outcomes of their performance, what the scores might mean, and whether they felt there was guidance in place for how decisions would be derived from TPA scores. As mentioned above, one difficulty with data collection and analysis was that TPA scores were not reported to the university or to candidates prior to the end of the term. Therefore, candidates were asked to provide their views without knowing their scores or overall performance on the TPA.

Findings. Like faculty, all case study candidates expressed frustration that they did not know when, if, and what type of feedback they would receive about their performance or what decisions might be made from those scores. Findings also reflect candidate confusion about completing a summative evaluation during a formative experience. Jackie says, “In order to learn you have to do something and then [get] feedback, immediate feedback” but “I don’t get anything back until August [four months after submission]. I have no idea how I did” (Phase 4).

Uncertainty about scoring and scorer interpretation of the handbook led to feelings of anger and frustration and accusations that the handbooks were not ready for field testing. This is illustrated in the excerpt below:

> Jennifer: “I am feeling kind of lackadaisical about it because it is not high stakes for me. I think I would feel very different if it was. I would not be super excited, especially just, you know, [because] in the end it kind of just comes down to little things. The fact that there is so many mistakes in the pamphlets that we were given, that we had to dig to figure out what they really wanted, the fact that they kind of wrote in a convoluted way. Those are things that students get upset about when they have to take a test from a teacher. It is what we are taught not to do. You make the expectations clear, you teach the test, and you don’t throw any curve balls and that is exactly what they were doing with the TPA, you know. You proof read your own stuff. You don’t judge kids for not knowing what to do when your test is thrown together and it looks like crap and that is kind of how I felt when I was writing the TPA. There were portions in there that were science requirements in my history pamphlet.”
>
> Researcher: “That is really embarrassing.”
>
> Jennifer: “I mean it is not little things. They were like glaring, ‘oh, my gosh, you have the wrong requirements in here.’ Can you imagine if you did something like that on a test and then gave them to students? . . . Then say, [to your student] ‘you’re not moving on to the 11th grade if you don’t pass this test that is convoluted and thrown together.’ No.” (Phase 4)

Many candidate questions were about the scoring procedure. Because faculty were not trained by Pearson, and the training materials were not available unless you were hired to score for Pearson, it was difficult for faculty and supervisors to advise candidates. Jill says, “Overall, I think the TPA is a pretty accurate representation or asking for accurate skills and knowledge from teachers [but] the scoring process I was a little confused about” (Phase 1). Candidates were asked if they believed that they met standard. Jennifer states that “it is hard for me to [know] until I see the scores, because I mean, obviously, I think I am doing what I am supposed to but I don’t know what a three really looks like, or a four, or a five” (phase 2). She goes on to clarify her question about scoring, “are they looking at my commentary, at my lesson plans, are they looking at it as a whole?” Jennifer’s worry is that she needed answers to every prompt and rubric requirement to appear in both the artifacts and commentary. “I am afraid that they might misunderstand something” (Phase 4).

Similarly, Jill reflects that, “I am not sure if any of it will meet standard. Just kind of across the board, [I am] really unsure of what it is that they are looking for. I have tried to be as articulate as possible in my commentary but there comes a point where I just can’t repeat myself anymore and I feel like they are either going to get it or they are not going to get it” (Phase 2). Lack of shared “gold-standard” samples made interpretation of the handbook more difficult. Jamie describes this difficulty, “The thing that got me through and really helped me in [earlier coursework] . . . was being able to look at somebody else’s [work] and see what it takes to meet standard” (Phase 4). She verbalizes that samples of performance are helpful, “not because I want to copy but I just want to see ‘okay they might be looking for this’ or ‘they might be looking for that’” (Phase 4).

Other concerns were expressed about how scores would be calculated for passing. For instance, would rubric scores be averaged? Would there be a cut score? Would candidates pass so long as they never scored a one on a rubric? What if they exceeded expectations on multiple tasks but failed one rubric (Jennifer, Phase 4)? Jennifer says, “At this point I don’t even know how I am being graded. I mean, do you have to get all three’s” (Jennifer, Phase 4)?

As stated earlier, Sterner decided to ask candidates to submit the TPA at the midterm point because it was not clear whether future candidates who did not pass would need to submit another TPA in the same term. The consequence of this decision was that candidate submissions revealed their teaching abilities in the middle of ST, rather than the end. Jill describes the problem:

> I think that the way that it [TPA] is put into the ST experience it is more of a formative assessment because it is early on and we haven't had much experience in the classroom. But they are using it as a summative assessment which is kind of one of the reasons that I feel it is unfair because it's measuring me at the start of my teaching and not when I am comfortable in the classroom, comfortable with the students and ready to videotape myself. (Phase 3)

Jamie says, "it would be great to . . . have somebody look through it . . . and give me feedback on the whole thing because then I could have maybe applied that to the rest of my student teaching" (Phase 4).

Support to validity. Some of the concerns expressed by candidates were a result of the timing of TPA submission, a decision made by Sterner, not SCALE or Pearson.

Threat to validity. The purpose of ST is to provide candidates a formative learning opportunity and to guide their practice as a teacher. Without feedback, or simply as a summative assessment, the TPA does not offer candidates opportunities for growth and is, therefore, a questionable pedagogical tool to introduce *during* ST.

Not enough guidance was in place for candidates to understand and interpret tests scores and their meanings. In fact, due to this, Sterner did not disclose raw scores to candidates. Instead the program provided candidates with areas of "strength" and "focus" based on mean scores on the tasks and categories. More guidance is needed for candidates to understand the scoring procedures (specifically, relationship between artifacts, commentaries and tasks in the scoring process), score outcomes, passing scores and requirements. In addition, samples of performance at different rubric levels are needed to facilitate the interpretation of the Handbook requirements.

Conclusion

This chapter discussed the validity assumptions of the WA TPA field test in spring 2012 using the ABV methodology articulated by Michael Kane and the Cambridge Reporting Framework developed by Shaw, Crisp, and Johnson. Of particular importance was the extent to which test scores could not be generalized across different ST contexts. The overall findings suggest that the operationalized construct was stable but scores were not generalizable and guidance was not in place regarding score meaning and use prior to the field test. Low correlation between the TPA and university instruments provided divergent evidence for the use of the TPA, indicating that decisions solely based on TPA may not be reliable. The risk of making the wrong decision is high. Other inferences (Inference 2 and 4) suggest potential weaknesses but not failures of validity. The next chapter presents the validity narrative and discussion, implications and next steps for study and a brief discussion on the ABV model.

Chapter Five

Synthesis

This study developed validity interpretations in order to argue that scores derived from the TPA can be used to determine teacher readiness and to evaluate the quality of teacher preparation programs. Validity questions, especially those for complex, high stakes performance evaluations, do not have simple answers. An argument establishing the central aspects that influence test score interpretations are necessary to narrow the seemingly limitless scope of any validity investigation. In argument-based validity investigations this begins in the form of inferences listed in an interpretive argument. Each inference prompts a particular investigation. Underlying inferences were examined by judging the assumptions of the validity claims using multiple types of evidence. The more plausible the assumptions, the more valid the test score interpretation. This chapter synthesizes those inferences in order to create a validity narrative of the interpretations of TPA score use. The five validity assumptions that informed the RQs are revisited and critically reevaluated in relation to the evidence presented (Chapter four) and the overarching validity argument. A discussion of the use and limitations of the ABV methodology are addressed. Finally, recommended TPA research and implications of this research for programs and future ABV studies are presented.

Summary of TPA validation findings and evaluation of the argument

The validity argument evaluates the interpretive argument (IA) through empirical checks on the assumptions essential for each of the inferences. The focus of the argument is to determine the weak links in the IA, in order to improve the assessment. In other words, ABV is primarily interested in whether there are weaknesses or potential flaws in the IA *in order to correct them*. Applying the Cambridge Framework, there are two final steps to determine, as a whole, if the IA is plausible. The rest of this chapter presents a summary of evidence in these two steps:

- *Step 1*: Evaluate the assumptions and statements included in the argument individually and decide whether each statement or assumption is *accepted*, *rejected*, or *not investigated*. This is reported in **Table 5.1** and **5.2**.

- *Step 2*: Assign an evaluation status to the inference as a whole: *justified*, *defeated*, or *unevaluated*. This evaluation is reported in **Table 5.3**.

Validity Assumptions Revisited (Step 1)

Because the TPA is a new instrument, SCALE has used research from the PACT as evidence to support use of the TPA. This study supports the findings published in the PACT literature; PACT weaknesses and strengths are also TPA weaknesses and strengths. The following findings reiterate PACT studies:

1. The TPA is overly time consuming and takes time away from candidate focus on instruction, professional development, coursework and personal life (Okhremtcouk, et al., 2009).
2. The TPA did support candidate perception of professional growth as a teacher (Okhremtcouk, et al., 2009).
3. IHE and placement site support networks were *essential* to candidate success (Okhremtcouk, et al., 2009).
4. Candidates believed their placement setting to mandate teaching decisions in such a way as to be limiting to success. This was most true of candidates in districts with larger populations (Pecheone & Chung, 2006).
5. Supervisor and TPA scores do not correlate (Sandholtz & Shea, 2011).
6. Participants have difficulty managing competing values of the formative nature of ST and summative aspect of high-stakes assessment (Lit & Lotan, 2013).
7. The TPA reduces academic freedom by limiting the intellectual diversity of competing theories of education which leads to homogenization of curriculum (Lit & Lotan, 2013).
8. The teaching event is often viewed not an opportunity for formative growth but a hoop to jump (Lit & Lotan, 2013).

One important difference is that the PACT has been described as a formative assessment. In developing the TPA, SCALE made a substantive move to label the TPA as a summative assessment (SCALE, 2013) in order to justify the semantic and policy assumptions of test score use. Paul Newton (2012) describes how this change alters the validation argument:

> The products sold by test publishers are procedures: assessment procedures, or, more broadly still, assessment-based decision-making procedures. The guarantee given by a publisher is that, when outcomes from a correctly administered assessment procedure are interpreted appropriately, they will be fit for making certain kinds of decisions. This would seem to be the essence of validity within the *Standards.* Declaring a particular procedure to be valid can therefore be understood as a speech act which promises, or guarantees, a good decision-making. The declaration of validity provides a green light to employ the procedure, as specified by the vendor for the purpose at hand. (p. 18)

Because SCALE claims the TPA is a high-stakes, summative assessment to be completed during ST, and that internship is a formative experience important to the development of the teacher, evidence to support this claim, and its underlying assumptions, must be sound and fully plausible to support good decision-making for teaching licensure.

Assumption 1: The tasks in the TPA represent the particular performance and skills needed to base a decision on teacher readiness.

Teaching is a complex, multi-faceted professional skill. Learning about teaching is not the same as learning to teach, which is different yet from actual teaching practice. Decisions of teacher readiness may be supported by the theoretical construct of the TPA. This construct was translated into performance requirements and these requirements are the operationalized construct. The operationalized construct represents most of the important skills that can be used to determine ability, especially weaknesses in performance, but the TPA also evaluates construct irrelevant traits. Therefore, scores need to be situated within a larger experience of ST in order to be used to base decisions of readiness. Alone, scores on individual tasks as operationalized may not represent particular performance and skills necessary for teacher readiness.

What the TPA does not sufficiently measure or evaluate are those activities, dispositions and expectations of teachers that are difficult to capture in any assessment. Among other criteria, these include management, relationships, leadership, advocacy, intercultural competence, and guardianship. The ability to connect with students, their families, colleagues, and administrators from differing socio-emotional and cultural backgrounds, and to advocate on behalf of those students may be criteria better evaluated after a candidate has had more experience with which to form judgments. However, because assessment "signals" what is important (Whittaker & Young, 2002), it is clearly an omission that these are *not* included in the TPA especially because the TPA has selected other tasks and traits that are developmentally advanced for teachers. Learning environments are social settings which have situated perspectives that locate a person, whether teacher, student, or evaluator, in a community of practice. The *situatedness* of human experience as socially constructed is amplified in a classroom, itself a socially constructed learning environment (Vygotsky, 1978). Situatedness is necessarily a component of the construct for which any performance assessment of authentic teaching practice must address. While the TPA does not ignore this, asking for candidates to describe the learning context of the classroom, it does not adequately address the differences in this context in order to compare candidates to the standard of performance. A troubling result of the operationalization of the TPA is that candidates had *less* time to practice teaching and less time to address important expectations of readiness not measured by the TPA. Even if the operationalized construct and use of scores for licensure were found to be sound, for the purposes of accountability these issues make it difficult to determine whether identified weaknesses are a failure of preparation, or a failure of application of learning.

Finally, mentor frustration about having to adapt classrooms and curriculums to support candidate work could also account for the change in view about the value of the TPA between the start of term (phase one) and its conclusion (phase three) (see VE 1). It would appear that mentors believed that, in theory, the tasks and categories evaluated were sound, but when applied to classroom instruction and teacher preparation, these traits proved too different from their teaching practices. One mentor explained that, "there is too much focus on the TPA and not enough 'real world' focus. Can the candidate really manage and teach? Does the candidate regularly include 'TPA like' lessons or do they only teach that way when being evaluated" (3/30/12)?

Given the evidence provided in this study, decisions about scores that are extrapolated to the larger domain of preparation and schooling are problematic. Assumptions for accountability decisions are not supported.

Assumption 2: The performance criteria on the TPA that measures teacher readiness should conform to the standards for teacher preparation in any state for which that tool was adopted.

The TPA is aligned to WA standards. The content assessed is representative of the current theoretical understanding of the professional activities of teaching. The TPA was designed to align with INTASC standards for teachers. There are ten INTASC standards, each with performance indicators. Standards one through seven align directly with TPA requirements. Some of the performance indicators in standard eight are met by while others are not. Standards nine does not align with the TPA. Similarly, WA PESB standards for teaching were aligned to the TPA. Nine of the thirteen *Standard V* requirements can be measured using the TPA. Because the INTASC standards are widely regarded to be the national standards of teacher preparation, such close alignment between the TPA and INTASC standards demonstrates strong construct definition for teaching readiness and sufficiently reflects the views of experts in the profession. Based on PESB *Standard V* requirements, the criteria of the TPA represent, to similar depth and breadth, the state standards for teacher readiness. Applying the researched INTASC and NBCT standards, criteria on TPA are aligned to established expectations of observable teacher readiness.

Conversely, feedback from experts in this study indicates that TPA are aligned to expectations of teacher readiness with the exceptions of classroom management, dispositions, and rapport with students and families. So long as part of a larger framework of candidate assessment that allows the university to make judgments that augment TPA scores, this assumption is plausible.

Assumption 3: Performance traits on the TPA are related to performance of the same traits measured in other assessments of teacher readiness.

Proficiency levels on each of the 15 sub-traits (rubrics) are generally adequate. However, at least eight (.50) rubrics show higher levels of difficulty because of lack of opportunity to learn (AL and SV) or construct irrelevant traits (videography, reading and writing ability). Histogram data demonstrate that the frequency and standard deviation follow a fairly normal distribution curve. However, because of the low cut rate score, score use is weighted toward passing indicating higher levels of proficiency than university evaluations suggest at the point submission. Several factors could explain the pass rate including the high quality of the candidates in the sample, the level of preparation the program undertook to prepare the candidates for the exam, scorer bias, (the full sample was not double-scored by trained scorers) or a low cut score.

Inter-rater agreement between the two scorers of the samples used in this study reveal low agreement rates, especially for a high-stakes evaluation. In addition, rubric analysis suggests that the majority of rubrics apply language that require professional judgment. Scorer training requirements are generally sound but qualification requirements allow for too great a disparity between the “gold standard” and scorer performance. Differences in rubric level scores (+1/-1) can influence pass rates, and thus licensure decisions. For this reason, score confidence is low if raters can differ on 40% of the rubrics and still qualify to score.

Reasons for differences in inter-rater agreement may indicate variance in professional opinion on the task or how the candidate performed the task, rater selection, differences in scorer training, a lack of understanding about scoring the rubric criteria, or a lack of understanding of the constructs being measured in the rubric criteria. It is also likely that the rubrics that required an examination of evidence or those rubrics where the videotaped segments were analyzed for scoring had greater variation (Graham, Milanowski, & Miller, 2012). This was true in the study for Task 3 assessment evidence, but not for the video. Inter-rater agreement differences could explain why SCALE recommended a fairly low cut score for licensure decisions (see VE 4).

Another explanation could be that, although specialists were sought who demonstrated experience in the subject area and in teaching and working with novices, several of the sub-traits were unfamiliar to raters (AL and SV). Because SV is a requirement for WA, but the raters are hired and trained across the US, SV is likely to be a new trait for scorers unfamiliar with WA standards. In fact, SV presents a host of scoring dilemmas. When this study was conducted many WA based teachers were not yet familiar with SV, its meaning, its practice, and what it would look like when done well. Scorers from WA, outside of higher education, may not be trained in SV. Scorers from other states may not be familiar with SV at all. Because the scoring process is meant to be double-blind (scorer does not know participant and participant does not know scorer), this creates a training issue independent of the rest of the TPA rubrics. If scorers do not receive separate SV scorer training, inter-rater agreement is likely to remain low. For this reason, it is essential that there be multiple-raters with agreed scores for every WA submission.

It is potentially the case that there is rater-difference in the definition of the construct of teacher readiness and the measures and traits that define teacher ability. For instance, expert opinion (see VE 1) highlighted classroom management and rapport as missing in the operationalized construct. PGP results indicate the largest differences in rater agreement occurred because of these two constructs (VE 12). Even if there were high levels of scorer agreement, the theoretical construct for readiness may not reflect good or bad candidate *preparation*. “Inferences from test scores to quality of schooling are problematic and must depend on a great deal of contextual information” (Haertel, 1999, p. 8) about which the operationalization of the construct in the TPA may not collect or standardize enough for program accountability decisions. This facet of test score use, for use in programmatic accountability, deserves more research.

It is unclear whether the cut score recommended by SCALE, and adopted by PESB, accurately measures whether a candidate met standards. In using a cut score, a candidate could score high on one task, struggle on others, but still pass the TPA. For instance, candidates are likely to score well in Task 1: Planning, as they will have had much practice planning in their coursework or the curriculum may be scripted. Consider this example, a candidate using scripted curriculum could score a proficient mark of “3” on planning rubrics, earning nine points. Then the candidate could proceed to do poorly on the actual instruction of those lessons earning an inadequate score of “1” on the following two rubrics, for two points. Because the assessment is already provided by the curriculum, the candidate would likely score well on the next task if they could analyze the data from the assessment and speak to next steps in their instruction (this is more of a writing and reflection task since candidates do not actually have to complete these next steps). The candidate could score at or just below proficient on these rubrics, a combination of “2s and 3s,” and the embedded components that are scored in both planning and assessment, earning twenty-four more points. The overall score would be thirty-five. This candidate would pass the TPA, but not have demonstrated an ability to actually instruct. This example, hypothetical yet possible, illustrates that cut score use, and the specific cut score adopted, may not indicate readiness. Conversely, this example could indicate good preparation since the readiness traits that can be practiced during preparation coursework (planning, analyzing data, written reflection) were the areas in which the candidate performed well.

MTMM results clearly indicate a method effect and no discernable pattern between summative PR scores and TPA scores. There are several possible explanations for high MTMM and PR method-effect findings. These include:

1. *Inter-method traits are highly correlated*. This would not be a surprising result. Distinguishing traits in the performance evaluation of integrated practice is difficult. For instance, candidates teach lessons they have planned, and assess students on the content of the lessons they have taught. Therefore, correlations between planning, teaching, and assessment seem likely.
2. *Two systematic, reliable methods are random (method effect)*. In this case, it may be that both instruments measure something relevant about teacher readiness, but not the same

traits. It would therefore be possible for a candidate to score highly on one method of teacher readiness and poorly on another. The construct of teacher readiness is complex and, even when aligned with state and national standards, traits selected for measurement can differ. In addition, the focus or interpretation of scales on the instruments may lead to low shared variance rates. Twenty-one traits were aligned between the two methods but scores show almost no correlation.

3. *The PR scores were influenced by the TPA*. As discussed above, the TPA had a dramatic impact on the ST term and, especially, candidates' ability to solo-teach for a meaningful length of time. In addition, as reported above, the more experienced candidates, mentors, and supervisors were with the TPA, the more disenchanted they became (see VE 1). It is possible that the change in the ST procedures, in combination with evaluators' frustration with the TPA, impacted PR scores (also creating the observed ceiling effect).
4. *The strong method effect may also indicate that either one or both the PR or the TPA are not a highly credible method of determining teacher readiness*.

In addition, pass rate comparisons between university evaluations (mentor and supervisor) and TPA scores indicate differences that would fail three additional students and highly endorse one student for licensure who was not recommended by the university. Therefore, candidate performance may not indicate levels of proficiency across the method. This assumption is not plausible.

The fourth and fifth assumptions address the potential impact of various random errors associated with the conditions under which participants completed their TPA and will be argued together.

Assumption 4 and 5: The criteria, rubrics, procedures, and scores derived from the TPA are generalizable across different candidates and handbooks. TPA proficiency does not depend on factors beyond the candidate's control. The criteria, rubrics, procedures, and scores derived from the TPA are generalizable across testing sites, placements and placement length.

Evidence suggests that candidate scores are likely influenced by scorer assignment, placement site and placement length (elementary compared to secondary). Candidates with long-term placements in the same classroom with the same students have an advantage in understanding the context of the taught segment. For this reason, elementary candidates may always have an advantage over secondary candidates whose students typically change courses and teachers each term. Scores are impacted by the triad partnership, particularly the relationship with the mentor teacher. TPA scores did not prove to be as reliable (or better) and indicator of achievement of teacher readiness as mentor and supervisors observations of candidate readiness. Scores of readiness on the TPA do not always correlate to scores provided by mentors partnered with candidates.

Test scores across candidate groups were comparable. However, there were identifiable candidate groups who received unequitable treatment with respect to the assessment. These groups include candidates placed in settings that apply scripted curricula that align with TPA performance expectations and candidates partnered with a more willing mentor. An important requirement of generalization is procedural fairness, or whether the same rules are applied to everyone in the same way using the same procedures, and that testing modifications are applied, when required by law. In the case of the field test, modifications were not addressed nor did SCALE or PESB have adequate guidance in place to assist universities in making supports available for candidates. Similarly, while handbooks suggest more similarities in procedural fairness than differences, the actual rules of completion were applied differently (e.g., some candidates had access to resources, including human resources, whereas others did not). This is probably most clearly seen in Task 2 when some candidates secured a mentor-videographer. This is an example of a procedurally unfair scenario. Access to resources, human and otherwise, create differences that make scores less comparable. It is also conceivably the case that candidates from some universities have greater access to resources than candidates from other universities (i.e., state IHE compared to private IHE, urban IHE verses rural IHE). Scores reported at the state and national level, for accountability purposes, may not reflect these important differences and are therefore not generalizable. The performance tasks or objective differences between handbooks also suggests some challenges to equivalent procedures when candidates are provided materials for planning in some subject areas and not in others. It is understandable that candidates would view the completion of Task 1 as more difficult when not provided a text or curriculum while others utilize published lesson plans and aligned assessments. Evidence also suggests that those candidates that complete the TPA at the end of their ST experience will be at an advantage over candidates who submit earlier. Because of procedural fairness concerns and lack of modifications provided this assumption is not plausible.

Another fairness question is one of substance, or whether the test design and criteria are reasonable and no new content was introduced by the test, for the test. Or, if this is the case, that candidates were provided an opportunity to learn the new content in an equitable way before tested. In the case of AL and SV, evidence suggests that new content was a significant portion of the assessment (six rubrics) and that candidates differed in their abilities to both learn and apply these concepts in their classrooms prior to the taught segments. For instance, some candidates were placed in classrooms already applying AL and SV while others were in settings unfamiliar with these practices. As stated earlier, candidates placed in longer-term placements (year-long) have greater opportunities to learn the classroom context and its students than candidates in semester long placements. Knowledge of the context of the classroom is a significant factor for success in the construct, as operationalized by the TPA, and the opportunity to learn and practice new test content. Candidates in shorter placements may not have an opportunity to learn the classroom context and fully understand student needs before completion. Because of substantive fairness concerns this assumption is not plausible.

Finally, for scores to be generalizable across groups, it is important that funding sources are in place and appropriate to the testing purpose. Pearson did not collect a fee during the field test. However, as of January 2014, it is estimated that candidates will pay 300 US dollars per submission. IHE must find resources to fund candidate, mentor, supervisor, and faculty training, revisions of courses and materials, procurement of videography equipment, scorer training, TPA score interpretation for programmatic improvement, updated evaluations of faculty, and many other considerations. It is not clear how IHE resources will be found to support candidates in their TPA completion, to provide aid for candidates who cannot afford the fee, or potential resubmission fees. Given the importance of the testing purpose (high-stakes for licensure), funding sources should already be understood and in place prior to implementation. Almost a year after the field test, many of these uncertainties and funding concerns remain. This assumption is not plausible. It is not clear that all students receive equal opportunity to succeed on the TPA and for this reason, the entirety of assumption four and five are not plausible.

VA synthesis. The validity argument assumptions and evidence, organized by inference, are summarized in Table 5.1.

Table 5.1

Summary of the validity argument (Step 1a)

Inference	Warrant Justifying Assumptions	Support to Validity Evidence (VE)	Threats to Validity (Rebuttals)
Construct Representation: Teacher Readiness *Domain of performance: Effective Teaching*	The tasks in the TPA represent the domain of performance and skills of teacher readiness. *Taking these assumptions as given:* Effective Teaching is of interest to stakeholders and candidates Performance assessment on effective teaching has positive influence on curriculum for teacher preparation and readiness Performance assessment is the most valid and reliable way to measure teaching effectiveness and readiness.	VE 2	VE 1
Evaluation / Scoring (Target Domain) *Target Domain: Observed Performance is representative sample of performance domain (TPA definition)*	Observed performance on the TPA can be considered performance in target domain (representative sample of effective teaching and teacher readiness). Scores indicate whether candidates can perform adequately on the tasks presented to them. TPA scores reflect the level of teacher quality and readiness on the assessment tasks. Therefore, a low score on the TPA indicates that the candidate cannot perform the tasks included in the assessment (Kane, 1994) Criteria used to score the performance are appropriate and have been applied as intended. The performance occurred under conditions compatible with the intended score interpretation (judgment/decision/consequence) in terms of the candidates' level of skill.	VE 3 VE 4 VE 6	VE 4 VE 5 VE 6 VE 7 VE 8
Generalization *Universe of Target Domain*	The tasks/scores adequately sample and reflect performance on all possible and relevant tasks for target domain. Scores include a representative sample of performance from the target domain. Interpretation of scores will emphasize levels of skill in sub-domains or tasks. All candidates are provided equal opportunity to succeed. (Procedural fairness) The test is procedurally fair. The same rules are applied to everyone in more or less the same way. All test takers are consistently treated essentially the same way, using the same (or essentially the same; equivalent) procedures and rules. If modifications are necessary (or required by law), they are applied in ways that create an equitable testing procedure for all. (Substantive fairness). Test design and criteria are reasonable in the context of the test procedure. No substantively new content was introduced by the test, for the test, or if so, equity in opportunity to learn that content was provided before tested. Score interpretations and decision rules are reasonable and appropriate for all test takers and sub-groups. Funding sources are appropriate to testing purpose.		VE 9 VE 10 VE 11

Extrapolation *Target Domain AND Performance Domain*	The skills assessed are necessary (if not sufficient) for effectiveness in the performance domain. The knowledge, skills and judgment required on the performance assessment are essential for effective teaching and teaching readiness in real-world practice. As such an interpretation and use of the results supports decisions about candidate future teaching effectiveness. Interpretation of scores allows for a judgment about the candidate's readiness to perform as an effective teacher. Anyone who performs well on the assessment should also be able to perform well in the target domain. Proficient scores on the TPA tasks likely reflect teaching readiness. Anyone who performs poorly on the assessment should also perform poorly in the target domain. Poor scores on the TPA tasks likely reflect a lack of teaching readiness.	VE 12	VE 12 VE 13
Decision Making	Uses of scores are clear. Uses of scores are appropriate.		VE 14 VE 15

The evidence in Table 5.1 is further summarized in Table 5.2 by applying either a *rejected*, *accepted*, or *not studied* evaluation. Only one of the assumptions was accepted, two were rejected and two were accepted, with concerns.

Table 5.2

Evaluation status for each assumption in the interpretive argument (Step 1b)

Inference	Evaluation Status
Construct Representation	Accepted
Evaluation / Scoring	Accepted, with concerns
Generalization	Rejected
Extrapolation	Accepted, with concerns
Decision Making	Rejected

The summary shows that the preponderance of validity evidence presents a serious threat to the assumptions of the IA for Inferences 3 and 5 and some concerns about test score use based on inferences 2 and 4. This conclusion suggests that TPA test score use for licensure are not credible until the weak inferences of generalization and decision-making are addressed.

Validity Interpretations Revisited

The second step in the validity argument addresses the interpretation connected to each inference (see **Appendix W**). All inferences are aligned to the central interpretation: *Scores provide a* *measure of relevant teaching readiness.* To justify this interpretation, evidence should support the following four claims (see IA, Chapter one). The claim and evaluation appear below.

- **The TPA is relevant and aligned.** This claim is sound.
- **The TPA is fair.** There is inconsistent evidence for this claim. Improvements are needed to address both procedural and substantive fairness and funding concerns before decisions made from scores can be considered credible, especially for comparison across different programs, cohorts and groups.
- **The TPA is based on adequate levels of proficiency.** Some evidence suggests that there are different levels of proficiency between rubrics and across the construct. Findings imply that some rubric levels are not developmentally appropriate criteria for proficiency of novice teachers. The analysis of rubric language suggests that proficiency decisions may involve high levels of scorer judgment, introducing EoM. Finally, differences between university methods on candidate proficiency and TPA scores suggests a method effect. This claim is questioned.
- **The TPA is consistent.** Evidence suggests that TPA scores did not prove to be as reliable (or better) and indicator of achievement of teacher readiness as mentor and supervisors observations of candidate readiness. Scores of readiness on the TPA do not always correlate with scores provided by mentors partnered with candidates in ST. This claim is not sound.

Of these four claims, one is sound, one is defeated, and two are questioned. Evidence to support the assumptions underpinning these claims were similarly mixed. Because the consequences of test score use are high-stakes, the burden to provide a strongly credible measure is similarly high-stakes. For this reason, when assigning an evaluation status any error must be made in favor of the test-taker and test-user, with the intent of improving the instrument. Table 5.3 assigns an evaluation status (Verheij, 2005) to the inference as a whole: *justified*, *defeated*, or *unevaluated*.

Table 5.3

Evaluation status for the inference

Claim	Evaluation
How appropriate are the intended interpretation and uses of TPA test scores?	
Interpretation 1. Scores provide a measure of relevant teaching readiness.	Defeated

The study conclusion is that the proposed uses of TPA test scores are not yet credible because many of the assumptions and claims underpinning the IA cannot be verified by the evidence.

Implications

The implications of this research include four areas of particular relevance: 1. implications for the construct of teacher readiness; 2. importance of thoughtful implementation of performance-based assessment for beginning teachers; 3. evidence of the TPA as an opportunity for teacher growth and evaluation; 4. for practice and programmatic revision in teacher preparation; and 5. use of argument-based validation in performance assessment with high-stakes licensure consequences.

Identification of the teacher readiness construct

As was discussed in Chapter two, the TPA is not just intended to be an instrument to inform licensure decisions. It is also an instrument that is meant to redefine the way in which teaching readiness should be evaluated. If researchers are correct in predicting that the operationalized construct of the TPA will serve as a hidden curriculum for teacher preparation, it may work to change the way teaching is practiced in the field. However, it may also serve to limit the way teachers are prepared, potentially restricting the construct of teacher readiness (for instance, classroom management and rapport). Applying Kendall Stansbury (1998) predictions about performance assessment, focus on TPA preparation may become justification for removing differing sets of ideas from the discussion of teacher preparation. Therefore, a hidden curriculum of operationalized readiness in the TPA might rightfully be perceived as detrimental (see RQ1) to the theoretical construct articulated in Sterner's conceptual framework. In fact, it seems likely this is already happening. For instance, at Sterner, prior to the TPA, candidates took a course on "Classroom Management" and another course on "Assessment." Now, candidates take one course called "Assessment, Management, and Differentiation." By combining these courses, candidates now have more time in ST to focus on TPA completion. It will be important to carefully consider whether the way in which the TPA has operationalized the theoretical construct captures what it is to be an effective teacher, ready to enter a social, situated, and diverse classroom.

Importance of thoughtful implementation of performance assessment for beginning teachers

Fred Hamel's (2013) paper, "Assuring Quality or Overwhelming Teachers? High Quality Performance Assessment in American Pre-Service Teacher Education" collected both qualitative and quantitative data on candidate and mentor experiences with the TPA in WA. Hammel writes:

> Our data suggests that the TPA is a work in progress, but with characteristic blind spots. Without greater attention to the actual work conditions of teacher, to the multiple demands of student teaching and how these interact with the TPA, and to the intricate learning relationships between mentors and candidates-the TPA will be a bit of what we call "a bull in a china shop." (Hamel, 2012, p. 27)

Like measurement investigators, Hamel recommends more research especially because of some of the seemingly contradictory claims in the literature to date. He goes on to write, "Is the TPA 'assuring quality' or 'overwhelming teachers'? The answer seems to be 'yes'" (pp. 27-28). During the pilot of the TPA in WA, many professionals, psychometrians, and politicians referred to the field test as "constructing the plane while it flew." [41] It may be that, in the rush to implementation, careful

[41] The consortium of states that had adopted the TPA, called the TPA Consortium, created an informative website and cyberspace where people could share questions, concerns, and their practice. In 2013, this group was renamed the edTPA and linked to the AACTE website. Assessment procedures, what was

strategies that would have facilitated procedural and substantive fairness were not followed and this practice could be corrected. For instance, were IHE given enough time to determine curricular needs, implement new or revised course offerings, and encourage faculty involvement across the university in the negotiations? Did candidates, mentors and supervisors receive adequate training to understand the handbooks? Were candidates provided enough opportunity to learn and practice the construct in the teaching setting, prior to the assessment? What funding resources and supports would IHE have to secure to implement the TPA in their programs? One implication of this study is that IHE were not ready to field test the TPA.

Evidence of the TPA as an opportunity for teacher growth and evaluation

This study expands the scope of previous research by critically evaluating claims made by SCALE and AACTE such as the following: "the Teacher Performance Assessment was designed by teachers and teacher educators to *support candidate learning*" (AACTE, 2013, emphasis added). Novice teachers experience predictable developmental stages (Moir, 2013). While all teachers and teacher educators want only the best trained, most effective teachers working with young people, the importance of experience in becoming an effective teacher cannot be underestimated. As the mentor indicated, sometimes the solution is "simply experience" (3/30/12). TPA expectations are not always appropriate for candidates who are in a first-time-ever teaching context. Careful mentorship, guidance, and expectations are key to training a confident and effective educator. This is similar in other licensed professional fields. Experienced doctors, experienced lawyers, experienced CEOs mentor beginning professionals until they are ready to practice in a solo setting. It is not a matter of qualification, both groups are qualified. We must be careful to celebrate what novice teachers bring to their environments (the latest in research and best practice and innovative

"acceptable support," deadlines, submission requirements and language clarifications required regular TPA updates though the consortium.

strategies to perennial problems) and not to punish them for what they cannot bring, namely experience, by holding them accountable to developmentally inappropriate standards.

As the TPA is implemented across the nation, the purpose of the ST experience, as a mentored, formative learning experience for beginning teachers, must remain the focus. Studies demonstrate the negative impact of the “sink or swim” model, where candidates are dropped into classrooms with little support (Moir, 2012; Rockquemore, 2011; Picanco, Darragh, Tully, & Henning, 2011). The operationalization of the TPA as a high-stakes assessment, with procedures that isolate the candidate, perpetuates the idea of ‘teacher’ as isolated and separate instead of a member of a collaborative team of practioners. Because ST is formative, thoughtful implementation of the TPA must allow for candidate imperfections and learning mistakes, in order to be authentic and formative. Any instrument that focuses too much on the summative, and not enough on the formative nature of learning to teach, does not belong in the ST term. Rather, it belongs *after* the ST term.

Research on the PACT focuses on its dual purposes as both a formative and summative instrument. However, SCALE has removed the language about formative assessment from TPA (edTPA) handbooks. For this reason, PESB and SCALE need to reexamine the focus of the TPA as a summative, rather than formative, assessment if it is to be conducted during ST. The positive outcome of ST is a candidate who is ready to teach. Candidates do not enter ST ready to demonstrate ability. Rather, teaching readiness grows throughout ST and at unpredictable rates. Learning is not a linear process. Implementation strategies that mandate particular due dates, selected by IHE prior to the start of the school year or by Pearson due to scoring demands, have not been thoughtful about developmental assessment, development of novice teachers, or the application of learning theories.

Finally, many educational practitioners and researchers, who are not allowed to have access to the TPA handbook or the scoring materials, both of which are proprietary to SCALE and Pearson, suggest that an openness to feedback is mostly just talk. The implication of this study is that SCALE, Pearson, AACTE, CAEP, and PESB need to revisit best practices for implementation.

Implications for practice and programmatic revision in teacher preparation

An inevitable result of high-stakes, competency exams (whether traditional pencil-and-paper or performance based) is that, "to survive, teacher education programs must adjust their curriculum to prepare teacher candidates to pass the test" and this is most problematic, "if the tests are not valid in evaluating teacher effectiveness to start with, [because] they have the potential to shape teacher preparation programs in ways that are also invalid, thereby undermining recent efforts to achieve meaningful teacher education reform" (Goodman, Arbona, & Dominquez de Rameriz, 2008, p. 27). This is a concern for validation, "if the introduction of a new high-stakes test resulted in teaching to the test, and a narrowing of the taught curriculum, the impact would have significant implications for score meaning" in that, "it would imply that test scores, no matter how high, could not represent attainment across the full curriculum, since students would not be learning across the full curriculum" (P. Newton, 2012).

University conceptual framework models identify constructs for successful preparation of candidates through graduates who are not just ready for teaching but ready to lead, advocate, critically question, and collaborate in their fields, and other behaviors not measured by the TPA. Accreditation (accountability) of IHE programs involve TPA pass rates **and** evidence of having met criteria articulated by the conceptual framework of the university. TPA scores will not measure attainment of the full curriculum if one implication of the TPA is to reduce the teaching domain to only what is the operationalized construct of the TPA. The implication is that preparation programs must adjust curriculum to meet testing requirements. Candidates received significantly different adaptations for TPA support between the two programs at Sterner. For instance, undergraduate candidates received three on-campus writing days and TPA feedback from supervisors. Graduate candidates had two extra weeks to complete their TPA, including a spring break holiday week where there were no classes or teaching responsibilities. It seems likely that variances in the way in which guidance is interpreted between programs or universities will result in differences in the level of support provided to candidates to complete the TPA. Candidates whose professors, supervisors, and mentors assist in developing or reviewing their learning segments, who are allowed or provided external readers for comment, or who have access and assistance with videography, will do better on the assessment than candidates whose programs choose not (or cannot for resource reasons) provide similar levels of support. Candidates are likely to submit a stronger sample when:

- in long-term placements in the same classroom with the same students
- placed with SOE adjuncts, mentors who have scored TPA, or mentors who are NB qualified
- placed in settings that already employ AL and SV
- partnered with mentors who participate in video collection
- partnered with mentors who participate in planning the taught segment
- provided with dedicated writing time
- teaching with prepared materials already aligned to prepared assessments
- given due dates at the very end of the term and not earlier. However, evidence also suggests that the TPA can hijack the term and reduce candidate opportunities to solo teach. Programs will have to carefully consider the benefits and disadvantages of the TPA submission date.

These implications indicate that programmatic revision to support TPA completion, especially around placement practices, are required for candidates to have equal opportunities for success but these same practices can diminish the program's ability to teach and evaluate other significant traits in the construct.

Limitations

It is important to note that the results of this study represent the performance of a small population. While the sample size is considered strong for statistical analysis, it remains a small number of candidates from one IHE, in one state. Results might be different for other candidates, especially those in programs with more extensive ST lengths (year-long) or less extensive (6 weeks), or in states less familiar with performance examinations for licensure. Similarly, those programs that have not yet embedded field experience throughout a candidates' duration of study may find their results differ. Results, too, might differ for groups of candidates based on background levels (e.g., graduate verses undergraduate, SES status, minority, or gender differences), language competency levels (this sample included all native English speakers), access to and comfort with technological resources, and support networks embedded within preparation programs not reflected in the population of this study.

The validity argument in this study presents conflicting evidence. Weighing conflicting evidence is a difficult, but necessary, part of validity investigations, especially those that apply ABV and multiple methods. This study has applied the notion of "preponderance of evidence" to the analysis of qualitative data. Only those patterns that are repeated throughout phases of collection or by a significant majority of respondents could "out-weigh" conflicting quantitative data. The researcher weighed the data understanding that it was collected during a mandated field test of a new, high-stakes measure of readiness with very little guidance. Bias on the part of participants is likely to be high, and not in favor of implementation. It is important to keep in mind that these participants were told to adopt a new instrument during a high-risk, high-stress situation. While not consequential, the lack of feedback made the experience neither formative nor summative. This fact hardly compensated for the challenge posed to candidates while teaching the segment and completing the TPA. A limitation of this study is that sometimes evidence were not always conclusive in the support or threat to the assumption or inference.

As a result of data SCALE collected during the field-test, the TPA was revised and renamed the edTPA in 2013. There are some significant differences between the TPA and the revised edTPA that may influence the applicability of the findings in this study. As of January 2014, requests to SCALE for permission to view the edTPA had not been returned. Therefore, differences between the two test versions could not be discussed. However, it is likely that there are some limitations of this study to the revised version.

Suggestions for Future Research

TPA

These results provide evidence of a need to further examine the TPA to determine how differing populations of candidates perform, if measured criteria are an authentic representation of teacher readiness, and whether scores predict higher rates of teacher effectiveness. A large-scale, longitudinal study needs to be developed with diverse candidates, IHE, programs, and scorers from across all states adopting the TPA. Significant research must be done to support decisions made from scores, especially beyond the ST experience, and whether scores can be used to make accountability decisions about preparation. Test-maker assumptions are not always clear, even to test-makers, and the assumptions identified in this study would benefit from other confirmatory or competing examinations from multiple stakeholders (Ryan, 2002). In addition, this study did not focus on the validity issues involved in portfolio use or electronic submission and scoring. These add a further complexity to the test and additional assumptions to be studied, especially those of construct representation, evaluation and generalization.

Like many performance assessments for high-stakes decision-making, "Follow-up studies are needed to examine whether or not teacher candidates who score high on these authentic assessment measures used in teacher education programs actually emerge as 'highly trained' professionals after their teacher education programs have ended" (Goodman, Arbona, & Dominquez de Rameriz, 2008, p. 37). Future research on the TPA would benefit from continued practice of ABV. In particular, Bennett, Kane, and Bridgeman (2011) have applied ABV to formative and summative accountability systems (p. 2). An examination dissecting the formative and summative ends for licensure and for accountability of the TPA would be beneficial for understanding its purposes for decision-making.

This study addressed the consequences and uses of test scores for licensure and candidate accountability. It will be important that a future study address the separate inference of test score use for programmatic accountability. Preliminary evidence shared in this study may challenge the practice of using TPA test scores for decision-making around teacher preparation and program accreditation. Whether scores provide a measure of relevant accountability of teacher preparation should be more carefully studied before consequences are implemented.

ABV

High-stakes performance assessment must serve the same purposes as all high-stakes assessment (Haertel, 1999; Sackett, 1998). The TPA as a large-scale, high-stakes exam with multiple interpretations (both accountability and licensure), claims, inferences and assessments, demonstrated the usefulness of the ABV model for exams of this nature. Chapelle, Enright, and Jamieson (2010) consider the ways in which Kane's ABV approach (2006) differs from the *Standards* (1999). Chapelle (2011) identified Kane's contribution for performance testing for language use. This study confirms similar contributions for the ABV model for performance testing for teaching readiness. These include:

- Association between test and domain: The IA identifies the domain, the sample, and the universe of generalization. Test makers and validators must distinguish between the inferences relevant to score meaning and evidence collected is used to support assumptions or improve the method.
- Scoring: The evaluation inference addresses scoring, training, and criteria relevant to meaning. Test makers and validators need to include assumptions that underlie the scoring procedures and scorer training.
- Construct definition: The construct can be articulated, defined, and included in the IA. This requires test-makers and validators who define constructs to include underlying assumptions and use research and evidence to support the construct defined.
- Consequences and uses: The IA assumes that implications and consequences of test score use are part of validation studies. The chain of argumentation provides a structure to examine these. Test-makers and validators need to "specify the intended decisions and consequences of test score use" (Chapelle, 2011, p. 20).

- Stakeholders: Validity arguments, like other argumentative structures, should be clear and persuasive and recognize all stakeholders. Test-makers and validators can use ABV to structure the examination of the plausibility of and clearly explain the argument for all the intended audience(s) (Chapelle, 2011).

However, ABV is not a quick or simple technology to validate test score consequence and uses. Though ABV provides a framework for analysis that offers a structure for decision-making, the inquiry process is complex. Establishing the assumptions and inferences that underpin the assessment is time-consuming and messy and made more so by the use of multiple data collection methods. A strength of Kane's model is that the framework necessitates that multiple types of evidence be used. However the practice of balancing the range of evidence to address the IA requires that investigators manage a vast array of data and examine their own biases toward "convincing evidence" and what will count as disconfirmatory (Chapelle, 2011).

Kane has described ABV as a two-step process when, in fact, the process of ABV in this study was circular. Step one, the interpretive argument, and step two, the validation argument, were constantly repeated. Step two simply informs another round and revision of step one. This practice is essential because the IA functions to make clear what test scores mean (see Chapter two). Conclusions, therefore, are always tentative because new evidence and insights cause changes in the IA and the cycle repeats.

In this study, development of the IA was a backward-planning process that ended in explicit inferences and assumptions and began with a host of deceptively simple questions. For instance:

1. What is the goal of the TPA?
2. Who are the stakeholders?
3. What does it mean to be an effective teacher? What behaviors or assessment tasks would indicate readiness for effective teaching?
4. What does it mean to say that a performance assessment is "fair"?

These questions were “grouped” together using Kane’s framework of Evaluation, Generalization, and Extrapolation. Because the construct of “readiness” and the domain of “effectiveness” are complex and integrated tasks, many questions posed demanded that Construct Representation also be evaluated. In addition, questions of fairness originally led to the inclusion of a sixth inference which was later merged with Generalization (see **Appendix I**).

While the interpretive argument is (necessarily) evolving, it must be specific enough to guide a manageable plan to address the key challenges of validity. In practice, this give and take of simultaneously creating and adapting an IA while also engaging in the study can prove overly-complicated and time consuming. The nature of the method involves collecting evidence to address each inference. However, inferences are often inter-connected and separating evidence for each is a complex step. It is difficult to plan in a way that can eliminate redundancies, wasted efforts, or hasten the pace. Because ABV assumes that validity is a never-ending jigsaw employing both a big-picture argumentative structure and puzzle piece evidence, detangling that process is not as simple in practice as Kane’s theory presents. For instance, Table 3.3 outlines the study design which includes twenty-seven unique sources of evidence. However, not all of the sources of evidence were ultimately used or helpful. Table 3.32 provides a listing of the evidence actually used to address the inferences. One of the strengths of ABV is that it can be used for any type of validity investigation and any type of assessment. The cost of the flexibility of such a broad structure is the lack of set procedures in how to apply it for specific inquiries. Like candidates and faculty describing the TPA, a researcher first using ABV might feel they are “constructing the plane while flying it.”

ABV are judged based on criteria of clarity, coherence and plausibility (Kane, 2009). Recently, Newton (2013; 2010) and Newton and Shaw (2014) have questioned whether ABV requires two separate arguments because the validity argument implies the development of interpretations. The IA specifies the reasoning and questions used to draw conclusions and to make decisions. The VA is the process of relating evidence to the IA. The best challenge to ABV is less a challenge to the VA (assuming that rigor and standards in data collection and analysis have been followed) but the establishment of defensible and plausible alternative interpretations to the IA. When constructing the IA, the type of arguments, and the process of development, are distinct from that of the VA (Kane, 2004; 2002). Though both are a part of what I have described as a circular process, each is a distinct step in that process.

Unlike other frameworks or validity approaches that begin by establishing a research design, ABV begins from the stance that validity processes should start with score meaning (Chapelle, Enright, & Jamieson, 2010, p. 6). The network of inferences developed set out to provide a clear rationale for the interpretation and use of TPA scores and a coherent reasoning to support use of candidate performance for licensure decisions that would be both persuasive and plausible for all stakeholders, but primarily those stakeholders for whom the TPA is most high-stakes (Kane, 2011). What ABV demands of the researcher is to take a step away from the research process to define terms, articulate complex relationships between inferences, assumptions, interpretations, and develop what is defensible in the domain and testing practice. This is an analytical and philosophical exercise and it results in the IA. The IA then places expectations and demands on the research design, validation process, properties of the test and testing procedures. Like Kane, I would suggest that without the separate step of the IA, it becomes all too easy for important claims and assumptions to be left implicit and to focus a study on data collection and analysis without a real yardstick with which to measure consequences and uses and, ultimately, score meaning. This has never been truer than in our current “assessment culture” where we find test-makers and stakeholders rushing to implement a test and planning to study it later. Separate from undertaking the validity study is the communication of the study process, evidence, and findings. Imbedded in ABV is that conclusions should be clear and meaningful for stakeholders. Identifying a separate reporting scaffold, such as that used by Shaw & Crisp (2012), to guide the structure of the evidence presentation augments the ABV framework in the writing stage.

Conclusion

Performance assessments are complex. The construct of teacher readiness and quality is multi-faceted. Licensure exams are high-stakes but vital to maintain the integrity of any professional discipline. Test score consequences derived from validation inquiries are influenced by an unlimited host of factors. The evidentiary reasoning process of the ABV model made it possible to address aspects of validity for an assessment with a complicated operationalized construct and to examine assumptions that involved different types of evidence with sometimes contradictory results. ABV confirmed that a coherent and unified set of procedures can assist researchers in making decisions about test score meaning, consequences and use (Kane, 2001).

Appendix A

Teacher Preparation Standards

The following appendices are the common standards for teacher preparation for licensure and certification in the US and the specific standards in Washington state.

Appendix A1

INTASC Standards

These 10 INTASC standards focus on a “Common Core of Teaching Knowledge” in the following nine domains:

1. Learner Development and Learning Differences
2. Learning Environments
3. Content Knowledge
4. Application of Content
5. Assessment
6. Planning for Instruction
7. Instructional Strategies
8. Professional Learning and Ethical Practice
9. Leadership and Collaboration

Within each standard, specific knowledge, disposition, and performance criteria are articulated in three quality levels or “progression indicators” (CCSSO, 2013).

Appendix A2

National Board Professional Teaching Standards (Certification Standards)

The Five Core Propositions form the foundation and frame the rich amalgam of knowledge, skills, dispositions and beliefs that characterize NB Certified Teachers (NBCTs)" (NBPTS, 2013). The Five Core Propositions are:

1. Teachers are committed to students and their learning.
2. Teachers know the subjects they teach and how to teach those subjects to students.
3. Teachers are responsible for managing and monitoring student learning.
4. Teachers think systematically about their practice and learn from experience.
5. Teachers are members of learning communities.

For each of these propositions, NB identifies knowledge, disposition and performance criteria that candidates for boards must meet.

Appendix A3

Washington Professional Educator Standards Board

PESB framework for effective teacher preparation for initial residency license were revised in 2010 and are called "Standard V." Standard V includes three components:

5.A. Effective Teaching

5.B. Professional Development

5.C. Teaching as a Profession

Each component includes sub-criteria assessed as unmet, met, or exemplary (PESB, 2013).

Appendix B

Overview of WA TPA Assessment – Assessment Tasks

WA TPA Task	What to Do	Scoring Rubrics	What to submit
Planning Instruction and Assessment	✓ Provide relevant information about your instructional context. ✓ Select a learning segment of sequential lessons[42] that develop students' competencies and knowledge to meet the standards and performance task. ✓ Determine what content and related academic language you will emphasize. ✓ Consider your students' strengths and needs and select a central focus and a key language demand for the learning segment. ✓ Provide opportunity for students to articulate the learning target(s), monitor their own progress, and identify resources needed to achieve the learning target(s). ✓ Create an instruction and assessment plan for the learning segment, and write lesson plans. ✓ Respond to commentary prompts to describe your students and teaching context, and explain your thinking in developing the plans and how they reflect what you know about your students as well as research/theory. ✓ Make daily notes about the effectiveness of your teaching for your students' learning (will be used in writing the Analyzing Teaching commentary in Task 4).	Planning (1, 2, 3)	□ Part A: Context for Learning Information □ Part B: Lesson Plans for Learning Segment ✓ Lesson Plans ✓ Instructional Materials ✓ Assessment tools and criteria □ Part C: Planning Commentary
Instructing and Engaging Students in Learning	✓ Collect permission forms from parents/guardians and prepare for video-recording. ✓ Review and identify lessons where you are engaging your students in understanding concepts. ✓ Submit video clip following handbook guidelines on length and quantity. ✓ Respond to commentary prompts to analyze your teaching and your students' learning in the video clip(s).	Instruction (4 and 5)	□ Part A: Video Clip(s) □ Part B: Instruction Commentary

[42] The length of the learning segment depends on how frequently you teach the same students. If daily, then it is 3-5 lessons. If weekly, it is 3-4 lessons. If in a block schedule, it is 3-5 hours of instruction.

WA TPA Task	**What to Do**	**Scoring Rubrics**	**What to submit**
Assessing Student Learning	✓ Analyze class performance from one assessment completed during the learning segment. ✓ Identify **three** student work samples that illustrate trends in student understanding within the class. ✓ Select and analyze the learning of **two** focus students in more depth, based on both work samples and related student-voice evidence articulating their own learning. Document your feedback on their work ✓ Respond to commentary prompts to report conclusions from your analysis and describe feedback given to the two focus students. ✓ Identify next steps in instruction based on your analysis and on student articulation of their learning. ✓ Provide the assessment task and evaluation criteria.	Assessment (6, 7, 8)	□ Part A: Student Work Samples □ Part B: Evidence of Feedback □ Part C: Assessment Commentary
Analyzing Teaching	✓ Using notes you have recorded throughout the learning segment, respond to commentary prompts to explain what you have learned about your teaching practice and two or three things you would do differently if you could teach the learning segment over. Explain why the changes would improve your students learning.	Analyzing Teaching (9)	□ Analyzing Teaching Commentary
Academic Language	✓ Select one key language demand related to the central focus. Explain how you will support students with varied language needs. ✓ Cite evidence of opportunities for students to understand and use the targeted academic language in 1) the video clip(s) from the Instruction task; OR 2) the student work samples from the Assessment task. ✓ Analyze the effectiveness of your language supports.	Academic Language (10,11,12)	□ Planning Commentary (Prompt 4) □ Instruction Commentary (Prompt 4) □ Assessment Commentary (Prompt 4)
Student Voice	✓ Explain how you will give students opportunities to express their understanding of the learning targets, identify resources to support and monitor their own learning progress, use student voice to raise awareness of where they are relative to the learning targets. ✓ Provide examples from the video clip(s) of strategies to elicit student voice of their understandings of the learning targets. ✓ Collect and analyze student reflections on their progress toward meeting the learning target(s), describe how you helped the two focus students understand their progress toward the learning targets, and use the reflection to inform your next steps in instruction.	Student Voice (13,14,15)	□ Planning Commentary (Prompts 1, 3e, 5 c-d) □ Video Clip(s) □ Student Reflections □ Assessment Commentary (Prompts 1b-c, 2b, 3c)

Appendix C

Description of TPA Categories, Tasks, Rubrics and Evidence Required

6 Categories	4 Tasks	15 Traits (Rubrics)	8 Evidences
Planning Instruction and Assessment		R1: Planning for [Content specific] Understandings R2: Using Knowledge of Students to Inform Teaching and Learning R3: Planning Assessments to Monitor and Support Student Learning	Lesson Plans, including assessments Commentary
Instructing and Engaging Students in Learning		R4: Engaging Students in Learning R5: Deepening Student Learning	Video Clips Commentary
Assessing Student Learning		R6: Analyzing Student Work R7: Using Feedback to Guide Further Learning R8: Using Assessment to Inform Instruction	Scored Student Work Samples Commentary
Analyzing Teaching		R9: Analyzing Teaching Effectiveness	Commentary
Academic Language	Embedded	R10: Understanding Students' Language Development and Associated Language Demands R11: Scaffolding Students' Academic Language and Deepening Content Learning R12: Developing Students' Academic Language and Deeping Content Learning	Lesson Plans & assessments Scored Student Work Samples Commentary

Student Voice (Washington Only)	Embedded	R13: Eliciting Student Understanding of Learning Targets R14: Supporting Student Use of Resources to Learn and Monitor their own Progress R15: Reflecting on Student-Voice Evidence to Improve Instruction	Lesson Plans & assessments Scored Student Work Samples Student Reflections Commentary

Appendix D

Research Timeline

Phase	**Pre TPA**	**During TPA Taught Segment**	**Post Taught Segment / Writing UP**	**Post Submission**
Timing	**February**	**March**	**Late March**	**April and May**
Data Sources	• Interviews with Case Study candidates • Survey to all candidates • West E scores for correlation • Weekly Self-Reflection statements (411) • Survey for supervisor and mentor • Observe in candidates block classes (prior to full-time student teaching)	• Visit Case Study placement site and observation during taught unit • Candidates lessons • Weekly Self-Reflection statements (411) • Case study interviews • Case study observation debrief and reflection meeting • Video consent form	• Observe writing up phase for cohort (those participating). These three ½ day sessions will be digitally recorded. • Weekly Self-Reflection statements (411) • Supervisor statements and notes • Survey for supervisor and mentor • Survey of Teacher Candidate experiences • Case Study interviews • Collect Case Study TPA Samples	• Survey to be completed immediately after submission • Interviews with Case Study • Group interviews with 10+ (2 groups) • TPA scores correlation • Supervisor and mentor evaluations of candidates • Survey for supervisor and mentor • Professional Growth Plan for cohort • Weekly Self-Reflection statements (411) • University Faculty survey • University Faculty group interview • University Supervisor Group Interview (2 groups of approx. 8-12)

Appendix E
Surveys

The following appendices are the electronic surveys sent to study participants. Surveys are provided by phase. All recipients received a link to the survey via their email address. This link was individualized to that only that participant through a program called Survey Monkey.

Appendix E1

TPA Pre-Experience Survey for Teacher Candidates

Directions: The following questions ask you to explain what you know and believe about the Teacher Performance Assessment (TPA) and how ready you feel to complete this assessment during your student teaching term. This survey should not take you more than 20 minutes. Thank you!

***1. Please provide your student ID number.**

***2. Are you male or female?**

Male

Female

3. Which category below includes your age?

17 or younger
18-20
21-29
30-39
40-49
50-59
60 or older

4. Are you White, Black or African-American, American Indian or Alaskan Native, Asian, Native Hawaiian or other Pacific islander, or some other race?

White
Black or African-American
American Indian or Alaskan Native
Asian
Native Hawaiian or other Pacific Islander
From multiple races
Some other race (please specify)

***5. In which program are you enrolled?**

MIT
TED
ETC

***6. What do you hope to learn about the teaching profession during student teaching?**

***7. As you enter the student teaching term, in which teaching responsibilities do you feel most uncertain?**

***8. As you enter the student teaching term, in which teaching responsibilities do you feel most confident?**

***9. How excited are you to begin the student teaching term?**

Which level of best captures your feelings?	Cannot wait!	Excited	Somewhat Excited	Not really excited

***10. What else could your program provide to help you be most successful in student teaching?**

***11. Choose your TPA subject area from the drop-down menu.**

***12. Have you attended a TPA orientation session?**

Yes

No

Other (please specify)

***13. How useful was the orientation session?**

Very

Somewhat

Not useful

***14. What could your program provide to support and assist you in understanding the TPA requirements?**

***15. Identify your level of understanding of the requirements of the Teacher Performance Assessment**

	Confused	Some Understanding	Fully, no questions
Task 1: Planning instruction and assessment			
Task 2: Instructing and Engaging Students in Learning			
Task 3: Assessing Student Learning			

Task 4: Analyzing Teaching			
Academic Language			
Student Voice			

∗16. In your own words, describe the requirements of the TPA:

∗17. Will the TPA help you grow as a teacher?

Yes

No

Other (please specify)

∗18. Describe the reason for your answer to the above question.

∗19. Please rank the following responses. I will complete the TPA because:

	(1) Most important reason	(2)	(3)	(4)	(5) least important reason
It is a professional development opportunity					
It is required for certification in WA state					
I want to challenge myself as a professional					
I will use the TPA to determine my strengths and areas for improvement as a future teacher					
It is required coursework in the SOE					
Other (please specify)					

∗20. Based on what you know about the TPA and your preparation as a teacher, which response best captures your view of the TPA?

I expect to meet standard

I expect to exceed standard

I am unsure how well I will meet standards

I am unsure of the standards

Other (please specify)

Appendix E2

Pre-Experience Survey for Supervisors

Directions:
The following questions ask you to explain your experience supervising or mentoring candidates during their student teaching term. In particular, this survey will focus on the process of preparing and completing the Teacher Performance Assessment (TPA). This survey should not take you more than 10-15 minutes to complete. Thank you!

1. What is your first name?

***2. What is your last name?**

***3. What is your gender?**

Male

Female

Undeclared

***4. Are you a National Board Certified Teacher?**

Yes

No

Currently in the process

Not sure

***5. In which University program are you mentoring/supervising?**

Master in Teaching

Teacher Education Department (Traditional Undergraduate)

Evening Teacher Certification

Both Master in Teaching and Traditional Undergraduate

Not sure

***6. Which description best captures your role?**

Mentor or Cooperating Teacher

University Supervisor

Other (please specify)

∗7. How many years have you supervised student teachers?

This is my first year

2-5

5-10

More than 10

Not sure

8. Describe any changes you have made in your supervision of candidates this year/semester?

∗9. Identify which TPA you are supervising this term (which TPA will your candidates complete?).

Elementary Literacy

Elementary Mathematics

Secondary Mathematics

Secondary English Language Arts

Secondary Social Studies

Secondary Science

World Languages

Physical Education

Music

Art

Don't know

All of them

Other (please specify)

∗10. Can portfolio based performance assessments such as the TPA accurately predict candidate readiness to teach?

Yes

No

Not sure

∗11. What is the best evidence of candidate readiness to teach?

∗12. What is the worst evidence of candidate readiness to teach?

***13. Rank the following responses. A candidate is ready to student teach when they:**

	Essential	Important	Somewhat Important	Not Important	Unsure	N/A
Understand the material presented						
Practice good classroom management						
Develop instruction aligned to standards						
Analyze assessment data to determine next steps with students						
Engage learners in critical thinking and meaningful dialog						
Reflect on their performance with an eye to determining strengths and weaknesses						
Encourage and utilize student voice to make decisions about instruction, assessment, and student needs						
Practice differentiation						
Develop and implement instruction that focuses on the content language needs of their students						
Communicate and collaborate with parents and colleagues						
Is culturally proficient and responsive to the cultural needs of the students in their classroom						
Other (please specify)						

***14. Is student teaching important to our profession?**

Yes

No

Not Sure

***15. Explain your answer to the above question (# 14).**

***16. Identify how well you understand the TPA requirements.**

	I understand the requirements	I feel confident but still have questions	I am confused about the requirements	I don't understand the requirements
Task 1: Planning instruction and				
Task 2: Instructing and Engaging				
Task 3: Assessing Student Learning				
Task 4: Analyzing Teaching				
Academic Language				
Student Voice				

***17. As you understand it, describe the requirements of the TPA.**

***18. Have you attended a TPA orientation session?**

Yes

No

Not sure

19. If you answered "Yes" to the question above, what in the orientation session was most beneficial or helpful as you learned about the TPA?

***20. Have you attended a scorer training or scored TPA?**

Yes

No

Not sure

21. If you answered "Yes" to the question above, what in the scorer training session was most beneficial or helpful as you learned about the TPA?

***22. Have you attended any other training where the TPA and your role was discussed?**

Yes

No

Not sure

23. If you answered "Yes" to the question above, what in the scorer training session was most beneficial or helpful as you learned about the TPA?

***24. Based on what you know about the TPA, which of the following statements best reflects your view. The TPA is:**

a good instrument measuring candidate readiness to teach

a good instrument but may not indicate candidate readiness to teach

not a good measurement

I am unsure

I do not understand the TPA well enough to have a view

Other (please specify)

***25. Has the Teacher Performance Assessment changed your mentoring experience?**

Yes

No

Not sure

26. If so, describe how.

***27. Do you anticipate that the TPA will change the mentoring experience?**

Yes

No

Not sure

28. If so, how do you predict it will change your experience mentoring candidates?

***29. How much time have you spent learning about the TPA?**

None

Less than one hour

1-2 hours

3-6 hours

6+ hours

***30. How much time have you spent discussing the TPA with each of your candidates?**

None

Less than one hour

1-2 hours

3-6 hours

6+ hours

∗31. Describe any changes you have made so that your candidates can complete the requirements of the TPA.

32. Is there anything you would like the University to know about your experience so far this semester?

33. What can the University provide to help support your work with our candidates this semester?

34. Would you like more information on becoming a TPA scorer?

∗35. Do you prefer that your comments remain anonymous?

Yes

No

I am unsure. Please contact me.

Appendix E3

Pre-Experience Survey for Mentors

Directions:
The following questions ask you to explain your experience supervising or mentoring candidates during their student teaching term. In particular, this survey will focus on the process of preparing and completing the Teacher Performance Assessment (TPA). This survey should not take you more than 10-15 minutes to complete. Thank you!

1. What is your first name?

***2. What is your last name?**

***3. What is your gender?**

Male

Female

Undeclared

***4. Are you a National Board Certified Teacher?**

Yes

No

Currently in the process

Not sure

***5. At what level are the majority of the students in your classroom performing:**

At grade level

Above grade level

Below grade level

Not sure

Other (please specify)

6. What constraints, if any, are placed on your instruction in the school district or at theschool where you teach? (For instance: scripted curriculum, no computer lab, no librarian).

7. What additional constraints, if any, are placed on the teacher-candidate's student teaching experience at your school?

***8. In which [University] program are you mentoring/supervising?**

Master in Teaching

Teacher Education Department (Traditional Undergraduate)

Evening Teacher Certification

Both Master in Teaching and Traditional Undergraduate

Not sure

***9. Which description best captures your role?**

Mentor or Cooperating Teacher

University Supervisor

Other

Other (please specify)

***10. To date, how many candidates have you mentored during student teaching?**

This is my first candidate

2-5

5-7

More than 7

Not sure

***11. Before this semester, how recently have you hosted a candidate for student teaching?**

last year

two to three years ago

three to five years ago

more than five years ago

12. Describe any changes you have made in order to mentor a candidate this semester?

***13. Describe the teacher-candidate with whom you are partnered this semester.**

***14. Choose your candidate's TPA subject area from the drop-down menu.**

***15. Did you attend the pairs training session at the start of this term?**

Yes

No

Not sure

***16. Describe the ways in which you have implemented the co-teaching strategies. With my candidate this spring, I have:**

tried them all

prepared several lessons to practice a few of the strategies

only just learned of the co-teach strategies in pairs training but plan to use a strategy in the next two weeks

I need to learn more about co-teaching before I implement it in my class(es)

I do not use co-teaching strategies

Other (please specify)

∗17. Can portfolio based performance assessments such as the TPA accurately predict candidate readiness to teach?

Yes

No

Not sure

∗18. What is the best evidence of candidate readiness to teach?

∗19. What is the worst evidence of candidate readiness to teach?

∗20. Rank the following responses. A candidate is ready to student teach when they:

	Essential	Important	Somewhat Important	Not Important	Unsure	N/A
Understand the material presented						
Practice good classroom management						
Develop instruction aligned to standards						
Analyze assessment data to determine next steps with students						
Engage learners in critical thinking and meaningful dialog						
Reflect on their performance with an eye to determining strengths and weaknesses						
Encourage and utilize student voice to make decisions about instruction, assessment, and student needs						
Practice differentiation						
Develop and implement instruction that focuses on the content language needs of their students						
Communicate and collaborate with parents and colleagues						
Is culturally proficient and responsive to the cultural needs of the students in their classroom						
Other (please specify)						

***21. Is student teaching important to our profession?**

Yes

No

Not Sure

***22. Explain your answer to the above question (# 21).**

***23. Identify how well you understand the TPA requirements.**

	I understand the requirements	I feel confident but still have questions	I am confused about the requirements	I don't understand the requirements
Task 1: Planning instruction and				
Task 2: Instructing and Engaging				
Task 3: Assessing Student Learning				
Task 4: Analyzing Teaching				
Academic Language				
Student Voice				

***24. As you understand it, describe the requirements of the TPA.**

***25. Have you attended a TPA orientation session?**

Yes

No

Not sure

26. If you answered "Yes" to the question above, what in the orientation session was most beneficial or helpful as you learned about the TPA?

***27. Have you attended a scorer training or scored TPA?**

Yes

No

Not sure

28. If you answered "Yes" to the question above, what in the scorer training session was most beneficial or helpful as you learned about the TPA?

***29. Have you attended any other training where the TPA and your role was discussed?**

Yes

No

Not sure

30. If you answered "Yes" to the question above, what in the scorer training session was most beneficial or helpful as you learned about the TPA?

***31. Based on what you know about the TPA, which of the following statements best reflects your view. The TPA is:**

a good instrument measuring candidate readiness to teach

a good instrument but may not indicate candidate readiness to teach

not a good measurement

I am unsure

I do not understand the TPA well enough to have a view

Other (please specify)

***32. Has the Teacher Performance Assessment changed your mentoring experience?**

Yes

No

Not sure

33. If so, describe how.

***34. Do you anticipate that the TPA will change the mentoring experience?**

Yes

No

Not sure

35. If so, how do you predict it will change your experience mentoring candidates?

***36. How much time have you spent learning about the TPA?**

None

Less than one hour

1-2 hours

3-6 hours

6+ hours

***37. How much time have you spent discussing the TPA with your candidates?**

None

Less than one hour

1-2 hours

3-6 hours

6+ hours

***38. Describe any changes you have made so that your candidate can complete the requirements of the TPA.**

39. Is there anything you would like the University to know about your experience so far this semester?

40. What can the University provide to help support your work with our candidates this semester?

41. Would you like more information on becoming a TPA scorer?

***42. Do you prefer that your comments remain anonymous?**

Yes

No

I am unsure. Please contact me.

Appendix E4

During TPA Survey for Teacher Candidates (Phase 2)

∗1. Please provide your student ID number.

∗2. List the first words that come to mind when you read "Teacher Performance Assessment"

3. Share anything you would like about your experience with the TPA so far.

∗4. Rank the following responses. I am completing the TPA because:

	(2) Most important reason	(2)	(3)	(4)	(5) least important reason
It is a professional development opportunity					
It is required for certification in WA state					
I want to challenge myself as a professional					
I will use the TPA to determine my strengths and areas for improvement as a future teacher					
It is required coursework in the SOE					
Other (please specify)					

∗5. Can portfolio based performance assessments such as the TPA accurately predict candidate readiness to teach?

Yes

No

Not sure

***6. Where are you in the process of completing the TPA? Select all that apply:**

Creating the lesson plans

Teaching the learning segment

Done teaching the learning segment

Writing the commentaries

I am done with my TPA

Other (please specify)

7. Describe the steps or process you used to complete the TPA requirements.

***8. When creating your TPA learning segment, did you utilize any methods or strategies that you learned about in your coursework (coursework should be considered classes you took prior to full-time student teaching)?**

Yes

No

Other (please specify)

9. If so, please list the course/method/strategy you used in your TPA lessons that you applied from your coursework.

***10. Did you find yourself preparing any differently for the TPA lessons than any other lessons you have taught?**

Yes

No

Other (please specify)

11. If so, please describe how the TPA lessons were different.

***12. How many lessons did you include in your learning segment?**

Three

Four

Five

Other (please specify)

***13. Would you characterize your TPA lessons as "typical" lessons in your classroom?**

Yes

No

Other (please specify)

14. If not, discuss how your TPA lessons are different from those you would typically teach in your classroom.

***15. Are there any pedagogical strategies or routines that are important to you that could not be implemented during your TPA learning segment?**

Yes

No

Other (please specify)

16. If so, can you describe what you felt you had to omit?

***17. Is there anything that you changed in your practice (planning, instruction, assessment, reflection) to meet the TPA requirements that you have continued to do in your non-TPA lessons?**

Yes

No

Other (please specify)

18. Please describe or list these.

***19. Did you experience any unexpected challenges or situations that impacted your TPA.**

Yes

No

Other (please specify)

20. If so, please describe the challenges or unexpected situations.

21. In what ways might the TPA reflect your preparation as a teacher?

***22. Did you utilize any co-teaching methods during your learning segment?**

Yes

No

Not sure

23. If so, describe the strategy(ies) you used in your learning segment.

***24. How many of your TPA lessons did you videotape?**

One

Two

Three

Four

All

∗25. After videotaping your lesson(s), did you opt to use the video from the lesson you had intended to be the Task 2 lesson?

Yes

No

I have not gotten that far yet

Other (please specify)

26. Why or why not?

∗27. How many of your students did you have permission to include in your video clip?

None

Close to 50%

More than 50%

All but a few (2-5 students)

All

Other (please specify)

∗28. How did you manage (what did you decide to do with) students who did not have permission to be on video in your lesson(s)?

∗29. Has your work on the TPA helped you to grow as a teacher?

Yes

No

Other (please specify)

∗30. How or why?

∗31. How confident are you that your work on the TPA will met standard (level 3 or above)?

Very confident

Confident

Somewhat confident

Not confident

Other (please specify)

∗32. Why do you think so?

***33. To date, how much time have you spent preparing the TPA (do not include time spent teaching)?**

I've not started

3-5 hours

5-10 hours

10-15 hours

More than 15 hours

Not sure

Other (please specify)

***34. How much time have you spent discussing the TPA with your mentor teacher and/or university supervisor?**

Less than one hour

1-2 hours

3-6 hours

6+ hours

***35. How much time have you spent discussing the TPA with university faculty?**

Less than one hour

1-2 hours

3-6 hours

6+ hours

***36. Did your mentor teacher review your TPA lessons?**

Yes

No

Other (please specify)

37. What feedback did you receive?

***38. Did your supervisor review your TPA lessons?**

Yes

No

Other (please specify)

39. What feedback did you receive?

***40. Did a faculty member review your lessons?**

Yes

No

Other (please specify)

41. What feedback did you receive?

42. As you taught your learning segment, what did you learn about yourself as a teacher?

∗43. What else could your program provide to help you be most successful in student teaching?

∗44. Using the criteria provided, rank the responses. As a result of my experience working on the TPA, my growth in the following areas was:

	None	Not significant	Significant	Very significant	N/A
Understand the material presented					
Practice good classroom management					
Develop instruction aligned to standards					
Analyze assessment data to determine next steps with students					
Engage learners in critical thinking and meaningful dialog					
Reflect on their performance with an eye to determining strengths and weaknesses					
Encourage and utilize student voice to make decisions about instruction, assessment, and student needs					
Practice differentiation					
Develop and implement instruction that focuses on the content language needs of their students					

Communicate and collaborate with parents and colleagues					
Is culturally proficient and responsive to the cultural needs of the students in their classroom					
Other (please specify)					

***45. Based on what you know about the TPA, which of the following statements best reflects your view. The TPA is:**

a good instrument measuring candidate readiness to teach

a good instrument but may not indicate candidate readiness to teach

not a good measurement

I am unsure

Other

Other (please specify)

***46. What advice do you have for candidates completing their learning segment next semester?**

47. Is there anything else you would like the University to know about your experience so far this semester?

Appendix E5

During TPA Task 1 Draft for Teacher Candidates (Undergraduate only) (Phase 2)

Directions: After completing your draft of Task 1, please take a moment to complete this survey. This survey should take approximately 15 minutes and should be submitted by midnight, March 2nd.

∗1. Describe the steps and process you used to complete the draft of your Task 1 requirements?

∗2. Would you characterize your TPA lessons as "typical" lessons in your classroom?

Yes

No

Other (please specify)

3. If not, discuss how your TPA lessons are different from those you would typically teach in your classroom.

∗4. How much time have you spent preparing Task 1?

I've not started

1-3 hours

3-5 hours

5-10 hours

More than 10 hours

Not sure

Other (please specify)

∗5. When planning your learning segment lessons, did you utilize any methods or strategies that you learned about from your coursework (coursework should be considered classes you took prior to full-time student teaching)?

Yes

No

Other (please specify)

6. If so, please list the course/method/strategy you used in your TPA lessons that you learned in coursework.

***7. Has completing the first task of the TPA helped you to grow as a teacher?**

Yes

No

Other (please specify)

∗8. Describe the reason for your answer to the above question.

***9. Has your mentor teacher reviewed your TPA lessons?**

Yes

No

Other (please specify)

***10. Has your supervisor reviewed your TPA lessons?**

Yes

No

Other (please specify)

***11. What advice do you have for candidates next semester for preparing for their TPA Task 1?**

12. Please share anything you would like about your experience with the TPA so far.

***13. How confident are you that the Task 1 draft you submitted on March 2nd will meet standard (level 3 or above)?**

Very

Somewhat

Not confident

Other (please specify)

***14. Identify any unexpected challenges or situations that have impacted your TPA.**

Yes

No

Other (please specify)

15. If so, please describe the challenges or unexpected situations.

16. As you prepare to teach your learning segment, what have you learned about yourself as a teacher?

17. What can the University provide to help support your work on the TPA?

Appendix E6

During TPA Writing Day 1 for Teacher Candidates (Undergraduate only) (Phase 2)

Thank you for your hard work and effort working on your TPA today. Please take a moment to answer this short 11 question survey. Don't forget, by the end of the week, to EMAIL [researcher] the drafts of your commentaries. These do not need to be the exact draft you plan to submit for scoring but what you have completed by Friday. Thank you!

***1. Where are you in the process of completing the TPA? Select all that apply:**

Writing the commentary for Task 1

Writing the commentary for Task 2

Writing the commentary for Task 3

Writing the commentary for Task 4

Editing my video

I am done with my TPA

Other (please specify)

***2. Where did you opt to work during your writing time today? Choose all that apply**

Library Computer lab

My home

My school

Other (please specify)

***3. What did you work on during your writing time today?**

***4. Do you feel you were productive during the writing time provided today?**

Yes

No

Other (please specify)

***5. As you worked, about which TPA task were of your questions?**

Task 1

Task 2

Task 3

Task 4

I did not have questions

Other (please specify)

***6. Can you describe these questions?**

7. Has a particular task of the TPA been more challenging? If so, which one and why?

∗8. What have you learned about the practice of teaching from your work on the TPA today?

∗9. What else can the University provide to help support your work this week?

10. If you found an error in your handbook, can you please note it here:

11. Is there anything else you would like to share about your experience today?

Appendix E7

During TPA Writing Day 2 for Teacher Candidates (Undergraduate only) (Phase 2)

Thank you for your hard work and effort working on your TPA today. Please take a moment to answer this short 11 question survey. Don't forget, by the end of the week, to EMAIL [researcher] the drafts of your commentaries. These do not need to be the exact draft you plan to submit for scoring but what you have completed by Friday. Thank you!

***1. Where are you in the process of completing the TPA? Select all that apply:**

Writing the commentary for Task 1

Writing the commentary for Task 2

Writing the commentary for Task 3

Writing the commentary for Task 4

Editing my video

I am done with my TPA

Other (please specify)

***2. Where did you opt to work during your writing time today? Choose all that apply**

Library Computer lab

My home

My school

Other (please specify)

***3. What did you work on during your writing time today?**

***4. Do you feel you were productive during the writing time provided today?**

Yes

No

Other (please specify)

***5. As you worked, about which TPA task were of your questions?**

Task 1

Task 2

Task 3

Task 4

I did not have questions

Other (please specify)

***6. Can you describe these questions?**

7. Has a particular task of the TPA been more challenging? If so, which one and why?

∗8. What have you learned about the practice of teaching from your work on the TPA today?

∗9. What else can the University provide to help support your work this week?

10. If you found an error in your handbook, can you please note it here:

11. Is there anything else you would like to share about your experience today?

Appendix E8

During TPA Writing Day 3 for Teacher Candidates (Undergraduate only) (Phase 2)

Thank you for your hard work and effort working on your TPA today. Please take a moment to answer this very short 11 question survey. Don't forget, by the end of the week, to EMAIL [researcher] the drafts of your commentaries. These do not need to be the exact draft you plan to submit for scoring but what you have completed by Friday. Thank you!

***1. Where are you in the process of completing the TPA? Select all that apply:**

Writing the commentary for Task 1

Writing the commentary for Task 2

Writing the commentary for Task 3

Writing the commentary for Task 4

Editing my video

Preparing to submit

I am done with my TPA

Other (please specify)

***2. Where did you opt to work during your writing time today? Choose all that apply**

Library Computer lab

My home

My school

Other (please specify)

***3. What did you work on during your writing time today?**

***4. Do you feel you were productive during the writing time provided today?**

Yes

No

Other (please specify)

***5. As you worked, about which TPA task were of your questions?**

Task 1

Task 2

Task 3

Task 4

I did not have questions

Other (please specify)

***6. Can you describe these questions?**

***7. If you could go back and change something about your TPA learning segment, would you?**

Yes

No

Maybe

Other (please specify)

***8. What would you change about your TPA learning segment?**

***9. Are there any aspects of your TPA about which you are uncertain will meet standard?**

Yes

No

Maybe

Other (please specify)

10. Please describe these.

***11. Are there any aspects of your TPA that you expect will exceed standard (rubric level 3)?**

Yes

No

Maybe

Other (please specify)

12. Please describe these:

***13. How confident are you that your work on the TPA will meet standard (level 3 or above on the rubrics)?**

	Not Confident	Somewhat Confident	Confident	Very confident
Task 1				
Task 2				
Task 3				
Task 4				
Academic Language				
Student Voice				
Other (please specify)				

14. Is there anything else you would like to share about your experience?

Appendix E9

Post-TPA Submission Survey for Teacher Candidates (Phase 3)

Congratulations on your TPA Submission! This survey (the 3rd of 4 this term) has 23 questions and should take approximately 20 minutes. Your input will help us better understand our programs and will be shared with others in to better the assessment system for pre-service teachers. Thank you for sharing your time and your thoughts!

***1. Please provide your student ID number.**

***2. Describe the purpose of the TPA.**

***3. List your top two strengths as a teacher.**

***4. List the top two areas you want to improve upon as a teacher.**

***5. Identify if and when you offered your students behavioral incentives (for instance, candy, points, recess time) in order to complete your TPA:**

	Yes	No	Not Sure
Permission slips			
Video clips			
Student work (assessments)			
Other (please specify)			

***6. Between the teaching of your learning segment and the submission of your TPA, did you modify your:**

	Yes	No	Not sure
context for learning			
lesson plan(s)			
instructional material(s)			
assessment tool(s) and procedure(s) used			
feedback on student work			
academic language			
student voice			
Other (please specify)			

***7. Has the TPA changed the way you think about teaching?**

Yes

No

Maybe

Other (please specify)

∗8. Did TPA measure your effectiveness as a teacher?

Yes

No

Maybe

Other (please specify)

∗9. How did the TPA measure your effectiveness?

∗10. Did TPA measure your impact on student learning?

Yes

No

Maybe

Other (please specify)

∗11. How did the TPA measure your impact on learning?

∗12. Rank the following responses. The TPA offered me an opportunity to grow in the following areas:

	Did not offer growth	Some Growth	Growth	Major Growth	N/A
Understand the subject area (literacy, math, science) of the learning segment					
Practice good classroom management					
Develop instruction aligned to standards					
Analyze assessment data to determine next steps with students					
Engage learners in critical thinking and meaningful dialogue					
Reflect on my performance with an eye to determining my strengths and weaknesses					

Encourage and utilize student voice to make decisions about instruction, assessment, and student needs					
Practice differentiation					
Develop and implement instruction that focuses on the content language needs of my students					
Communicate and collaborate with parents and colleagues					
Be culturally proficient and responsive to the cultural needs of the students in my classroom					
Other (please specify)					

***13. All together, how many hours have you spent completing the TPA (do not include time spent teaching)?**

Less than 5 hours

5-10 hours

10-20 hours

20-30 hours

30-40 hours

More than 40 hours

I don't know

Other (please specify)

***14. Refer to the following definitions to answer this question:**

Summative Assessment: testing for accountability
Formative Assessment: testing for learning improvement

Which of the following statements is true? In my experience, the TPA was:

a formative assessment

a summative assessment

both formative and summative

neither formative nor summative

I don't know

Other (please specify)

***15. Based on my learning and professional development, the TPA was worth the effort.**

True

False

Maybe

Other (please specify)

***16. How confident are you that your work on the TPA will meet standard (level 3 or above on the rubrics)?**

	Not Confident	Somewhat Confident	Confident	Very confident
Task 1				
Task 2				
Task 3				
Task 4				
Academic Language				
Student Voice				
Other (please specify)				

***17. If you could go back and change something about your TPA would you?**

Yes

No

Maybe

Other (please specify)

18. If you answered "Yes" please describe what you would change.

***19. What do you wish you knew before starting the TPA (that you know now)?**

∗20. Identify the level of challenge you experienced completing the following requirements of the Teacher Performance Assessment (include all aspects of the process from start to submission):

	Very Challenging	Challenging	Somewhat challenging	Not challenging
Task 1: Planning instruction and assessment				
Task 2: Instructing and Engaging Students in Learning				
Task 3: Assessing Student Learning				
Task 4: Analyzing Teaching				
Academic Language				
Student Voice				
Other (please specify)				

∗21. Based on your completion of the TPA, which of the following statements best reflects your view. The TPA is:

a good instrument measuring candidate readiness to teach

a good instrument but may not indicate candidate readiness to teach

not a good measurement

I am unsure

Other

Other (please specify)

22. What else could your program provide to help you be most successful in student teaching?

23. Is there anything else you would like the University to know about your experience?

Thank you for your time and sharing your thoughts with us!

Appendix E10

Post –TPA Submission Survey for Mentors (Phase 3)

Directions: The following questions ask you to explain your experience mentoring a candidate during their student teaching term. In particular, this survey will focus on the process of preparing and completing the Teacher Performance Assessment (TPA). This survey has 24 questions and should not take you more than 15 minutes to complete. Thank you!

***1. What is your first name?**

***2. What is your last name?**

***3. In which [University] program are you mentoring/supervising?**

Master in Teaching

Teacher Education Department (Traditional Undergraduate)

Not sure

4. Describe any benefits you or your students have received by hosting a teacher-candidate this semester.

5. Describe any challenges you or your students have experienced by hosting a candidate this semester.

***6. Which best describes your candidate's performance so far this term?**

Doing harm

Struggling

Emergent

Meeting my expectations

Successful

Surpassing my highest expectations

Other (please specify)

***7. List two of your candidate's greatest strengths as a teacher.**

***8. List two of your candidate's areas for continued growth as a teacher.**

***9. In comparison to past terms, have you changed your mentoring practice this semester?**

Yes

No

Maybe

This is my first semester as a mentor

Other (please specify)

10. Please describe the changes you have made in your mentoring practice.

∗11. List the first words that come to mind when you read "Teacher Performance Assessment"

∗12. Based on your experience mentoring during the TPA, which of the following statements best reflects your view. The TPA is:

a good instrument measuring candidate readiness to teach

a good instrument but may not indicate candidate readiness to teach

not a good measurement

I am unsure

Other

∗13. What was YOUR experience, as the mentor teacher, with the Teacher Performance Assessment?

∗14. Did you review TPA lessons?

Yes

No

Other (please specify)

15. What feedback did you provide?

∗16. How much time have you spent working on the TPA with your candidate?

None

Less than one hour

1-2 hours

3-6 hours

6+ hours

∗17. Identify if you or your candidate offered your students behavioral incentives (for instance, candy, points, free time) in order to complete the TPA:

	Yes	No	Not sure
Permission slips			
Video clips			
Student work (assessments)			
Other (please specify)			

∗18. Refer to the following definitions to answer this question:

Summative assessment: testing for accountability
Formative assessment: testing for learning improvement

Which of the following statements is true? For my candidate, the TPA was:

a formative assessment

a summative assessment

both formative and summative

neither formative nor summative

I don't know

Other (please specify)

∗19. When is the best time in the student teaching semester to ask candidates to complete the TPA?

EARLY in the term so a candidate can address any area of failure before the end of the term

At or by the MIDTERM point so we can move on to full-time student teaching as soon as possible

AFTER the midterm when candidates have a had a chance to fully phase-in

As LATE as possible in the term

Other (please specify)

∗20. Which of the following best captures your view of the type of support a candidate will need from a mentor teacher to complete their TPA (check all that apply):

Logistical aid (collecting permission slips, accessing student records, video-taping lessons)

Assistance understanding the classroom context and student needs

Review of lesson plans

Co-teaching

Waiting to phase-in until the TPA is submitted

Willingness to adjust classroom routines or style to meet TPA requirements

Willingness to adjust curriculum to meet TPA requirements

Time away from the classroom to complete the TPA

The same support as a candidate who is not completing a TPA

No support is needed

Other (please specify)

21. Do you have any suggestions for ways we can improve our support for candidates while they complete their TPA?

22. What could [University] provide to support and assist you as a mentor teacher?

23. Is there anything else you would like the University to know about your experience mentoring our candidates?

***24. Do you prefer that your comments remain anonymous?**

Yes

No

I am unsure. Please contact me.

Appendix E11

Post –TPA Submission Survey for Supervisors (Phase 3)

Directions:
The following questions ask you to explain your experience supervising and mentoring candidates during their student teaching term. In particular, this survey will focus on the impact of the Teacher Performance Assessment (TPA). This survey should not take more than 15 minutes to complete. Thank you!

∗1. How many teacher-candidates are you currently supervising?

This was my first candidate.

Between 2 and 5 candidates.

Between 5-10 candidates.

More than 10 candidates.

∗2. In which [University] program are you mentoring/supervising?

Master in Teaching (MIT)

Teacher Education Department (Traditional Undergraduate/TED)

Evening Teacher Certification (ETC)

Both Master in Teaching and Traditional Undergraduate

Not sure

∗3. Identify how well you understand the TPA requirements.

	I understand the requirements	I feel confident but still have questions	I am unsure about the requirements	I don't understand the requirements
Task 1: Planning instruction and assessment				
Task 2: Instructing and Engaging Students in				
Task 3: Assessing Student Learning				
Task 4: Analyzing Teaching				
Academic Language				
Student Voice				

***4. Have you attended a scorer training?**

Yes, through [University]

Yes, through Pearson

Yes, through another university or organization

No

Not sure

***5. To what extent did the TPA develop your candidates' knowledge of the teaching profession?**

Very much

Somewhat

Very little

Not at all

Other (please specify)

***6. How much time have you spent discussing the TPA with each of your candidates?**

None

Less than one hour

1-2 hours

3-6 hours

6+ hours

7. Describe any positive impact of the TPA on the student teaching term this semester.

8. Describe any challenges or difficulties with the TPA this term.

***9. Did your role as a supervisor change as a result of the TPA?**

Yes, for the better

Yes, for the worse

No

This is my first semester/year supervising

Not sure

10. What about your role has changed?

11. Is there anything you would like the University to know about your experience supervising this semester?

12. What can the University do to help support your work with our candidates?

***13. Do you prefer that your comments remain anonymous?**

Yes

No

I am unsure. Please contact me.

Thank you for completing this survey.

Appendix E12

Post-ST Survey for Teacher Candidates (Phase 4)

Directions: The following questions ask you to reflect on your current teaching experiences which may, or may not, have been affected by the completion of the TPA. Please answer with the response which most closely mirrors your experience as a result of completing the TPA. Thank you!

***1. Please provide your student ID number.**

***2. To what extent did the TPA**

	Very much	Somewhat	Very little	Not at all	N/A
challenge you to reflect upon your teaching practice					
develop your abilities to adjust your teaching practice to the students your classroom					
refine your ability to plan subject specific lessons					
shape your teaching knowledge through the completion of the rationale (commentary) portions					
influence your analysis of student work to inform instruction					
influence habits of reflection					

***3. To what extent did the inclusion of the TPA affect your:**

	Improve	Somewhat for the better	Not at all	Somewh at for the worse	Harm/ Worsen	N/A
relationship with your mentor teacher						
relationship with your supervisor						

relationship and engagement with students						
knowledge of the teaching profession						
engagement with the curriculum/ content matter taught during student teaching						
sharing of primary responsibility for planning and teaching subjects or sections						
ability to practice co-teaching methods						
Current teaching practice						
Other: please specify						

***4. To what extent did the TPA**

	Very much	Somewhat	Very little	Not at all	N/A
influence how you think about using assessment to inform instruction					
encourage you to learn about students in your classroom					
shape your habit to reflect upon your teaching practice					
refine your ability to analyze student work					
encourage collaboration with other teachers when faced with an instructional challenge					
develop your ability to incorporate and use voice data					
encourage a habit of planning for student language needs					

∗5. When is the best time in the student teaching semester to ask candidates to complete the TPA?

EARLY in the term so a candidate can address any area of failure before the end of the term

At or by the MIDTERM point so we can move on to full-time student teaching as soon as possible

AFTER the midterm when candidates have a had a chance to fully phase-in

As LATE as possible in the term

I am not sure

Other (please specify)

6. Describe any positive impact of the TPA on your student teaching term.

7. Describe any challenges or difficulties with the TPA this term.

∗8. Please complete the following statement: When I think about the TPA I feel or I wish

9. Is there anything else you would like to share about your experience with the TPA?

10. Is there anything else you would like to share about your experience student teaching?

∗11. May we contact you after you receive your scores for a follow up survey?

Yes

No

Not sure, contact me later

Thank you for participating!

Appendix E13

Post-ST Survey for Mentors (Phase 4)

Directions: The following questions ask you to explain your experience mentoring a candidate during their student teaching term. In particular, this survey will focus on any potential impact of the Teacher Performance Assessment (TPA). This survey should not take you more than 15 minutes to complete. Thank you!

∗1. To what extent did the inclusion of the TPA affect:

	Improved	Somewhat for the better	Not at all	Somewhat for the worse	Harm/ Worsen	N/A
your role as a mentor teacher						
your relationship with the teacher candidate						
the candidate's relationship and engagement with your students						
the candidate's knowledge of the teaching profession						
the candidate's engagement with the totality of the content matter taught during their time in your classroom						
the release and/or sharing of primary responsibility for planning and teaching subjects or sections						
your ability to practice co-teaching methods						
Other (please specify)						

∗2. Which best describes your overall assessment of your candidate's readiness to be an effective first year teacher? (Your response will NOT be factored into any evaluation of or shared with the candidate)

Will exceed expectations for beginning teacher (highly accomplished)

Solid foundation in knowledge and skills for effective teaching

Acceptable level to begin teaching effectively

Some skill but needs more practice to be a teacher of record

Will struggle, not ready to teach

Not sure

Other (please specify)

3. Describe any positive impact of the TPA on the student teaching term this semester.

4. Describe any challenges or difficulties with the TPA this term.

∗5. Did your candidate have a positive impact on the students in your classroom?

Yes

No

Not sure

Other (please specify)

∗6. Which description best captures your feelings about hosting a future teacher candidate during the completion of their TPA?

Now more than ever it is important that I host a candidate

The TPA did not impact my thinking about future mentoring

Somewhat, I am still thinking it over

I do not plan to host another candidate because of the TPA

I will wait to decide whether to host another candidate until the TPA is fully understood

Not sure

Other (please specify)

7. What suggestions do you have that would help us improve the student teaching experience for you, as the mentor teacher?

8. Is there anything else you would like to share about your experience as a mentor teacher this year/semester?

∗9. Are you willing to participate in a mentor teacher group interview to discuss the impact of the TPA on the student teaching term?

Yes

No

Maybe

∗10. Do you prefer that your comments remain anonymous?

Yes

No

I am unsure. Please contact me.

Appendix E14

Post-TPA Survey for Faculty (Phase 4)

Directions: The following questions ask you to describe your experience as a [University] instructor during the initial implementation of the Teacher Performance Assessment. This survey should not take more than 15 minutes to complete. Thank you!

∗1. Did you teach a course for [the] University School of Education this year?

Yes

No

∗2. Did you teach a course that addressed the TPA?

Yes, 1 course

Yes, more than 1 course

No

Not sure

3. How many of the courses you taught this year addressed the TPA?

***4. How much total class time have you devoted to discussing the TPA?**

No time

Less than 1 hour

1-3 hours

3-5 hours

More than 5 hours

Other (please specify)

∗5. In the 2011-2012 academic year, how much time have you spent on TPA related teaching activities such as revising syllabi, modifying instructional materials, and/or responding to candidate inquiries?

None

Less than 1 hour

1-5 hours

5-10 hours

10-20 hours

20-40 hours

More than 40 hours

Other (please specify)

6. Describe any changes you have made to your course(s) to address the TPA.

7. What, if any, instructional activities or concepts have you removed or given up in your courses to accommodate the TPA?

***8. During the 2011-2012 academic year, how much time have you spent on your own professional development for the TPA (activities such as reading, thinking, preparing, or training)?**

None

Less than 1 hour

1-5 hours

5-10 hours

10-20 hours

20-40 hours

More than 40 hours

Other (please specify)

9. If teacher-candidates contacted you to ask questions about the TPA while they were completing the assessment, please describe the type of questions you were asked (e.g., assessment, student voice, academic language, standards and alignment).

***10. Have you read any of the the current TPA handbooks (revised in November and distributed to faculty in January)?**

Yes

No

Not sure

∗11. Identify how well you understand the TPA requirements.

	I don't understand the requirements	I am unsure about the requirements	I feel confident but still have questions	I understand the requirements
Task 1: Planning instruction and assessment				
Task 2: Instructing and engaging students in learning (video component)				
Task 3: Assessing student learning				
Task 4: Analyzing teaching				
Academic language				
Student voice				

12. What could be provided to support your work preparing candidates for the TPA?

∗13. Have you received any TPA scoring training?

Yes, from [University]

Yes, from Pearson

Yes, from another institution

No

Not sure

∗14. Are you willing to serve as a internal TPA scorer for pre-service teacher undergraduate (TED, ETC) or masters programs (MIT) (not through Pearson)

Yes

Yes, if it is part of my course load or I am compensated

No

Maybe

∗15. Refer to the following definitions to answer this question:

Summative assessment: testing for accountability

Formative assessment: testing for learning improvement

Which of the following statements best captures your view of the TPA as an assessment tool?

a formative assessment

a summative assessment

both formative and summative

neither formative nor summative

I don't know

Other (please specify)

∗16. Based on what you know about the TPA, which of the following statements best reflects your view. The TPA is:

a good instrument measuring candidate readiness to teach

a good instrument but may not indicate candidate readiness to teach

not a good instrument

I am unsure

I do not understand the TPA well enough to have a view

Other (please specify)

∗17. Do you prefer that your comments remain anonymous?

Yes

No

I am unsure. Please contact me.

Thank you for completing this survey.

Appendix F

Case Study and Group Interview Protocols

The following appendices are the study interview protocols used for case study and group interviews. Protocols are provided by phase. Due to IRB requirements at Sterner, interviews were conducted by the both an investigator and a second interviewer using the protocols developed by the study investigator.

Appendix F1

Pre TPA Phase - Case Study First Interview (Phase 1)

Dear Investigator: This initial interview is semi-structured and should take approximately 30 minutes for each candidate. Please videotape or audiotape your conversation and submit it to Kaitlyn Rebbe. A flip cam has been reserved for your continual use for both February and March and can be collected from Sarah Seidel.

Ask each of the questions below (some may not be answerable at this point depending upon the candidate's experiences and readiness. If the interviewer opted not to ask a question, indicate why a question was skipped) and any appropriate follow up or additional questions.

Collect Candidate Data

Candidate:	Program: □ MIT □ TED
Placement Site:	Mentor Teacher:
University Supervisor:	Student ID#:
TPA Subject Area: □ Elementary Literacy □ Elementary Mathematics □ Secondary Mathematics □ Secondary English Language Arts □ Secondary Social Studies □ Secondary Science	

□ World Languages

□ Music

□ Art

□ PE

Describe the case study process and the investigator’s role in the research and (non)role in the evaluation of the candidate. Reiterate that the investigator will observe lessons but does not replace the mentor or the supervisor in the student teaching term.

Interview Questions

Student Teaching and Placement Questions:

1. Describe your ideal teaching experience.
2. Tell me about your relationship with your mentor teacher.
3. Tell me about your student teaching classroom and your students.
4. What would you say are your greatest strengths as a teacher?
5. What are the areas you want to improve upon as a teacher?

Teacher Performance Assessment Questions:

6. When did you first hear about the TPA? What were your initial impressions?
7. As you understand it, describe the requirements of the TPA.
8. Have you attended a TPA orientation session?
    a. What in the orientation session was beneficial or helpful?
    b. Which requirements of the TPA do you feel ready to begin?
    c. How do you feel about the TPA requirements and scoring process?
    d. Do you have any questions about the TPA, in general?
9. Tell me about your TPA.
    a. Where are you in the preparation process for the TPA?
    b. What is the big idea for your learning segment?
    c. What will your lessons cover?
    d. When will you teach your segment?
    e. Are you using a unit or lessons that you created during coursework?
    f. Have you talked with your mentor teacher about the TPA? What was that conversation like?
10. How do you think you might do on the TPA?
11. In what ways might the TPA reflect your preparation as a teacher?
12. Do you think the TPA will help you grow as a teacher? Why or why not?

Appendix F2

During TPA Phase - Case Study Second Interview (Phase 2)

Dear Investigator: This mid-process interview is semi-structured and should take approximately 30 minutes for each candidate. Please videotape or audiotape your conversation and submit it to [research assistant]. Ask each of the questions below (some may not be answerable at this point depending upon the candidate's completion of certain steps of the TPA. If the interviewer opted not to ask a question, indicate why a question was skipped) and any appropriate follow up or additional questions.

Collect Candidate Data

Date:	Candidate:	Student ID#:

Interview Questions

General:

1. Please share anything you would like about your experience with the TPA.
2. What are the first words that come to mind when you hear "TPA"?
3. Where are you in the process of completing the TPA? Describe the steps and process you used to complete the requirements.
4. How much time have you spent preparing the TPA?

Constraints:

5. Describe any changes you have made to your instruction in order to teach your TPA?
6. What constraints, if any, are placed on your instruction in the school district or at the school where you teach?
7. Identify any unexpected challenges or situations that have impacted your TPA.
8. Would you characterize your TPA lessons as "typical" lessons in your classroom? If not, discuss how your TPA lessons are different from those you would typically teach in your classroom.
9. Are there any pedagogical strategies or routines that are important to you that you felt you could not implement in your TPA lessons?

Support and Prep:

10. When planning your learning segment lessons, did you utilize any methods or strategies from your coursework (coursework should be considered classes you took prior to full-time student teaching)?
11. Did anyone review your TPA lessons? What feedback did you receive?

Video:

12. How many of your TPA lessons did you video tape? Why did you make that decision?
13. Which lesson in the learning segment did you intend to be your Task 2 lesson?
14. What did you decide to do with students who did not have permission to be on video in your lesson(s)?

Impact:

15. After preparing for the TPA, is there anything that you have continued to practice in other lessons as a result of the TPA requirements?

16. Has completing the TPA helped you to grow as a teacher? Why or why not?
17. How confident are you that your work on the TPA will meet standard (level 3 or above)?
18. What can the University provide to help support your work on the TPA?
19. What advice do you have for candidates next semester for preparing for their TPA?

Logistics:

We will have two more interviews (writing up and post submission). The next interview will be during the writing up phase. When will you write your commentaries? Let’s schedule our next interview appointment.

Appendix F3

Writing Up TPA Phase - Case Study Third Interview (Phase 3)

Dear Investigator: This interview is semi-structured and should take approximately 30 minutes for each candidate. Please videotape or audiotape your conversation and submit it to [research assistant]. Ask each of the questions below (some may not be answerable at this point depending upon the candidate's completion of certain steps of the TPA. If the interviewer opted not to ask a question, indicate why a question was skipped) and any appropriate follow up or additional questions.

Collect Candidate Data

Date:	Candidate:	Student ID#:

Interview Questions

General:

1. How would you describe your student teaching experience so far this semester?
2. As a student teacher, do you feel you are meeting the expectations of your program? Why or why not?
3. Where are you in the process of completing the TPA commentaries? Describe how you approached writing up the commentaries.
4. How much time have you spent writing the commentaries for the TPA?
5. Has a particular task of the TPA been more challenging? If so, why?
6. Which tasks in the TPA do you feel were **most** helpful? Why?
7. Which tasks in the TPA do you feel were **least** helpful? Why?

Assessment:

8. Describe the process you used to collect and analyze student work in your learning segment.
9. Is this process typical of how you review student work in non-TPA lessons?
10. Describe the assessment you used for Task 3. Why did you select that assessment?
11. Is this assessment typical of the kind of assessments you use in your classroom? Why or why not?
12. What type of feedback did you provide your students on this assessment?
13. Is that typical of the feedback you generally provide students?

Reflection:

14. Did the TPA give you an opportunity to examine your teaching practice? Why or why not?
15. What did you learn **about the practice** of teaching from the TPA?
16. What have you learned **about yourself** (as a teacher) as a result of completing the TPA?
17. In the lessons since the learning segment, what have you continued to do (planning, instruction, assessment, reflection) as a result of the TPA?
18. Has writing the commentaries for the TPA helped you to grow as a teacher? Why or why not?

Evaluation:

19. Are there any aspects of your TPA about which you are uncertain will meet standard?
20. Are there any aspects of your TPA that you expect will exceed the standard?
21. How confident are you that your work on the TPA will meet standard (level 3)? (SCALE: 1: Not-5: Very)

Program Advice:

22. What can the University provide to help support your work on the TPA?
23. What advice do you have for candidates next semester for writing their commentaries?
24. Is there anything else you would like to share about your experience with the TPA at this time?

Logistics: We will have one more interview. The next interview will occur after you submit the TPA. Let’s schedule our next interview appointment.

Appendix F4

Case Study Final Interview (Phase 4)

Dear Investigator: This interview is semi-structured and should take approximately 30 minutes for each candidate. Please videotape or audiotape your conversation and submit it to department secretary (before May 11) or Suzie Henning (after May 11).

Collect Candidate Data

Date:	Candidate:	Student ID#:

Interview Questions

General:

1. How do you feel about your student teaching experience, as a whole?
2. What experiences in your student teaching semester were most beneficial (helped you grow the most)? (Please ask candidates to be specific and to limit their examples to this semester.)
3. How prepared do you feel to be an effective teacher next year?

   (If candidates struggle offer these potential choices: 1. Highly prepared, 2. Solid foundation in knowledge and skills, 3. Acceptable level to begin teaching effectively, 4. Some skill but I need more practice, 5. Struggling, not ready to teach)

   a. What would you say are your greatest strengths as a teacher?
   b. What are the areas you want to improve upon as a teacher?
   c. Are there any correlations between the strengths/weaknesses you have identified and your expectations for your TPA scores/performance?

TPA: "At different points in the semester, we have asked you to consider the TPA based on your progress completing the assessment. Now, for the following questions, we ask you to think about your experience with the TPA as a whole."

4. Describe any positive impact of the TPA on your student teaching term.
5. Describe any challenges or difficulties with the TPA.
6. Has the TPA impacted your feelings about teaching? If so, how?

TPA Feedback:

7. Do you feel like the TPA balanced the dual concerns of teacher-accountability and teaching for learning (summative and formative assessment)? Why or why not?
8. Have you received feedback on your TPA?
   a. If so, what did you learn about yourself as a teacher from the feedback you received?
   b. If not (or in addition) what type of feedback would have been beneficial to you?

Program Advice:

9. At what point does the TPA belong in the preparation program? Where should teacher candidates be introduced to, prepared for, and complete the assessment?
10. What is NOT addressed in the TPA that is still important for teacher preparation?
11. If you had known what you know now about the TPA, would you have made any different choices about teaching?
12. Is there anything else you would like to share about your experience with the TPA?

Logistics:

13. May we contact you after you receive your scores for a follow up interview?
14. May we contact you after your first year as a teacher for a follow up interview?
15. Once it is completed, would you like us to share a copy of our research project?

Thank you so very much for sharing your ideas and your time with us.

Appendix F5

TPA Group Interview – Teacher Candidates

Dear Investigator: This is a structured interview that should take approximately 30 minutes. You have been partnered with 4-7 candidates who agreed to participate in the interview and for their responses to be a part of our research. Audiotape your conversation using the Recorder Pro application and submit the recording to [R.A.] for transcription and storage.

Prior to the interview, determine whether you want to call on candidates or have candidates volunteer and arrange the seats in a circle or semi-circle. During the interview, ask each of the questions in the order in which they appear. Do not move to the next question until each candidate has responded or asked to be skipped. The number of questions in this interview is based on an average of a one-minute response per student, per question. You will want to monitor time to make sure that each respondent has time to share. Please advise a note-taking device (number the students or use their initials) and take notes during the interview on the order of the respondents in answering the questions for transcription purposes. (For instance, Q1: RB, JD, SH, JK, NB; Q2: NB, SH, JK, RB, JD, etc.)

Introductions:

- Please start promptly at 4:00.
- Introduce yourself and your role in the project.
- Explain the purpose of the interview:

> "This group interview is part of a larger project to study teacher candidate experiences with the TPA. This interview will take approximately 30 minutes. Your name will not be attributed to your comments in our published research. I will start by asking a question and then we will progress around the circle/volunteers can share their responses. There are no correct or incorrect responses to these questions. However, if you would prefer not to answer a question, simply say "SKIP." Thank you for sharing your experiences with us today."

Interview Questions:

1. Describe your experience with the Teacher Performance Assessment.

2. What is the purpose of the TPA?

3. If you could have selected only one of the four tasks of the TPA to complete, which would you have chosen? Why?

4. Was a particular task of the TPA more challenging than the others?... Please explain.

5. Will your practice as a teacher change because of your TPA experience? Please explain. *[If candidates ask for clarification on "practice" it is meant to include both teacher thinking about and actions taken in preparing, instructing, assessing and reflecting on teaching.]*

6. Is there anything else you would like to share about your experience with the TPA?

Appendix F6

TPA Group Interview – SOE Faculty

Dear Investigator: This is a structured interview that should take approximately 30 minutes. Participants include faculty who agreed to participate in the interview and for their responses to be a part of our research. Please videotape or audiotape your conversation and submit it Suzie Henning. A flip cam has been reserved and can be collected from [department secretary].

Prior to the interview, determine whether you want to call on faculty or have faculty volunteer and arrange the seats in a circle or semi-circle. During the interview, ask each of the questions in the order in which they appear. Do not move to the next question until each participant has responded or asked to be skipped. The number of questions in this interview is based on an average of a one-minute response per faculty, per question. You will want to monitor time to make sure that each respondent has time to share. Please advise a note-taking device (number the students or use their initials) and take notes during the interview on the order of the respondents in answering the questions for transcription purposes. (For instance, Q1: RB, JD, SH, JK, NB; Q2: NB, SH, JK, RB, JD, etc.)

Introductions:

- Please start promptly at 1:00.
- Introduce yourself and your role in the project.
- Explain the purpose of the interview:

> "This group interview is part of a larger project to study teacher candidate experiences with the TPA. This interview will take approximately 45 minutes. Your name will not be attributed to your comments in our published research. I will start by asking a question and then we will progress around the circle/volunteers can share their responses. There are no correct or incorrect responses to these questions. However, if you would prefer not to answer a question, simply say "SKIP." Thank you for sharing your experiences with us today."

Interview Questions

1. Describe any positive impact of the TPA this year.
2. Describe any challenges or difficulties of the TPA.
   a. What was most challenging for you?
   b. What did you think our students would struggle with the most?
3. What are student perceptions of the TPA?
4. We have only recently received d scored feedback on candidate TPA work from this term and are in the process of analyzing what it means for our programs. Given what you know about our programs, what do you anticipate the data will reveal about our strengths and weaknesses in preparing candidates to complete the TPA?
5. In what ways, if any, has the TPA changed the role of the methods instructor? Any good ideas or strategies you have included in your courses to address the TPA?
6. *At what point does the TPA belong in our programs? Where should TC feel most prepared to address the TPA?
7. What isn't addressed in the TPA that still matters?
8. Is there anything else you would like to share about your experience with the TPA?

Appendix F7

TPA Group Interview – University Supervisors

Dear Investigator: This is a semi-structured interview that should take approximately 50 minutes. Participants include faculty who agreed to participate in the interview and for their responses to be a part of our research. Please videotape or audiotape your conversation and submit it Suzie Henning. A flip cam has been reserved and can be collected from [department secretary].

Prior to the interview, determine whether you want to call on participants or have them volunteer. Arrange the seats in a circle or semi-circle. During the interview, ask each of the questions in the order in which they appear. Do not move to the next question until each participant has responded or asked to be skipped. You will want to monitor time to make sure that each respondent has time to share. Please advise a note-taking device (number the students or use their initials) and take notes during the interview on the order of the respondents in answering the questions for transcription purposes. (For instance, Q1: RB, JD, SH, JK, NB; Q2: NB, SH, JK, RB, JD, etc.)

Introductions

- Introduce yourself and your role in the project.
- Explain the purpose of the interview:

> "This group interview is part of a larger project to study teacher candidate experiences with the TPA. This interview will take approximately 60 minutes. Your name will not be attributed to your comments in our published research. I will start by asking a question and then we will progress around the circle/volunteers can share their responses. There are no correct or incorrect responses to these questions. However, if you would prefer not to answer a question, simply say "SKIP." Thank you for sharing your experiences with us today."

Interview Questions

1. Describe your experience supervising candidates this term (TED) / year (MIT).
2. Has your role as the supervisor changed this year?
    a. How does it compare to your experience supervising / scoring with the Pedagogy Assessment (PPA)?
    b. What was most challenging for you? What was most beneficial?
    c. What did our students struggle with the most?
3. What are student perceptions of the TPA?
4. What did students report that they learned from their work on the TPA?
5. What did mentor teachers report about their experiences working with candidates during their TPA?
6. Describe any effect of the TPA on student learning in the classrooms you visited.
7. To date, we have not received scored feedback on candidate TPA work from this term. Given what you know about your candidates, what do you anticipate the data will reveal about our strengths and weaknesses in preparing candidates to complete the TPA?
8. *At what point does the TPA belong in the student teaching term? Where should TC feel most prepared to address the TPA at the beginning, middle, end?
9. What isn't addressed in the TPA that still matters?
10. What can we do to support your work with candidates next year?
11. Is there anything else you would like to share about your experience with the TPA?

*Skip if running out of time.

Thank you for sharing your thoughts and your time with us today.

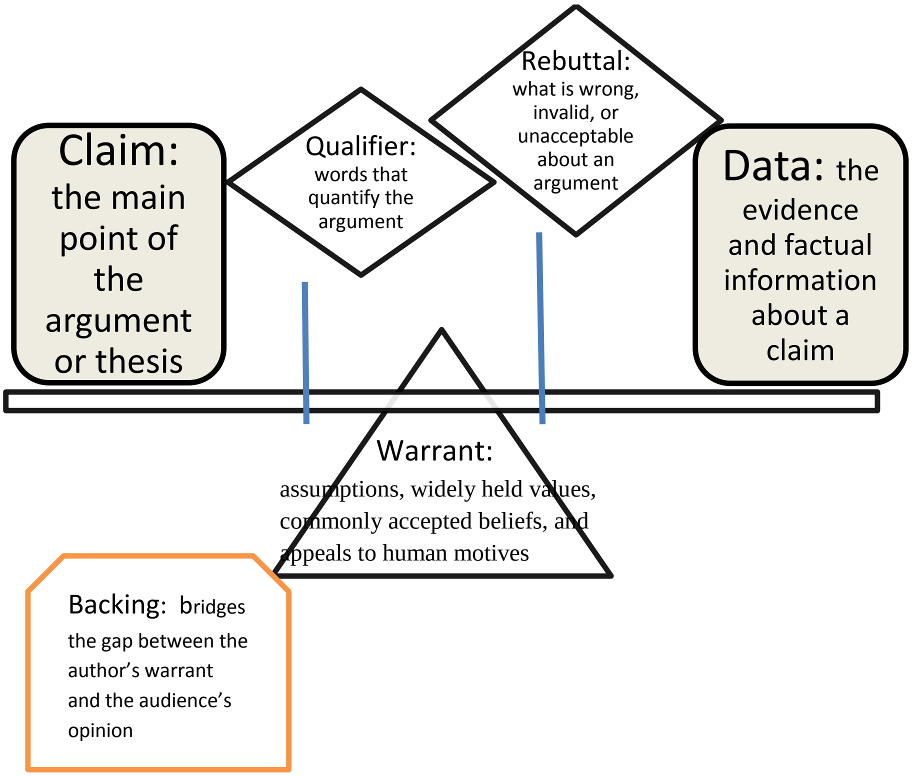
Appendix G

Codes and Descriptors Used for Qualitative Analysis

1. **Teaching Readiness**
   - A. Participant experiences
   - B. Preparing for TPA
   - C. Reflecting
   - D. Perceptions around Success/Failure
   - E. Motivation
   - F. Learning Value (formative)
   - G. Challenges
   - H. Purpose
2. **Construct Representation**
   - A. Content Validity
     - Missing curriculum
   - B. Construct-Irrelevancy
3. **Scoring/ Evaluation**
   - A. Rubric Expectations
   - B. Meaning and use of scores
4. **Generalizability/ Fairness**
   - A. Placement
   - B. Mentor
   - C. Video
   - D. Life-Impact
   - E. Resources
   - F. Writing
5. **Decision-Making**
   - A. Purpose for TC
   - B. Purpose for program
   - C. Funding

Appendix H

Toulmin's Argumentative Model (from LeTourneau University, 2002)


Claim: the main point of the argument or thesis
Qualifier: words that quantify the argument
Rebuttal: what is wrong, invalid, or unacceptable about an argument
Data: the evidence and factual information about a claim
Warrant: assumptions, widely held values, commonly accepted beliefs, and appeals to human motives
Backing: bridges the gap between the author's warrant and the audience's opinion

Appendix I

Interpretive Argument and Data Sources

The following table identifies the original interpretive argument and connects that argument to the research questions, number of instruments that provide evidence to address that research question and the type of data that instrument provides. Note that a later iteration of the IA merged the Generalization and Fairness inferences.

Interpretive Argument		
Inference/RQ	**Type of Data Sources**	
	Qualitative Data	*Quantitative Data*
Construct Representation: Teaching Effectiveness / Teacher Readiness 1. Do the four TPA tasks represent the six categories relevant to the intended construct (teacher readiness)?	*Expert Consensus* *TPA Handbook*	
Evaluation / Scoring 2. Are the scoring procedures sound and reliable? 3. Are the rubric score levels achieved by the candidate actually representative of what that candidate performed on the TPA? (Does candidate performance correlate to the assigned test score)?	*Case Study* *Interviews* *Surveys*	*TPA scores* *West E scores* *Surveys*
Generalization 4. Are the score levels achieved on the rubrics a true representation of a candidate's performance? In other words, are the scores a candidate earned consistent and generalizable with other samples of that candidate's teaching performance? 5. Does poor performance on the TPA imply a lack of adequate mastery of the construct?	*Case Study* *Interviews* *Surveys*	*Candidate, Mentor, Supervisor Evaluations* *Test Scores* *Surveys*
Fairness 6. How generalizable are the criteria, rubrics, procedures, and scores derived from the TPA across different candidates and handbooks? *7.* Does TPA proficiency depend upon factors beyond the candidate's control? How generalizable are the criteria, rubrics, procedures, and scores derived from the TPA across testing sites, placements and placement length and programs?	*Case Study* *Surveys*	*Candidate, Mentor, Supervisor Evaluations* *Surveys* *TPA Scores*
Extrapolation 8. Do TPA test scores provide reliable indicators of a readiness to teach? Are the TPA scores, as a whole, a true measurement of teaching ability?	*Surveys* *Interviews*	*Mentor and Supervisor Evaluations* *Surveys*

Decision-Making 9. Guidance is in place so that all stakeholders know what scores mean and how the outcomes will be used?	*Document Analysis* *Interviews*	

Appendix J

Expert responses to best evidence of candidate readiness to teach

Appendix J1

Supervisor Responses to Traits of Teaching Readiness

Best Evidence of Candidate Readiness, as listed by Supervisor, Phase 4 n=14	Number of Supervisors	%	# of TPA Rubrics
Performance evaluation or direction observation of teaching	6/14	42%	2
Evidence of effective instruction (assessment)	5/14	35%	3
Collaboration	4/14	28%	0
Culturally Responsive	4/14	28%	6
Effective Planning	4/14	28%	3
Rapport	4/14	28%	0
Content knowledge	3/14	21%	5

Appendix J2

Mentor teacher ranking of characteristics that exhibit candidate's readiness to teach

(Pre-TPA Experience)

The table below and an asterisk (*) indicates the inclusion of a construct that is not directly measured by the TPA.

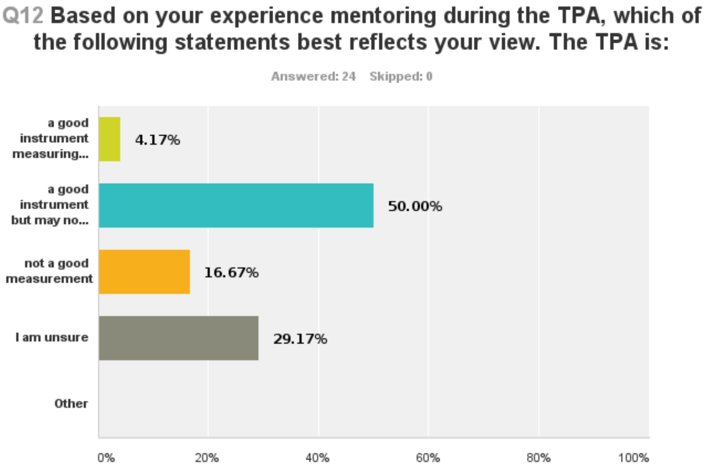
Rank the following responses. A candidate is ready to teach when they: **n=55**	**Essential**	**Important**	**Somewhat important**	**Not important**	**Unsure**
Understand the material presented	**58.18%** 32	**38.18%** 21	**3.64%** 2	**0%** 0	**0%** 0
***Practice good classroom management**	**76.36%** 42	**21.82%** 12	**1.82%** 1	**0%** 0	**0%** 0
Develop instruction aligned to standards	**38.18%** 21	**54.55%** 30	**7.27%** 4	**0%** 0	**0%** 0
Analyze assessment data to determine next steps with students	**43.64%** 24	**41.82%** 23	**14.55%** 8	**0%** 0	**0%** 0
Engage learners in critical thinking and meaningful dialog	**54.55%** 30	**40%** 22	**5.45%** 3	**0%** 0	**0%** 0
Reflect on their performance with an eye to determining strengths and weaknesses	**69.09%** 38	**29.09%** 16	**1.82%** 1	**0%** 0	**0%** 0
Encourage and utilizes student voice to make decisions about instruction, assessment, and student needs	**29.09%** 16	**45.45%** 25	**23.64%** 13	**1.82%** 1	**0%** 0
Practice differentiation	**32.73%** 18	**41.82%** 23	**23.64%** 13	**1.82%** 1	**0%** 0
Develop and implements instruction that focuses on the content language needs of their students	**27.27%** 15	**52.73%** 29	**18.18%** 10	**1.82%** 1	**0%** 0
***Communicate and collaborates with parents and colleagues**	**38.18%** 21	**41.82%** 23	**18.18%** 10	**1.82%** 1	**0%** 0
Is culturally proficient and responsive to the cultural needs of the students in their classroom	**32.73%** 18	**47.27%** 26	**18.18%** 10	**1.82%** 1	**0%** 0

Appendix J3

Mentor teacher views of TPA

(Post-TPA Experience)

n=24 Answer Choices	Responses
a good instrument measuring candidate readiness to teach	4.17% 1
a good instrument but may not indicate candidate readiness to teach	50% 12
not a good measurement	16.67% 4
I am unsure	29.17% 7


Q12 Based on your experience mentoring during the TPA, which of the following statements best reflects your view. The TPA is:

Answered: 24 Skipped: 0

Appendix J4

University supervisor ranking of characteristics that exhibit candidate's readiness to teach (Pre-TPA Experience)

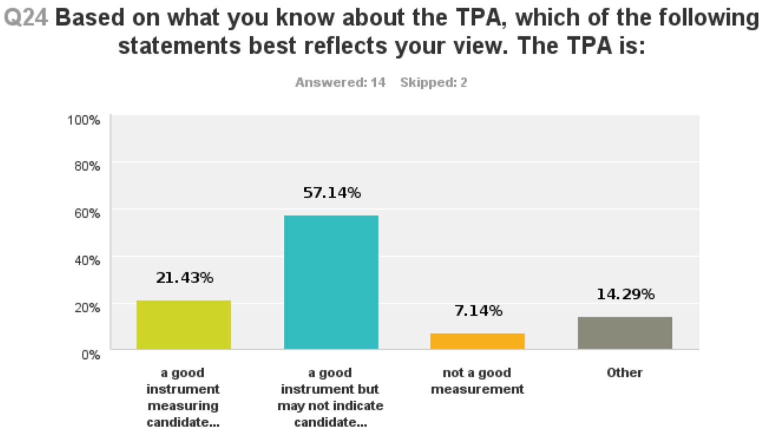
The table below and an asterisk (*) indicates the inclusion of a construct that is not directly measured by the TPA.

Rank the following responses. A candidate is ready to teach when they: **n=14**	**Essential**	**Important**	**Somewhat important**	**Not important**	**Unsure**
Understand the material presented	**78.57%** 11	**21.43%** 3	**0%** 0	**0%** 0	**0%** 0
***Practice good classroom management**	**57.14%** 8	**42.86%** 6	**0%** 0	**0%** 0	**0%** 0
Develop instruction aligned to standards	**57.14%** 8	**42.86%** 6	**0%** 0	**0%** 0	**0%** 0
Analyze assessment data to determine next steps with students	**78.57%** 11	**21.43%** 3	**0%** 0	**0%** 0	**0%** 0
Engage learners in critical thinking and meaningful dialog	**64.29%** 9	**35.71%** 5	**0%** 0	**0%** 0	**0%** 0
Reflect on their performance with an eye to determining strengths and weaknesses	**78.57%** 11	**21.43%** 3	**0%** 0	**0%** 0	**0%** 0
Encourage and utilize student voice to make decisions about instruction, assessment, and student needs	**42.86%** 6	**57.14%** 8	**0%** 0	**0%** 0	**0%** 0
Practice differentiation	**42.86%** 6	**57.14%** 8	**0%** 0	**0%** 0	**0%** 0
Develop and implement instruction that focuses on the content language needs of their students	**35.71%** 5	**64.29%** 9	**0%** 0	**0%** 0	**0%** 0
***Communicate and collaborate with parents and colleagues**	**57.14%** 8	**35.71%** 5	**7.14%** 1	**0%** 0	**0%** 0
Is culturally proficient and responsive to the cultural needs of the students in their classroom	**50%** 7	**42.86%** 6	**7.14%** 1	**0%** 0	**0%** 0

Appendix J5

University Supervisor views of TPA Pre-TPA Experience

n=14 Answer Choices–	Responses–
a good instrument measuring candidate readiness to teach	21.43% 3
a good instrument but may not indicate candidate readiness to teach	57.14% 8
not a good measurement	7.14% 1
Other	14.29% 2



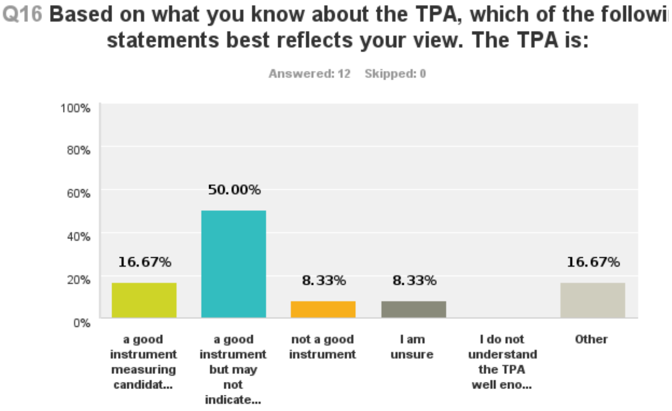
Other:

1. "At first review the TPA looks to be a good instrument. However, it cannot stand alone. Evaluation and support of the mentor teacher and university supervisor are essential for a more complete profile of the candidate. I've yet to fully experience the full process."
2. "I think the TPA is an instrument through which teacher candidates can look at their teaching practices and reflect deeply about them. However, the timing of it is questionable since the candidates need to spend so much time on student teaching requirements and classroom activities. It places a huge stress on them and takes time away from daily activities in the classroom. I feel it is too intense of an assessment during the student teaching experience. The students are learning [sic] about so many aspects of being a good teacher, and I have noticed this has been a distraction to their teaching experience. The quality of their teaching has actually dropped when this assessment was added to their workload."

Appendix J6

Faculty views of TPA Post-TPA Experience

n=12 Answer Choices	Responses
a good instrument measuring candidate readiness to teach	16.67% 2
a good instrument but may not indicate candidate readiness to teach	50% 6
not a good instrument	8.33% 1
I am unsure	8.33% 1
I do not understand the TPA well enough to have a view	0% 0
Other	16.67% 2


Q16 Based on what you know about the TPA, which of the followi statements best reflects your view. The TPA is:

Other:

1. "Since the instrument has scorer subjectivity, it is an instrument that concerns me if used as the only or the definitive determination for certification."
2. "It has good points, one them being a familiarity with current trends and some current language. It is so high-stakes that it exlipses [sic] everything else about student teaching. I could be wrong, but it appears that Pearson and Stanford are butting heads about their roles, and I believe that word is out that this instrument is talking all the enjoymnent [sic] out of teaching and is so obtuse and difficult that it is not worth the hassle to go through, particularly in today's job markets. The extra projected costs are going to be hard on students and the perception is that Pearson is making big-bucks. (The fact that they are now in the GED market is disconcerting, as well. I think the TPA concerpt [sic] has merit, when it is administered by a for-profit company that has been engaged in some rather unethical behavior makes me very uncomfortable for our students and faculty."

Appendix K

Supervisor Assessment of TPA Missing Traits from Phase 4 Interview

Missing from TPA	#	%
Management	3/14	21%
Dispositions	8/14	57%
Collaboration	8/14	57%

Appendix L

Supervisor Assessment of Construct Irrelevant Traits Measured by TPA from Phase 4 Interview

Not relevant, but measured by TPA	#	%
Writing/Reading ability	8/14	57%
Ability to use technology, especially videography and digital editing	9/14	64%
Ability to use scripted plans	10/14	71%
A mentor teacher(s) understanding of and willingness to assist with TPA	12/14	86%

Appendix M

Alignment of the Teacher Performance Assessment (edTPA) with Washington Standard V for Teachers

Reproduced from **http://assessment.pesb.wa.gov/assessments/edtpa/standard-5-tpa**).

Standard V (WAC 181-78A-270(1))	**edTPA Washington rubric**
a. Effective teaching	
(i)Using multiple instructional strategies, including the principles of second language acquisition, to address student academic language ability levels and cultural and linguistic backgrounds	EM 4: How does the candidate identify and support language demands associated with a key mathematics learning task? EM14: How does the candidate analyze students' use of language to develop content understanding?
(ii) Applying principles of differentiated instruction, including theories of language acquisition, stages of language, and academic language development, in the integration of subject matter across the content areas of reading, mathematical, scientific, and aesthetic reasoning	EM 4: How does the candidate identify and support language demands associated with a key mathematics learning task? EM 14: How does the candidate analyze students' use of language to develop content understanding?
(iii) Using standards-based assessment that is systematically analyzed using multiple formative, summative, and self-assessment strategies to monitor and improve instruction	EM 5: How are the informal and formal assessments selected or designed to monitor students' conceptual understanding, procedural fluency, and reasoning/problem solving skills? EM 10: How does the candidate use evidence to evaluate and change teaching practice to meet students' varied learning needs? EM 11: How does the candidate analyze evidence of student learning of conceptual understanding, procedural fluency, and reasoning/problem solving skills? EM 12: What type of feedback does the candidate provide to focus students? EM 13: How does the candidate provide opportunities for focus students to use the feedback to guide their further learning? EM 15: How does the candidate use the analysis of what students know and are able to do to plan next steps in instruction?

(iv) Implementing classroom/school centered instruction, including sheltered instruction that is connected to communities within the classroom and the school, and includes knowledge and skills for working with others	EM 2: How does the candidate use knowledge of his/her students to target support for students to develop conceptual understanding, procedural fluency, and mathematical reasoning/problem solving skills? EM 3: How does the candidate use knowledge of his/her students to justify instructional plans? EM 6: How does the candidate demonstrate a respectful learning environment that supports students' engagement in learning? EM 7: How does the candidate actively engage students in developing understanding of mathematical concepts?
(v) Planning and/or adapting standards-based curricula that are personalized to the diverse needs of each student	EM 2: How does the candidate use knowledge of his/her students to target support for students to develop conceptual understanding, procedural fluency, and mathematical reasoning/problem solving skills? EM 3: How does the candidate use knowledge of his/her students to justify instructional plans?
(vi) Aligning instruction to the learning standards and outcomes so all students know the learning targets and their progress toward meeting them	EM16: How does the candidate focus student attention on the learning targets? EM17: How does the candidate support students to access resources for learning and to monitor their own learning progress? EM18: How does the candidate use student-voice evidence to identify instructional improvements?
(vii) Planning and/or adapting curricula that are standards driven so students develop understanding and problem-solving expertise in the content area(s) using reading, written and oral communication, and technology	EM1: How do the candidate's plans build students' conceptual understanding, procedural fluency, and mathematical reasoning/problem solving skills? EM7: How does the candidate actively engage students in developing understandings of mathematical concepts? EM8: How does the candidate elicit responses to promote thinking and develop understanding of mathematical concepts? EM 9: How does the candidate use representations to develop students' mathematical concepts?
(viii) Preparing students to be responsible citizens for an environmentally sustainable, globally interconnected, and diverse society	NA
(ix) Using technology that is effectively integrated to create technologically proficient learners	NA
(x) Informing, involving, and collaborating with families/neighborhoods, and communities in each student's educational process, including using information about student cultural identity, achievement and performance	EM 2: How does the candidate use knowledge of his/her students to target support for students to develop conceptual understanding, procedural fluency, and mathematical reasoning/problem solving skills? EM 3: How does the candidate use knowledge of his/her students to justify instructional plans? EM 7: How does the candidate actively engage students in developing understanding of mathematical concepts?

b. Professional development	
Developing reflective, collaborative, professional growth-centered practices through regularly evaluating the effects of his/her teaching through feedback and reflection	EM 10: How does the candidate use evidence to evaluate and change teaching practice to meet students' varied learning needs? EM 15: How does the candidate use the analysis of what students know and are able to do to plan next steps in instruction?
c. Teaching as a profession	
(i)Participating collaboratively and professionally in school activities and using appropriate and respectful verbal and written communication	NA
(ii)Demonstrating knowledge of professional, legal, and ethical responsibilities and policies	NA

Appendix N

Rubric Score Statistics for all TPA Participants

Rubric Scores By Candidate

	1	2	16	3	4	5	6	7	8	9	10	11	12	13	14	15	Sum	Mean	Pass/Fail?
1	2	3	2	2	2	2	3	2	3	2	3	3	2	2	3	2	38	2.38	P
2	3	3	3	3	3	3	2	3	2	3	2	3	3	2	3	3	44	2.75	P
3	3	3	4	2	3	3	3	3	3	2				5	4	3	41	3.15	P
4	4	4	4	1	3	3	3	4	3	3	4	3	3	2	1	2	47	2.94	P
5	3	3	2	3	3	3	2	2	3	3	2	3	3	3	2	3	43	2.69	P
6	4	4	3	3	3	3	3	2	3	3	3	3	3	2	2	3	47	2.94	P
7	4	4	4	3	3	3	3	2	3	3	3	3	3	1	1	1	44	2.75	P
8	4	3	4	4	3	3	3	3	4	4	4	4	3	4	3	3	56	3.50	P
9	3	3	4	4	4	3	3	3	2	2	3	3	3	2	3	3	48	3.00	P
10	3	2	2	2	2	2	3	2	3	2	3	3	3	2	3	3	40	2.50	p
11	2	1	2	2	2	2	3	3	3	2	2	1	2	2	2	3	34	2.13	F
12	4	3	4	4	4	4	4	4	3	4	4	4	3	5	2	3	59	3.69	P
13	4	4	3	4	1	2	3	3	3	3	3	2	4	3	3	3	48	3.00	P
14	3	4	3	3	3	3	3	3	3	3	4	3	3	3	2	3	49	3.06	P
15	3	4	3	4	4	4	4	4	4	4	4	4	4	4	4	4	62	3.88	p
16	4	4	5	4	2	2	3	4	3	3	3	3	3	3	3	3	52	3.25	p
17	4	4	4	3	3	3	3	2	3	3	2	3	3	2	2	2	46	2.88	p
18	4	4	4	4	3	3	3	2	4	3	4	4	4	4	3	3	56	3.50	p
19	4	4	3	4	4	3	3	3	4	4	4	5	3	4	4	5	61	3.81	p
20	3	3	3	3	2	3	2	2	2	3	2	3	3	2	2	3	41	2.56	p
21	2	2	2	2	3	3	3	2	4	3	2	3	2	2	2	4	41	2.56	P
22	2	3	3	3	3	2	2	3	3	3	2	2	3	3	2	3	42	2.63	P
23	3	3	3	3	3	2	3	4	3	3	2	3	3	3	3	3	47	2.94	P
24	3	3	3	4	3	3	3	3	3	3	2	3	3	4	4	3	50	3.13	p
25	3	3	4	3	3	4	4	3	3	2	3	3	3	3	3	3	50	3.13	p
26	3	2	3	1	3	2	3	2	2	2	2	2	2	2	2	2	35	2.19	P
27	4	4	4	3	3	3	3	3	3	3	3	2	2	4	2	3	49	3.06	p
28	4	4	4	3	3	3	3	2	3	3	3	3	3	3	3	1	48	3.00	p
29	3	2	2	3	2	1	2	3	2	2	2	2	1	1	2	1	31	1.94	F
30	3	3	4	3	4	3	3	3	4	3	2	3	3	3	4	3	51	3.19	P
31	4	3	4	4	4	3	4	3	3	4	4	3	4	3	2	3	55	3.44	P
32	3	3	3	3	2	1	3	3	3	3	3	4	3	2	2	2	43	2.69	P
33	2	2	1	1	2	2	2	1	2	1	E	2	2	2	2	2	26	1.73	F
34	4	4	4	3	3	3	3	2	3	2	3	3	3	3	2	2	47	2.94	P
35	4	4	4	4	3	3	4	3	4	3	3	3	3	3	3	2	53	3.31	P
36	5	4	5	4	4	4	5	4	4	4	5	5	5	5	4	5	72	4.50	P
37	3	3	3	3	4	4	4	4	4	3	3	4	4	3	4	4	57	3.56	P
38	3	2	3	3	2	2	3	2	2	3	2	3	2	2	3	3	40	2.50	P

39	3	2	2	3	3	2	2	2	3	3	2	3	2	3	2	3	40	2.50	P
40	4	3	4	4	3	3	3	3	3	3	4	3	4	3	E	4	51	3.40	P
41	4	3	4	3	4	3	4	3	4	3	4	4	3	3	3	4	56	3.50	P
42	4	3	4	3	2	3	3	3	3	3	3	3	3	2	2	3	47	2.94	P
43	2	1	2	2	3	2	2	2	1	1	2	2	2	2	2	1	29	1.81	F
44	4	4	5	4	4	4	3	3	3	2	3	3	4	4	2	3	55	3.44	P
45	3	3	4	3	3	3	5	4	4	3	2	3	3	2	2	3	50	3.13	P
46	4	2	2	4	3	2	3	3	2	3	2	3	2	2	3	1	41	2.56	P
47	4	3	3	4	4	3	4	3	4	3	3	4	3	4	3	5	57	3.56	P
48	4	4	4	4	4	4	3	4	4	4	4	3	3	4	3	3	59	3.69	P
49	4	3	4	3	2	3	3	4	4	2	3	2	3	3	4	4	51	3.19	P
50	4	4	4	3	2	2	3	3	2	3	3	4	3	2	3	3	48	3.00	P
51	4	4	3	4	3	3	4	4	4	3	4	4	4	2	4	3	57	3.56	P
52	3	3	3	3	3	2	2	3	2	3	2	2	2	2	2	3	40	2.50	P
53	4	4	3	4	3	3	4	4	4	4	4	3	4	2	4	5	59	3.69	P
54	3	3	2	3	3	4	3	3	4	4	3	3	3	5	3	5	54	3.38	P
55	3	4	4	4	3	3	4	4	4	3	4	3	4	3	3	3	56	3.50	P
56	3	2	2	3	3	2	3	3	3	3	2	3	3	1	2	1	39	2.44	P
57	2	2	2	3	3	3	3	3	3	2	2	1	2	1	3	2	37	2.31	P
58	4	4	4	3	4	3	2	2	2	3	3	4	3	3	2	2	48	3.00	P

Appendix O

Inter-rater Agreement Indexes for TPA scores **(Graham, Milanowski, & Miller, 2012)**

Index	**High Agreement Statistic for Consequential Use**	**Minimum Agreement Statistic for Consequential Use**	**Average Agreement Reported in the Literature**	**Average agreement for TPA**
% **Absolute** Agreement	.90	.75	.70	.71
% +1/-1 Adjacent **Variance**	-	-	-	.79 (23% of total difference in agreement)
% +2/-2 Adjacent **Variance**	-	-	-	.18 (5% of the total difference in agreement)
% +3/-3 Adjacent **Variance**	-	-	-	.03 (1% of the total difference in agreement)
Cohen's Kappa	.81	.61	.54	.65
Intra-class correlation	.90	.80	.81	.14

Appendix P

Inter-rater agreement for TPA Scores

Appendix P1

Inter-rater agreement by Task/Category and Rubric

Task/Category	**% Agreement**
R1	67
R2	71
R2b (16)	69
R3	64
Task 1 Average	**68**
R4	74
R5	71
Task 2 Average	**73**
R6	74
R7	71
R8	69
Task 3 Average	**71**
R9	74
Task 4 Average	**74**
R10	66
R11	64
R12	80
Academic Language Average	**70**
R13	68
R14	75
R15	74
Student Voice Average	**72**
Average across Tasks/Categories	**71**

Appendix P2

Inter-rater scoring differences, by Rubric

	R1	R2	R16	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	
Total Scored	58	58	58	58	58	58	58	58	58	58	56	56	56	58	56	58	920
# Difference	19	17	18	21	15	17	15	17	18	15	18	20	11	20	15	15	271
%	**33**	29	**31**	**36**	26	29	26	29	**31**	26	**32**	**36**	20	**34**	27	26	**0.29**
Total-1	1	4	2	3	7	8	2	6	5	2	5	4	3	12	6	3	0.27
Total +1	17	10	7	13	6	9	10	10	10	9	7	14	5	2	5	6	0.52
																	.79
Total -2				1	1					1				1	1	3	0.03
Total +2	1	3	8	4	1		2	1	2	3	4	1	2	5	2	3	0.15
																	.18
Total -3																	0.00
Total +3			1				1		1		2	1	1		1		0.03
																	.3

Appendix Q

Rubrics with Potential Language or Judgment Issues that Could Lead to Scorer Error

(Secondary History/Social Studies)

Rubric	Rubric Language	Potential Judgment or Scoring Issue	In all Rubrics?
1: Planning	Level 2: "Standards, objectives, learning tasks and materials are loosely or inconsistently aligned with each other" Level 3: ""Standards, objectives, learning tasks and materials are consistently aligned with each other and with the central focus for the learning segment"	The terms loosely and consistently are difficult to quantify. Consistently does not indicate always and the extent to which standards must be aligned is not clear. It is possible, then, that one rater may determine that a candidate's submission met standard while another may say it did not based on different notions of "consistently." Also, the term "loosely or inconsistently aligned" requires professional judgment. Raters are told to look at the "preponderance of evidence" to make a judgment. However, this too requires professional decisions that may differ for the same candidate's work. The differences between levels require candidates not just to align their lessons with standards but with an additional central focus that addresses the assigned performance focus, based on the subject area. What if the candidate aligns the standards, objectives and tasks well and consistently throughout the learning segment but the learning segment does not address the performance focus required by the subject endorsement? Is the skill assessed alignment or content pedagogy?	yes
2: Planning	Level 3: "Learning tasks draw on students' prior learning **and** experience **and** social/emotional development **or** interests."	Meeting standard requires that candidates complete multiple skills and it is not clear, in the case that both tasks are not completed, which would give preference to a passing score. The use of "AND" to link two behaviors in this rubric requires professional judgment.	yes
3: Planning	When identifying candidate abilities to plan assessment: Level 2: uses the term "limited evidence" Level 3: uses the term "evidence" Level 4: uses the term "multiple forms of evidence"	Determining what qualifies as limited evidence, evidence, and multiple forms of evidence requires professional judgment that could differ across raters.	yes

4: Instruction	Level 2: “attempts to link… but the links are unrelated to understandings…” Level 3: “links new content to student’s prior learning” Level 2: “students are participating…focusing solely on facts” Level 3: “students are intellectually engaged”	When viewing the video, determining if a candidate attempts to link new content to prior learning or actually links to new content learning requires professional judgment. In addition, determining whether students are participating or intellectually engaged is a difficult professional determination, especially in a short video clip.	yes
6: Assessment	Level 2: “criteria are generally aligned”; “analysis is supported …in a general way” Level 3: “Criteria are clearly aligned”;	It is unclear what constitutes “generally” or “general way.”	yes
7: Assessment	Level 3: “Candidate describes how students will use feedback to improve their performance” Level 4: “...students will use feedback to deepen their [performance objective]”	It is unclear if this is a planning expectation. Is it the case that the candidate needs to provide an opportunity for the student to improve based on feedback or that the candidate should create an environment in which the student will want/value the feedback for their own improvement. How and whether this criterion is met will differ by rater. To exceed standard, the candidate must provide feedback that will “deepen” the student’s understanding of the performance objective. Measuring the extent to which a “deepening understanding” can occur from feedback provided is unclear.	yes
8: Assessment	Level 2: “Next steps propose general support” Level 3: “targeted support”	Whether a candidate has provided general or targeted support is a professional judgment. In fact, it seems likely that the type of support needed by a student is highly contextual, situational, and individualized.	yes
11: Academic Language	Level 2: “limited support for students” Level 3: “provides support”	The difference between providing support and providing limited support requires professional judgment.	yes
14: Student Voice	Level 2: “creates or adopts a tool” Level 3: “creates one or more tools”	This language penalizes students for adopting a tool or strategy when it may, in fact, be a stronger instrument, than one they might create. Similarly, some candidates do not have the freedom to use tools of their own creation.	yes

Appendix R

Rubrics with Potential Development Issues that Could Lead to Scorer Error (Secondary History/Social Studies)

- Rubric 5: Instruction: A candidate will not meet standard if they ask “surface-level” or “correct or incorrect” questions. Candidates meet standard (Level 3) if they “elicit student responses that require [performance expectation]”. One case study candidate stated it best when she said, “It is completely realistic and they need to know [that candidates can meet the performance expectation]. I understand why they picked primary documents. It just seems …that they picked the skill that was probably the most challenging because orchestrating a lab and doing data analysis for science is probably one of the most challenging pieces and definitely primary sources was probably the most challenging for history” [Jennifer, Phase 2, 3/8/12].
- Rubric 7: Assessment: A candidate whose feedback to students focuses on identifying and correcting errors will not meet standard. To meet standard, a candidate must demonstrate that feedback identifies both successes and errors and describe how students will use feedback to improve. However, candidates who are first learning how to apply assessment will tend to focus and reflect on improvements needed.
- Rubric 9: Reflection. Initial surveys of candidates entering student teaching revealed that their primary concern about teaching was classroom management [TPA Pre-Experience Survey, 1/17/12]. It is likely, when reflecting on their performance, they would think about those issues that they had already determined to be weaknesses, or previously established goals for their teaching practice. However, in this rubric (the only reflection rubric) candidates that focus on classroom management will not meet standard. Only those reflections that address learning needs, evidence of student learning, or students’ prior learning will address the rubric criteria. In addition, candidates are asked not to reflect on

their performance instead focusing on what they would change. In fact, if a candidate focused on their own teaching practice, they would earn a Level 1: "Candidate proposes changes unrelated to knowledge of students and their learning." This is a classic sample of how the TPA looks at the participant not as a candidate who is learning about teaching by teaching but as a practicing teacher with some level of prior experience. The lack of measurement of classroom management skills during student teaching was considered a missing trait by experts (see VE 1).

- Rubric 10: Academic Language: candidates who primarily focus on their students language development needs will not meet standard. Similarly, candidates who identify unfamiliar vocabulary without considering language demand will not meet standard. To meet standard, this is out of sync with the development of a novice teacher who, before they can understand any patterns or exceptions, must first understand the whole class and what is "typical." Until that happens, it is a natural focus to determine and focus on student needs. While an ability to also focus on student strengths is the mark of a stronger teacher, it is also the mark of a more experienced teacher. This rubric holds candidates to an unrealistic expectation of prior experience with students and with language needs.
- Rubric 13: Student Voice: this rubric evaluates the candidates' ability to communicate learning targets in ways that students can understand. It requires a level of understanding of class response and behavior that is very difficult without prior experience with that grade level and experience with those students. A reminder that candidates were asked to complete this assessment in the first month of student teaching is a reminder that such understanding of the students and the context is unlikely. Similarly, SV is not a practice of many classroom teachers in WA. Having been adopted as a part of Standard V, few teachers with more than three years of teaching experience (required for mentoring candidates) would have experienced a preparation program that incorporated SV. Some may have had in-service training, but not all. It is unlikelihood that SV practices (such as

communication of “targets”) have been integrated for P-12 students prior to the candidate’s arrival.

- Rubric 15: Student Voice: asks candidates to use this student data to reflect on their own practice. However, as several candidates indicated during the case study interviews, this data proved unreliable. Students over or under evaluated their own progress in meeting the targets and, given other assessment data, it was not a good basis with which to identify any instructional implications (other than more SV training might be needed for students and their mentors).

Appendix S

Inter-rater Agreement Indexes for Teacher Candidate Final Progress Report

Index	**High Agreement Statistic for Consequential Use**	**Minimum Agreement Statistic for Consequential Use**	**Average agreement Reported in the Literature**	**Average agreement for Final Progress Reports**
% **Absolute** Agreement	.90	.75	.70	.87
% +1/-1 Adjacent **Variance**	-	-	-	.12 (90% of total difference in agreement)
% +2/-2 Adjacent **Variance**	-	-	-	.01 (9% of the total difference in agreement)
% +3/-3 Adjacent **Variance**	-	-	-	.00 (1% of the total difference in agreement)
Cohen's Kappa	.81	.61	.54	.65

Appendix T

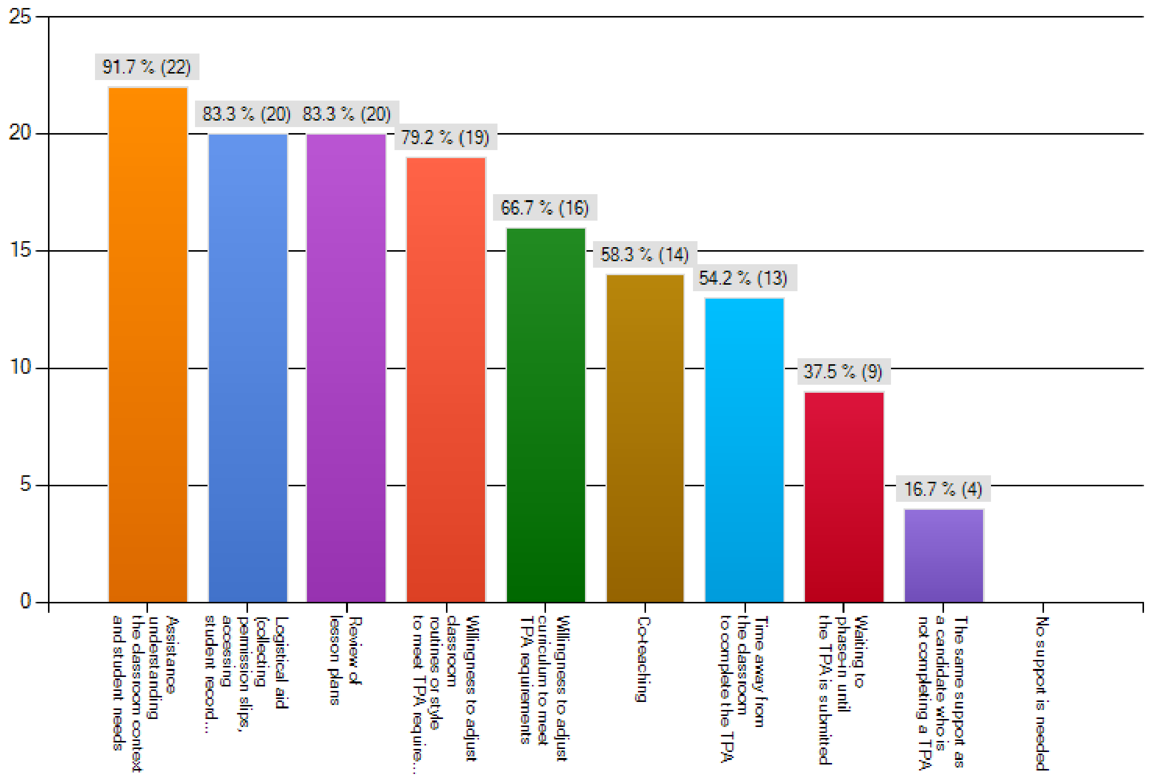
Inter-rater agreement by Standard/Trait for Teacher Candidate Final Progress Report

Standard	% Agreement
1.1	94
1.2	94
1.3	94
1.4	84
1.5	90
1.6	97
Standard 1 Average	**92**
2.1	78
2.2	90
2.3	91
2.4	89
2.5	91
2.6	78
Standard 2 Average	**86**
3.1	85
3.2	85
3.3	84
3.4	85
3.5	79
3.6	87
3.7	77
3.8	79
3.9	78
Standard 3 Average	**82**
4.1	90
4.2	90
4.3	82
4.4	88
4.5	87
4.6	93
Standard 4 Average	**88**
Average Across Standards	**87**

Appendix U

Required Mentor Support for Candidate TPA Completion

Which of the following best captures your view of the type of support a candidate will need from a mentor teacher to complete their TPA (check all that apply):



Appendix V

edTPA Policies **(Professional Educator Standards Board, 2013)**

The list below represents the policy decisions that have been made to date. This document is reproduced from the PESB website (http://www.pesb.wa.gov/)/

Policy	Board meeting
Established Board's plan for use of field trial data related to all or portions of edTPA as initially consequential to candidates	September 2011
Approved award of clock hours for educators scoring the edTPA	January 2012
Determined continued use of the existing performance-based pedagogy assessment for purposes of adding endorsements via Pathway 2 until March 2016. In the meantime, charged staff with searching for other measures of working with Stanford on use of portions of edTPA for this purpose	March 2012
Established an unlimited retake policy	May 2012
Candidates are assigned to either the literacy or mathematics elementary education edTPA. Data of institutions' balance of edTPA literacy and mathematics will be delivered to the Board annually.	updated May 2013
Preservice candidates exiting their teacher prep program with multiple endorsements are required to take one edTPA in one of the endorsements they are seeking. Board approved the recommendations for the edTPA alignment table, with the understanding the tables are recommendations and not requirements. The Board will be updated with stakeholders' feedback in one year. Staff was directed to prepare scenarios (top 5 frequently combined endorsements) which might guide candidates who are contemplating exiting their program with multiple endorsements.	March 2013
Board confirmed edTPA consequential date of January 2014 for all preservice candidates. Some subjects will be consequential beginning January 2014: Elementary Literacy, Elementary Math, Early Childhood Education, Secondary Math, Secondary ELA, Secondary Science, Secondary History/Social Studies, K12 Performing Arts, Special Education, K12 Physical Education, and World Language. Candidates will continue to submit portfolios in all areas so assessments may be scored and institutions will use vendor scores to determine if candidates have satisfactorily met the assessment requirement for program completion.	September 2013
Institutions will provide a scenario document for candidates exiting a program with multiple endorsements as guidance for candidates.	September 2013

During the November meeting the following items were approved by the Board: • The student voice rubrics will not be consequential for candidates but o Candidates will continue to submit portfolios that address the student voice prompts, o Vendor will continue to score student voice rubrics, o Staff will return annually with student voice data, o Members will consider taking action on student voice rubrics in three years. • The cut score for the edTPA is 35 and will be consequential beginning January 2014. • The cut score for the World Language and Classical Language edTPAs will be 30 and will be consequential beginning January 2014.	November 2013
Amendment to WAC 181-78A-264 and WAC 181-78A-270: **WAC 181-78A-264** 1 (f)(i) A: Programs shall administer the ((~~pedagogy~~)) teacher performance assessment adopted by the professional educator standards board to all candidates in a residency certificate program. **WAC 181-78A-270 (1)** (d) Performance Assessment: Beginning January 1,2014, all candidates will complete and pass the teacher performance assessment per WAC 181-78A-264 as authorized by the professional educator standards board: Provided, that candidates who participated in the teacher performance assessment field trials or took the pedagogy assessment prior to January 1, 2014, may be recommended for certification by the preparation program.	November 2013

Appendix W

Validation of Proposed Interpretations

Interpretive Argument	Validity Argument	Proposed Interpretation
Inference	Validation Questions	
Construct Representation: Teaching Effectiveness / Teacher Readiness *<u>Domain of performance:</u> Effective Teaching*	1. Do the four TPA tasks represent the six categories relevant to the intended construct (teacher readiness)?	1: Scores provide a measure of relevant teacher readiness/ effectiveness
Evaluation / Scoring (Target Domain) *<u>Target Domain:</u> Observed Performance is representative sample of performance domain (TPA definition)*	2. Are the scoring procedures sound and reliable? 3. Are the rubric score levels achieved by the candidate actually representative of what that candidate performed on the TPA? (Does candidate performance correlate to the assigned test score)?	1: Scores provide a measure of relevant teacher readiness/ effectiveness
**Generalization* *Universe of Target Domain*	4. Are the score levels achieved on the rubrics a true representation of a candidate's performance? In other words, are the scores a candidate earned consistent and generalizable with other samples of that candidate's teaching performance? 5. Does poor performance on the TPA imply a lack of adequate mastery of the construct? 6. How generalizable are the criteria, rubrics, procedures, and scores derived from the TPA across different candidates and handbooks? 7. Does TPA proficiency depend upon factors beyond the candidate's control? How generalizable are the criteria, rubrics, procedures, and scores derived from the TPA across testing sites, placements and placement length and programs?	1: Scores provide a measure of relevant teacher readiness/ effectiveness
Extrapolation *Target Domain AND Performance Domain*	8. Do TPA test scores provide reliable indicators of a readiness to teach? Are the TPA scores, as a whole, a true measurement of teaching ability?	1: Scores provide a measure of relevant teacher readiness/ effectiveness
Decision Making	9. Guidance is in place so that all stakeholders know what scores mean and how the outcomes will be used?	1: Scores provide a measure of relevant teacher readiness/ effectiveness

Bibliography

AACTE. (2013). *Participation Map.* Retrieved September 25, 2013, from edTPA: http://edtpa.aacte.org/state-policy

AERA, APA, NCME. (1999). *Standards for Educational Psychological Testing.* Washington, DC: AERA.

American Association of Colleges of Teacher Education. (2011). *Transformations in Educator Preparation: Effectiveness and Accountability.* Washington, D.C.: AACTE.

American Association of Colleges of Teacher Education. (2013). *Using edTPA.* Retrieved May 13, 2013, from edtpa.aacte.org

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1999). *Standards for Educational and Pyschological Testing.* American Educational Research Association: Washington, D.C.

Arends, R. (2006). Performance assessment in perspective: History, opportunities, and challenges. In S. Castle, & B. Shaklee (Eds.), *Assessing Teacher Performance: Performance Based Assessment in Teacher Education* (pp. 3-22). Lanham, MD: Roman & Littlefield Education.

Arends, R. (2006). Summative performance assessments. In S. Castle, & B. Shaklee (Eds.), *Assessing Teacher Performance: Performance-Based Assessment in Teacher Education* (pp. 93-123). Lanham, MD: Rowman & Littlefield Education.

Bell, J. (2010). *Doing your Research Project: A Guide for First-Time Researchers in Education, Health and Social Science* (4th ed.). UK: Open University Press.

Bellack, A., & Hersen, M. (1984). *Research Methods in Clinical Psychology.* New York: Pergamon.

Bennett, R. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*, 70-91.

Bennett, R. E., Kane, M., & Bridgeman, B. (2011). Theory of action and validity argument in the context of through-course summative assessment. *Invitational Research Symposium on Through-Course Summative Assessment* (pp. 1-53). Atlanta, GA: Educational Testing Service.

Berlak, A. (2010, Summer). Coming soon to your favorite credential program: National exit exams. *Rethinking Schools, 24*(4).

Billett, S. (2012). Workplace curriculum: Practice and propositions. In F. Dochy, D. Gijbels, M. Segers, & P. Van den Bossche, *Theories of Learning for the Workplace: Building Blocks for Training and Professional Development Programs* (pp. 17-36). New York: Routledge Psychology in Education.

Borsboom, D. (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics.* Cambridge: Cambridge University Press.

Borsboom, D., & Mellenbergh, G. (2007). Test validity in cognitive assessment. In J. Leighton, & M. Gierl, *Cognitive Diagnostic Assessment for Education: Theory and Applications* (pp. 85-115). New York: Cambridge University Press.

Borsboom, D., Cramer, A., Keivit, R., Scholten, A., & Franic, S. (2009). The end ofconstruct validity. In R. Lissetz, *The Concept of Validity: Revisions, New Directions and Applications* (pp. 135-70). Charlotte, NC: Inforamtion Age Publishing.

Borsboom, D., Mellenbergh, G., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-71.

Brennan, R. L. (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice, 20*(4), 6-18.

Briggs, D. C. (2004). Comment: Making an argument for design validity before interpretive validity. *Measurement, 2*(3), 171-191.

Campbell, D., & Fisk, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

CCSSO. (2013). *InTASC Learning Progressions for Teachers 1.0: A Resource for Ongoing Teacher Development.* Retrieved 07 29, 2013, from http://www.ccsso.org/Resources/Programs/Interstate_Teacher_Assessment_Consortium_(InTASC).html

Chapelle, C. A. (2011). Validity argument for language assessment: The framework is simple... *Language Testing, 29*(1), 19-27.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3-13.

Chung, R. R. (2005). *The Performance Assessment for California Teachers (PACT) and Beginning Teacher Development: Can a Performance Assessment Promote Expert Teaching Practice?* Stanford University.

Chung, R. R. (2007, 04). *Beyond the ZPD: When do beginning teachers learn from a high-stakes portfolio assessment?* Retrieved 02 13, 2013, from SCALE: Teacher Performance Assessment Publications: https://scale.stanford.edu/resources/teacher-publications

Chung, R. R. (Winter 2008). Beyond assessment: Performance assessments in teacher education. *Teacher Education Quarterly, 35*(1), 7-28.

Cizek, G. J. (Winter 2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19-27.

Cochran-Smith, M. (2003). Learning and unlearning: the education of teacher educators. *Teaching and Teacher Education, 19*, 5-28.

Cole, N. (1988). A Realist's Appraisal of the Prospects for Unifying Instruction and Assessment. *Assessment in the service of learning: Proceedings of the 1987 ETS invitational conference.* Princeton, NJ: Educational Testing Service.

Commission on Teacher Credentialing. (2009). *Current and Prospective Educators*. Retrieved from http://www.ctc.ca.gov/

Cronbach, L. (1971). Test validation. In R. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443-507). Washington, D.C.: American Council on Education.

Cronbach, L. (1988). Five perspectives on validity argument. In H. Wainer, & H. Braun (Eds.), *Test Validity* (pp. 3-17). Hillsdale, NJ: Laurence Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, Vol 52*(4), 281-302.

Csıkszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience.* New York: Harper and Row.

Darling-Hammond, L. &. (2012, November 5). A better way to grade teachers. *The Los Angeles Times*, p. 3. Retrieved May 13, 2013, from http://scale.stanford.edu/sites/default/files/A%20better%20way%20to%20grade%20teachers%20-%20latimes%20Nov.5.2012.pdf

Darling-Hammond, L. (2006). Assessing teacher education: The usefulness of multiple measures for assessing program outcomes. *Journal of Teacher Education, 57*(2), 120-138.

Darling-Hammond, L. (2006). Securing the right to learn: Policy and practice for powerful teaching and learning. *Educational Researcher, 35*(7), 13-24.

Darling-Hammond, L. (2010). *Evaluating Teacher Effectiveness: How Teacher Performance Assessments Can Measure and Improve Teaching.* Center for American Progress.

Darling-Hammond, L., & Loewenberg Ball, D. (1998). *Teaching for High Standards: What Policymakers Need To Know and Be Able To Do.* Philadelphia, PA: Consortium for Policy Research in Education.

Darling-Hammond, L., & Snyder, J. (2000). Authenitic assessment of teaching in context. *Teaching and Teacher Education, 16*, 523-545.

Darling-Hammond, L., Chung Wei, R., & Johnson, C. M. (2009). Teacher preparation and teacher learning: A changing policy landscape. In G. Sykes (ed.), *The Handbook of Educational Policy Research* (pp. 613-636). Washington, D.C.: American Educational Research Association.

Darling-Hammond, L., Chung Wei, R., Althea, A., Richardson, N., & Orphanos, S. (2009). *Professional Learning in the Learning Profession.* Stanford, CA: National Staff Development Council.

Darling-Hammond, L., Holtzman, D., Gatlin, S., & & Heilig, J. (2005). Does teacher preparation matter? Evidence about teacher certification, Teach for America, and teacher effectiveness.

Education Policy Analysis Archives, 13(42). Retrieved from http://epaa.asu.edu/epaa/v13n42/

Data Quality Campaign. (2013, November). *Action Issues*. Retrieved from Data quality campaign: http://www.dataqualitycampaign.org/

Davies, A. (2011). Kane, validity and soundness. *Language Testing, 29*(1), 37-42.

Davies, A., & Elder, C. (2005). Validity and validation in language testing. In E. Hinkel, *Handbook of Research in Second Language Teaching and Learning* (pp. 795-813). Mahway, NJ: Lawrence Erlbaum.

Delandshere, G., & Arens, S. A. (2001). Representations of teaching and standards-based reform: Are we closing the debate about teacher education? *Teaching and Teacher Education, 17*, 547-66.

Denzin, N. K., & Lincoln, Y. S. (2005). *The Handbook of Qualitative Research.* Thousand Oaks, CA: Sage.

Duckor, B., Castellano, K., Téllez, K., & Wilson, M. (2013). Examining the internal structure of the Performance Assessment for California Teachers: The elementary-literacy teaching event. *American Educational Research Association*, (pp. 1-41). San Francisco.

Ericsson, K. (2004). The influence of experience and deliberate practice on the development of superior expert performance. In K. Ericsson, N. Charness, P. Feltovich, & R. Hoffman, *Cambridge Handbook of Expertise and Expert Performance* (pp. 685-706). Cambridge, UK: Cambridge University Press.

Ericsson, K., & Ward, P. (2007). Capturing the naturally occurring superior performance of experts in the laboratory toward a science of expert and exceptional performance. *Current Directions in Psychological Science, 16*(6), 346-50.

Feinberg, L. (1990). Multiple-choice and its crtics: Are the alternatives any better? *Commentaries from the College Board* (pp. 3-15). New York, NY: College Board.

Fenderson, S. (2010). *Instruction, Perception, and Reflection: Transforming Beginning Teachers' Habits of Mind (Ed.D Thesis).* San Francisco: University of San Francisco.

Field, A. (2009). *Discovering Statistics Using SPSS.* London: SAGE.

Gladwell, M. (2008). *Outliers: The Story of Success.* New York: Little, Brown, and Co.

Goodman, G., Arbona, C., & Dominquez de Rameriz, R. (2008). High-stakes, minimum-competency exams: How competent are they for evaluating teacher competence. *Journal of Teacher Education, 59*(1), 24-39.

Gotch, C. M., & Perie, M. (2012). Using Validity Arguments to Evaluate the Technical Quality of Local Assessment Systems. Vancouver, B.C.: American Educational Research Association.

Graham, M., Milanowski, A., & Miller, J. (2012, February). Measuring and promoting inter-rater agreement of teacher and principal performance ratings. Washington, D.C.: Center for Educator Compensation Reform, Department of Education. Retrieved December 12, 2013, from Center for Educator Reform: http://www.cecr.ed.gov/

Greenberg, J., Pomerance, L., & Walsh, K. (2011). *Student Teaching in the United States.* Washington D.C.: National Council on Quality Teacher Education.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), 255-274.

Grenfell, M. (1998). *Training Teachers in Practice.* Clevedon: Multilingual Matters.

Guion, R. (1980). On trinitarian doctrines of validity. *Professional Psychology, 11*(3), 385-98.

Haertel, E. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5-9.

Haertel, E. (2004). Interpretive argument and validity argument for certification testing: can we escape the need for psychological theory? *Measurement: Interdisciplinary Research and Perspectives, 2*, 175-178.

Hamel, F. (2012). Assuring quality or overwhelming teachers? High quality performance assessment in American pre-service teacher education. *Presentation at the Annual Meeting of the Japan-US Teacher Education Consortium.* Tokyo: Unpublished.

Hammerness, K., & Darling-Hammond, L. (2002). Meeting old challenges and new demands: The redesign of the Stanford teacher education program. *Issues in Teacher Education, 11*(1), 17-30.

Hammersley, M., & Atkinson, P. (1995). *Ethnography: Principles in Practice.* NY: Routledge.

House, E. (1977). *The Logic of Evaluative Judgement.* LA: Center for the Study of Evaluation, University of California.

Humphrey, D., Koppich, J., & Hough, H. (2005). Sharing the Wealth: National Board Certified Teachers and the Students Who Need Them. *13*. Retrieved September 26, 2013, from http://epaa.asu.edu/ojs/article/view/123/249

Jacob, B., & Lefgren, L. (2005). *Principals as Agents: Subjective Performance Measures in Education.* Retrieved from NBER Harvard Faculty Research Working Paper Series: http://www-personal.umich.edu/~bajacob/files/Teacher%20Labor%20Markets/principals%20as%20agents.PDF

Kane, M. (1982). A sampling model for validity. *Applied Psycological Measurement, 6*(2), 125-160.

Kane, M. (1986). Reliability in criterion-referenced tests. *Journal of Educational Measurement, 23*(3), 221-224.

Kane, M. (1990). *An Argument-Based Approach to Validation.* Iowa: ACT Research Report Series.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kane, M. (1992). The assessment of professional competence. *Evaluation and the Health Professions, 15*(2), 163-182.

Kane, M. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluation and the Health Professions, 17*(2), 133-159.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practices, 21*(1), 31-41.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*(3), 135-170.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (pp. 17-64). American Council on Education/Praeger.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (pp. 17-64). American Council on Education/Praeger.

Kane, M. (2009). Validating the interpretations and uses of test scores. In R. Lissitz, *The Concept of Validity: Revisions, New Directions and Applications* (pp. 39-64). Charlotte, NC: Information Age Publishing.

Kane, M. (2011). The errors of our ways. *National Council on Measurement in Education, 48*(1), 12-30.

Kane, M. (2011). Validating score interpretations and uses: Messick lecture, language testing research colloquium, Cambridge April 2010. *Language Testing, 29*(1), 3-17.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Kellor, E. M. (2002). *Performance-Based Licensure in Connecticut.* Madison, WI: Consortium for Policy Research in Education.

King, B. (1991, January). Teachers' views on performance-based assessments. *Teacher Education Quarterly, 18*(3), 109-119.

Kolen, M. J., & Tong, Y. (2010). Psychometric Properties of IRT Proficiency Estimates. *Educational Measurement: Issues and Practice, 29*, 8-14.

LaDuca, A. (1994). Validation of professional licensure examinations: Professional theory, test design, construct validity. *Evaluation and the Health Professions, 17*(2), 178-97.

Lai, E. R., Wei, H., Hall, E. L., & Fulkerson, D. (2012). *Establishing an Evidence-Based Validity Argument for Performance Assessment.* New Jersey: Pearson.

Lang, W. S., & Wilkerson, J. R. (2005). Easy Approaches to Establishing Validity in a Task-Based Teacher Performance Assessment System. Amercian Association of Colleges of Teacher Education.

LeTourneau University. (2013, October 14). *Toulmin's analysis.* Retrieved from Owlet: On-line Writing and Learning: http://owlet.letu.edu/contenthtml/research/toulmin.html

Likert, R. (1932). *A Technique for the Measurement of Attitudes.* New York: Columbia University Press.

Linn, R., Baker, E., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Lit, I. W., & Lotan, R. (2013). A Balancing act: Delimmas of implementing a high-stakes performance assessment. *The New Educator, 9*, 54-76.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*(3), 635-94.

Lomask, M., Seroussi, M., & Budzinski, F. (March 1997). The validity of portfolio-based assessment of science teachers. *NARST* (pp. 1-23). Chicago: Connecticut State Department of Education Bureau of Research and Teacher Assessment.

Lyons, N. (1996). A grassroots experiment in performance assessment. *Educational Leadership, 53*(6), 64-7.

Maxwell, J. A. (1992). Understanding and validity in qualitative research. *Harvard Educational Review, 62*(3), 279-300.

Mayer, D. (2005). Reviving the "policy bargain" discussion: Professional accountability and the contribution of the Teacher Performance Assessment. *The Clearing House, 78*(4), 177-181.

McCormick, R. (2001). How do beginning teachers perceive their development as reflective practitioners? *American Educational Research Association.* Seattle.

McQuillan, J. (2011). *Teacher performance assessment: An opportunity for collaboration, learning, and effecting change in teacher preparation.* Retrieved July 12, 2012, from Success in High-Need Schools Journal: http://www.acifund.org/new/images/pdf/successvol9.pdf

Messick, S. (1975). The standard problem: meaning and values in measurment and evaluation. *American Psychologist, 30*(10), 955-66.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-104). NY: McMillian.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher, 23*(2), 13-23.

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. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Miles, M., & Huberman, A. (1994). *Qualitative Data Analysis* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Mitchell, D., Scott, L., Hendrick, I., & Boyns, D. (1998). *The California Beginning Teacher Support and Assessment Program: 1998 statewide evaluation study.* Riverside, CA: California Educational Research Cooperative.

Moir, E. (2012, September 10). *Op-Ed: For first-year teachers, It's sink or swim.* Retrieved from Takepart: www.takepart.org

Moir, E. (2013, September). *Beginning Teacher Learning Communities: Practice Beliefs.* Retrieved from The New Teacher Center: http://www.newteachercenter.org/sites/default/files/ntc/main/resources/NTCPracticeBrief-BTLC.pdf

Moir, E., Baron, W., Freeman, S., & Petrock, L. (2002). *A Developmental Continuum of Teacher Education.* Santa Cruz, CA: The New Teacher Center.

Morrell, E. (2010). Critical literacy, educational investment, and the blueprint for educational reform: An analysis of the reauthorization of the elementary and secondary education act. *Journal of Adolescent and Adult Literacy, 54*(2), 146-49.

National Board for Professional Teaching Standards. (2013). *2013 Guide to National Board Certification.* Retrieved 07 29, 2013, from http://www.nbpts.org/sites/default/files/documents/Candidate-Center/Guide_to_NB_Certification%203.25.13.pdf

Newman, J. H. (1907). *The Idea of a University.* New York: Logmans, Green and Co.

Newton, P. (2010). Conceptualizing comparability. *Measurement, 8*, 172-9.

Newton, P. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspective, 10*(1-2), 1-29.

Newton, P. (2013). Two kinds of argument? *Journal of Educational Measurement, 50*(1), 105-9.

Newton, P., & Shaw, S. (2013). Standards for talking and thinking about validity. *Psychological Methods, 18*(3), 301-19.

Newton, P., & Shaw, S. (2014). *Validity in Educational and Psychological Assessment.* London: SAGE: Cambridge Assessment.

Newton, S. (2010). *Preservice Performance Assessment and Teacher Early Career Effectiveness: Preliminary Findings on the Performance Assessment for California Teachers.* Stanford, CA: Stanford University, Stanford Center for Assessment, Learning, and Equity.

Office of the Superintendent of Public Instruction. (2013, September). *Certification*. Retrieved from OSPI: https://www.k12.wa.us/

Okhremtcouk, I., Seiki, S., Gilliland, B., Atch, C., Wallace, M., & Kato, A. (2009). Voices of pre-service teachers: Perspectives of the performance assessment for California teachers (PACT). *Issues in Teacher Education, 18*(1), 39-61.

Park, J. (2012). *Developing and Validating an Instrument to Measure College Students' Inferential (Doctoral Dissertation).* Retrieved from http://iase-web.org/documents/dissertations/12.Park.Dissertation.pdf

Partnership for Assessment of Readiness for College and Careers. (2010). *The Partnership for Assessment of Readiness for College and Careers (PARCC) application for the Race to the Top Comprehensive Assessment Systems Competition.* Retrieved from http://www.fldoe.org/parcc/pdf/apprtcasc.pdf

Pearson. (2013, July). *West E test validity and reliability.* Retrieved from Eastern Washington University: http://www.ewu.edu/Documents/CALE/Cross-Campus%20Group/WA_Title%2011%20Validity_Reliability_Pearson.pdf

Pecheone, R. L., & Stansbury, K. (1996). Connecting teacher assessment and school reform. *Elementary School Journal, 97*(2), 163-77.

Pecheone, R., & Chung, R. (2006). Evidence in teacher education: The performance assessment for California teachers (PACT). *Journal of Teacher Education, 57*(1), 22-36.

Pecheone, R., & Chung-Wei, R. (1997). *PACT Technical Report: Summary of Validity and Reliability Studies of the 2003-2004 pilot year.* Stanford, CA: Stanford University.

Peck, C. A., Gallucci, C., & Sloan, T. (May 2010). Negotiating implementation of high-stakes performance assessment policies in teacher education: From compliance to inquiry. *Journal of Teacher Education, 20*(10), 1-13.

Picanco, K., Darragh, J., Tully, D., & Henning, A. (2011). When teachers collaborate, good things happen: Teacher candidate perceptions of the co-teach model for the student teaching internship. *The Association of Independent Liberal Arts Colleges for Teacher Education, 8*, 83-104.

Popham, W. (1990). *Modern Educational Measurement: A Practicioners Perspective (2nd ed).* Englewood Cliffs, NJ: Prentice Hall.

Popham, W. (2005). All About Accountability / Instructional Quality: Collecting Credible Evidence. *Educational Leadership, 62*(6), 80-1.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

Professional Educator Standards Board. (2013, July 23). *Program Support: Standard V (2010) - Residency Teacher.* Retrieved July 29, 2013, from http://program.pesb.wa.gov/review/site-visits/rubrics/2010/standard-5/teacher

Rearick, M. (1997). Portfolio talk: Educational researchers, teachers, teacher educators, and members of the Connecticut Department of Education talk about B.E.S.T. practice in Professional Development: Beginning Educator Support and Training: Teacher educator's perspective. *American Educational Research Association.* Chicago.

Rennert-Ariev, P. (2008). The hidden curriculum of performance-based teacher education. *Teachers College Record, 110*(1), 105-138.

Resnick, L., & Tucker, M. (1990). *Setting a New Standard: Toward an Examination System for the United States.* Pittsburg, PA; Washington, D.C.: Learning Research and Development Center; and National Center on Education and the Economy.

Riggs, M. L., Verdi, M. P., & Arlin, P. K. (2009). A local evaluation of the reliability, validity, and procedural adequacy of the teacher performance assessment exam for teaching credential candidates. *Issues in Teacher Education, 18*(1), 13-37.

Rockoff, J., & Speroni, C. (2011, October 17). *The Value of Subjective and Objective Evaluations of Teacher Effectivness.* Retrieved from PhysOrg: http://phys.org/news/2011-10-subjective-teacher-effectiveness.html

Rockquemore, K. A. (2011, October 17). *Sink or Swim.* Retrieved from Inside Higher Ed: www.insidehighered.com

Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. *Educational Measurement: Issues and Practice*, 7-15.

Sackett, P. (1998). Performance assessment in education and professional certification: Lessons for personnel section? In M. Hakel (Ed.), *Beyond Multiple Choice: Evaluating Alternatives to Traditional Testing for Selection* (pp. 113-129). Hillsdale, NJ: Lawrence Erlbaum Inc.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*(4), 215-227.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2009). Responses to issues raised about validity, bias, and fairness in high-stakes testing. *American Psychologists*, 285-287.

Sandholtz, J. H., & Shea, L. M. (2011). Predicting performance: A comparison of university supervisors' predictions and teacher candidate scores on teaching peformance assessment. *Journal of Teacher Education, 63*(1), 39-50.

Sato, M., Chung, R. R., & Darling-Hammond, L. (2008). Improving teachers' assessment practices through professional development: The case of national board certification. *American Educational Research Journal*, 1-32.

Sawchuk, S. (2012, May 11). *Teacher Performance Assessment Under Scrutiny.* Retrieved 7 16, 2012, from http://blogs.edweek.org/edweek/teacherbeat/2012/05/teacher_performance_assessment

Schilling, S. (2004). Conceptualizing the validity argument: An alternative approach. *Measurement:, 2*, 178-182.

Schön, D. A. (1983). *The Reflective Practitioner: How Professionals Think in Action.* New York: Basic Books.

Schultz, S. E. (2002). Assessing growth in teaching knowledge. *Issues in Teacher Education, 11*(1), 49-63.

Shavelson, R., Baxter, G., & Pine, J. (1992). Performance assessment: Political rhetoric and measurment reality. *Educational Researcher, 38*(3), 22-27.

Shaw, S., & Crisp, V. (2012). An approach to validation: Developing and applying an approach for the validation of general qualifications. *Research Matters: A Cambridge Assessment Publication* (Special Issue 3: An approach to validation), 1-44.

Shaw, S., Crisp, V., & Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy and Practice, 19*(2), 159-176.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-24.

Shimberg, B. (1981). Testing for licensure and certification. *American Psychologist, 36*(10), 1138-46.

Shulman, L. (1992). Toward a pedagogy of cases. In J. Shulman (Ed.), *Case Methods in Teacher Education.* New York: Teacher College Press.

Singer, A. (2013, October 14). *Problems with Pearson's Student Teacher Evaluation System -- It's Like Déjà Vu All Over Again.* Retrieved January 25, 2014, from The Huffington Post: http://www.huffingtonpost.com/alan-singer/problems-with-pearsons-te_b_4093772.html

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477-481.

Smith, K., & Lev Ari, L. (2005). The place of practicum in pre-service teacher education. *Asian Pacific Journal of Teacher Education, 33*(3), 289-302.

Stanford Center for Assessment, Learning and Equity. (2013). *2013 edTPA Field Test: Summary Report.* Stanford, CA: SCALE.

Stanford Center for Assessment, Learning, and Equity. (2012). *edTPA*. Retrieved May 13, 2013, from SCALE: http://scale.stanford.edu/teaching/edtpa

Stanford Center for Assessment, Learning, and Equity. (2012). *Teacher Performance Assessment.* Stanford: Stanford University.

Stansbury, K. (1998). *What is Required for Performance Assessment of Teaching?* San Francisco, CA: WestEd.

Stewart, J. B. (2013, September 14). New metric for colleges: Graduates' salaries. *The New York Times*, p. B1.

Stone, B. (1998). Problems, pitfalls, and benefits of portfolios. *Teacher Education Quarterly, 25*(1), 105-114.

Stotsky, S. (2006). Who should be accountable for what beginning teachers need to know? *Journal of Teacher Education, 57*(3), 256 – 268.

Suen, H. K., & Davey, B. (1990). Potential theoretical and practical pitfals and cautions of the performance assessment design. *American Educational Research Association.* Boston, MA.

Talbot, R. M., & Briggs, D. C. (2007). Does theory drive the items or do items drive the theory? *Measurement: Interdisciplinary Research and Perspectives, 5*(2-3), 205-208.

Taylor, C. (1994). Assessment for measurement or standards: The peril and promise of large-scale assessment reform. *American Educational Research Journal, 31*(2), 231-262.

Toulmin, S. (1958). *The Uses of Argument.* Cambridge: Cambridge University Press.

Uhlenbeck, A., Verloop, N., & Beijaard, D. (2002). Requirements for an assessment procedure for beginning teachers: Implications from recent theories on teaching and assessment. *Teachers College Record, 104*(2), 242-272.

US Department of Agriculture. (2013, July 10). *Supplemental Nutrition Assistance Program (SNAP)*. Retrieved from http://www.fns.usda.gov/

Verheij, B. (2005). Evaluating arguments based on Toulmin's scheme. *Argumentation, 19*, 347-371.

Volmer, M. L., & Crek, R. J. (1993). Teacher Asesment: A Continuing Controversy. *American Asociation of Coleges for Teacher* (pp. 1-13). San Diego, CA: ERIC: ED 356 211.

Vygotsky, L. (1978). Interaction between learning and development. In L. Vygotsky, *Mind and Society* (pp. 79-91). Cambridge, MA: Harvard University Press.

WACTE. (2012, April). *TPA special session*. Pullman, WA: Washington Association of Colleges of Teacher Education.

Walling, B., & Lewis, M. (2000). Development of professional development pre-service teachers: Longitudinal and comparative analysis. *Action Teacher Education, 22*(2a), 63-67.

Washington Professional Educator Standards Board. (2013). *The Purpose and Roles of the Professional Educator Standards Board*. Retrieved May 13, 2013, from http://www.pesb.wa.gov/mission/purpose

Wei, R. C. (2010). Assessment for learning in preservice teacher education: Performance based assessments. In M. Kennedy (Ed.), *Teacher Assessment and the Quest for Teacher Quality: A Handbook* (pp. 69-132). San Francisco, CA: Jossey Bass.

Whittaker, A., & Young, V. M. (2002). Tensions in assessment design: Professional development under high-stakes accountability. *Teacher Education Quarterly, 29*(3), 43-60.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9).

Wiggins, G., & McTighe, J. (2000). *Understanding by Design.* Boston, MA: Pearson Education.

Wilson, M., Hallam, P., Pecheone, R., & Moss, P. (n.d.). *Using Student Achievement Test Scores as Evidence of External Validity for Indicators of Teacher Quality: Connecticut's Beginning Educator Support and Training Program.* Retrieved 07 29, 2013, from https://scale.stanford.edu/resources/teacher-publications

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147-70.

Yin, R. (1994). *Case study research: Design and methods (2nd ed.).* Thousand Oaks, CA: Sage.

Youngs, P., Odden, A., & Porter, A. C. (2003). State policy related to teacher licensure. *Educational Policy, 17*(2), 217-236.

Zeichner, K. (2003). The adequacies and inadequacies of three current strategies to recruit, prepare, and retain the best teachers for all students. *Teachers College Record, 105*(3), 490-519.