BUKHARI, NURLIYANA, Ph.D. An Examination of the Impact of Residuals and Residual Covariance Structures on Scores for Next Generation, Mixed-Format, Online Assessments with the Existence of Potential Irrelevant Dimensions under Various Calibration Strategies. (2017)
Directed by Dr. Richard Luecht and Dr. Micheline Chalhoub-Deville. 287 pp.

In general, newer educational assessments are deemed more demanding challenges than students are currently prepared to face. Two types of factors may contribute to the test scores: (1) factors or dimensions that are of primary interest to the construct or test domain; and, (2) factors or dimensions that are irrelevant to the construct, causing residual covariance that may impede the assessment of psychometric characteristics and jeopardize the validity of the test scores, their interpretations, and intended uses. To date, researchers performing item response theory (IRT)-based model simulation research in educational measurement have not been able to generate data, which mirrors the complexity of real testing data due to difficulty in separating different types of errors from multiple sources and due to comparability issues across different psychometric models, estimators, and scaling choices.

Using the context of the next generation K-12 assessments, I employed a computer simulation to generate test data under six test configurations. Specifically, I generated tests that varied based on the sample size of examinees, the degree of correlation between four primary dimensions, the number of items per dimension, and the discrimination levels of the primary dimensions. I also explicitly modeled the potential nuisance dimensions in addition to the four primary dimensions of

interest, for which (when two nuisance dimensions were modeled) I also used varying degrees of correlation. I used this approach for two purposes. First, I aimed to explore the effects that two calibration strategies have on the structure of residuals of such complex assessments when the nuisance dimensions are not explicitly modeled during the calibration processes and when tests differ in testing configurations. The two calibration models I used included a unidimensional IRT (UIRT) model and a multidimensional IRT (MIRT) model. For this test, both models only considered the four primary dimensions of interest. Second, I also wanted to examine the residual covariance structures when the six test configurations vary. The residual covariance in this case would indicate statistical dependencies due to unintended dimensionality.

I employed Luecht and Ackerman's (2017) expected response function (ERF)-based residuals approach to evaluate the performance of the two calibration models and to prune the bias-induced residuals from the other measurement errors. Their approach provides four types of residuals that are comparable across different psychometric models and estimation methods, hence are 'metric-neutral'. The four residuals are: (1) e0, which comprises the total residuals or total errors; (2) e1, the bias-induced residuals; (3) e2, the parameter-estimation residuals; and, (4) e3, the estimated model-data fit residuals.

With regard to my first purpose, I found that the MIRT model tends to produce less estimation error than the UIRT model on average (e2MIRT is less than e2UIRT) and tends to fit the data better than the UIRT model on average (e3MIRT is

less than e3UIRT). With regard to my second research purpose, my analyses of the correlations of the bias-induced residuals ($r_{e1_i, e1_h}$) provide evidence of the large impact of the presence of nuisance dimension regardless of its amount. On average, I found that the residual correlations ($r_{e1_i, e1_h}$) increase with the presence of at least one nuisance dimension but tend to decrease with high item discriminations.

My findings shed light on the need to consider the choice of calibration model, especially when there are some intended and unintended indications of multidimensionality in the assessment. Essentially, I applied a cutting-edge technique based on the ERF-based residuals approach (Luecht & Ackerman, 2017) that permits measurement errors (systematic or random) to be cleanly partitioned, understood, examined, and interpreted—in-context and in relative to difference-that-matters criteria—regardless of the choice of scaling, calibration models, and estimation methods. For that purpose, I conducted my work based on the context of the complex reality of the next generation K-12 assessments and based on my effort to maintain adherence to the established educational measurement standards (American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), 1999, 2014); International Test Commission (ITC) (ITC, 2005a, 2005b, 2013a, 2013b, 2014, 2015)).

AN EXAMINATION OF THE IMPACT OF RESIDUALS AND RESIDUAL COVARIANCE STRUCTURES ON SCORES FOR NEXT GENERATION, MIXED-FORMAT, ONLINE ASSESSMENTS WITH THE EXISTENCE OF POTENTIAL IRRELEVANT DIMENSIONS UNDER VARIOUS CALIBRATION STRATEGIES

by

Nurliyana Bukhari

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2017

Approved by

Richard M. Luecht
Committee Co-Chair

Micheline B. Chalhoub-Deville
Committee Co-Chair

To my family in Malaysia, Jordan, and America.

APPROVAL PAGE

This dissertation written by BUKHARI NURLIYANA has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Co-Chairs _____
Richard M. Luecht

_____
Micheline B. Chalhoub-Deville

Committee Members _____
Randall D. Penfield

_____
Allison J. Ames

_____
Rosna Awang-Hashim

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

In the Islamic tradition, acquiring knowledge is the most important activity a person can embark on to live a meaningful life. In addition to the seekers of knowledge, the ones who help pave the way, provide empathy, love, and facilitation to the seekers of knowledge are equally important and are elevated in status and in quality. As a graduate student in America, I have received unwavering support and encouragement for this research and this important journey from several individuals who I wish to acknowledge.

First, I would like to express my sincere gratitude to Dr. Micheline Chalhoub-Deville and Dr. Ric Luecht, who are the best mentoring team, role models, advisors, supervisors, critics, and Professors I could have hoped for. I am truly grateful for their continuous support for my Ph.D. studies and related research, for their patience, motivation, flexibility, immense knowledge, and guidance in coursework, researches, and in writing this dissertation. I would also like to thank my dissertation committee—Dr. Randy Penfield, Dean of School of Education at UNCG; Dr. Allison Ames, an external member from James Madison University, Virginia; and Dr. Rosna Awang-Hashim, an overseas external member from Universiti Utara Malaysia (UUM), Malaysia—for their undivided support over the past several years in committee activities and as exemplars of the academic profession.

In addition, other professors have provided invaluable advice and learning opportunities, including Dr. Bob Henson, Dr. John Willse, Dr. Terry Ackerman, Dr. Jill

Chouinard, Dr. Holly Downs, Dr. Devdass Sunnassee, Dr. Nicholas Myers, Dr. Soyeon Ahn, Dr. Valentina Kloosterman, and Dr. Jaime Maerten-Rivera.

My journeys as a graduate student would not have been smooth were it not for Dr. Mohamed Mustafa Ishak, Vice Chancellor of UUM; Dr. Yahya Don, Dean of School of Education and Modern Languages (SEML), at UUM; Dr. Mohd Izam Ghazali, former Dean of SEML at UUM; Christina Groves, the Administrative Support Associate (ASA) in the Educational Research and Methodology (ERM) department at UNCG; Valeria Cavinnes, the Electronic Thesis & Dissertation Administrator from the UNCG Graduate School; Rachel Hill, former ASA in the ERM department; Cheryl Kok Yeng, a Malaysian graduate student at University of Miami who helped me a lot when I first arrived in America; and Dr. Norhafezah Yusof, a friend at UUM. Thank you very much to my funders, the Government of Malaysia (through the Ministry of Higher Education) and the Universiti Utara Malaysia as well as to my financial guarantors for Ph.D., Mohamad Raffizal Mohamed Yusof and Norshamsul Kamal Ariffin Abdul Majid (who is also my maternal uncle).

I would also want to extend my gratitude to Jonathan Rollins and Shuying Sha, who are my very good friends and close colleagues in the ERM department, for their contributions to my academic progress and helping me smile even when I was not sure I could. Other graduate students in and outside of the department have also played an integral role in my journey and it would be remiss of me to not mention: Ayu Abdul Rahman, Jian Gou, Chai Hua Lin, Zhongtian Lin, Tini Termitini, Sharifah

Nadiah Syed Mukhiar, Muhammad Halwani Hasbullah, Cheryl Thomas, Oksana Naumenko, Shufen Hung, Emma Sunnassee, Lindsey Varner, Yan Fu, Meltem Yumsek, Julianne Zemaitis, Juanita Hicks, Robyn Thomas, Tala Mirzaei, Gilbert Ngerano, Tyler Strachan, Bruce Mccollaum, Thomas McCoy, Jia Lin, and Saed Qunbar.

Finally, and most importantly, my family who has provided me with the love I needed to finish. My husband, Ahmed Rbeihat, who has been very supportive and understanding, helping me celebrate each moment, providing unconditional love during the process. My parents, Bukhari Abdullah and Norkazimah Abdul Majid, and my in-laws, Omer Rbeihat, Nadia Nasah, and Maha Harasis, have been my biggest cheerleaders, as well as helping me with their continuous prayers and words of wisdom. Also, to all my brothers and sisters in Malaysia (Norsyamimi Bukhari and Rahmat Asyraf Bukhari), Jordan (Maram Rbeihat, Abdul Rahman Rbeihat, Fatimah Rbeihat, and Rifqah Rbeihat), and America (The Uzzaman's family, The Abdel Karim's family, and The Ghazali's family) who constantly motivate me. Yatie Thaler in Sunrise, Florida; Ummu and Abu Naufal in Miami; Gurinder Kaur and family in Albany, New York; The Rbehat family in Raleigh; Munib Nasah in Boston; and Nor Othman-Lesaux and husband in Greensboro have provided second homes for me and much-needed escapes here in America. Last but not least, to my four-year old son, Umar, who has endured a three-year separation away from his parents during this pursuit of knowledge, Ummi (and Abbi) will see you and be with you soon! *InshaaAllah*.

TABLE OF CONTENTS

Page

LIST OF TABLES ix

LIST OF FIGURES xii

CHAPTER

I. INTRODUCTION 1

Concept of Validity: A Brief Description 3
The Next Generation Assessments 7
Universal Design Principle 29
Description of Problem 31
Purposes of Research 35
Research Questions 37
Organization of the Study 38
A Note on Terminology 39

II. LITERATURE REVIEW 40

Item Response Theory 42
Mixed-Format Tests 97
Potential Sources of Construct-Irrelevant Variance in Scores Reporting 111
English Language Proficiency and Content Performance for K-12 English Language Learner Students 121
Summary in the Context of Current Research 127

III. METHODS 134

Constant of Study 135
Conditions of Study 135
Data Generation 140
Structure of the Generated Response Data 148
Item Parameter Estimation 154
Estimation/Scoring of Latent Ability 154
Criteria for Evaluating the Results: Luecht and Ackerman's Expected Response Function Approach 155

IV. RESULTS 166

Results for Research Question 1: Comparison of the ERF-Based Residuals for MIRT and UIRT Models 168
Results for Research Question 2: Examination of Bias-Induced Residual Covariance (*r*e1*i*,e1*h*) 194

V. CONCLUSIONS 219

Summary of Findings and Implications for Practice 219
Discussion 225
Limitations and Directions for Future Research 230

REFERENCES 233

APPENDIX A. SELECTED-RESPONSE ITEM FORMAT FROM SMARTER BALANCED ASSESSMENT CORPORATION ELA ITEM DESIGN TRAINING MODULE (RETRIEVED ON DECEMBER, 2015) 282

APPENDIX B. TECHNOLOGY-ENHANCED ITEM FORMAT FOR ELA. PEARSON EDUCATION: PARTNERSHIP FOR ASSESSMENT OF READINESS FOR COLLEGE AND CAREERS (PARCC) ASSESSMENT (2015) 283

APPENDIX C. TWO TYPES OF SBAC ITEM FORMATS: (A) TECHNOLOGY-ENABLED ITEM FORMAT, (B) TECHNOLOGY-ENHANCED ITEM FORMAT FROM SBAC MATHEMATICS ITEM DESIGN TRAINING MODULE (RETRIEVED ON DECEMBER, 2015) 284

APPENDIX D. GRIDDED RESPONSE ITEM FORMAT (STATE OF FLORIDA DEPARTMENT OF EDUCATION, 2013) 285

APPENDIX E. DESCRIPTIVE STATISTICS FOR CONDITIONAL e0 (BASED ON PERCENTAGE SCORES) 286

LIST OF TABLES

Page

Table 1. Sample of Technology-Enhanced (TE) Item Formats based on Examinees' Interactions 13

Table 2. Two by Two Table for Observed Frequencies 76

Table 3. Example of a Two by Two Table for Observed Frequencies 76

Table 4. Two by Two Table for Expected Frequencies 77

Table 5. Summary of Item Formats from Partnership for Assessment of Readiness for College and Careers Consortium (PARCC) Assessments 98

Table 6. Summary of Item Formats from Smarter Balance Assessment Consortium (SBAC) Assessments 99

Table 7. Partial Correlations among Language Domain & Content Scores from Wolf & Faulkner-Bond (2016) Study 126

Table 8. Summary of Potential Issues in Educational Assessments (Not Limited to Next Generation Assessments) 130

Table 9. Summary of the Relevant Literature to Provide Context for Simulation Study 131

Table 10. Complete Simulation Design 136

Table 11(a). Structure of Sigma for Test Format when There is No Nuisance Dimension or One Nuisance Dimension 145

Table 11(b). Structure of Sigma for Test Format when There are Two Nuisance Dimensions 146

Table 12. Summary & Corresponding Rationale of the Constant & Conditions of Simulation Study 153

Table 13. Descriptions & Operational Definitions of Residuals used as Criteria to Answer the Research Question 167

Table 14. Descriptive Statistics for Conditional e2MIRT (based on Percentage Scores) for All Crossed Conditions 171

Table 15. Descriptive Statistics for Conditional e2UIRT (based on Percentage Scores) for All Crossed Conditions 172

Table 16. Descriptive Statistics for Conditional e3MIRT (based on Percentage Scores) for All Crossed Conditions 173

Table 17. Descriptive Statistics for Conditional e3UIRT (based on Percentage Scores) for All Crossed Conditions 174

Table 18. Descriptive Statistics for e2 and e3 Residuals from MIRT and UIRT Calibrations Given the Amount of Nuisance Dimension 176

Table 19. Descriptive Statistics for e2 and e3 Residuals from MIRT and UIRT Calibrations Given the Strength of Correlations between Nuisance Dimensions 179

Table 20. Descriptive Statistics for e2 and e3 Residuals from MIRT and UIRT Calibrations Given the Strength of Correlations between Primary Dimensions 182

Table 21. Descriptive Statistics for e2 and e3 Residuals from MIRT and UIRT Calibrations Given Different Item Discrimination Levels on the Primary Dimensions 185

Table 22. Descriptive Statistics for e2 and e3 Residuals from MIRT and UIRT Calibrations Given Number of Items in each Primary Dimensions 188

Table 23. Descriptive Statistics for e2 and e3 Residuals from MIRT and UIRT Calibrations Given Different Sample Sizes 190

Table 24. Summary Table for Two-Factorial ANOVA on the Conditional Mean e2UIRT 192

Table 25. Descriptive Statistics for Conditional e1 (based on Percentage Scores) for All Crossed Conditions 198

Table 26. Descriptive Statistics for the Bias-Induced Residual Correlations for All Crossed Conditions 199

Table 27. Descriptive Statistics for Bias-Induced Residual Correlations Given the Amount of Nuisance Dimension 203

Table 28. Descriptive Statistics for Bias-Induced Residual Correlations Given the Strength of Correlations between Nuisance Dimension 204

Table 29. Descriptive Statistics for Bias-Induced Residual Correlations Given the Strength of Correlations between Primary Dimensions 206

Table 30. Descriptive Statistics for Bias-Induced Residual Correlations Given Different Item Discrimination Levels on the Primary Dimensions 207

Table 31. Descriptive Statistics for Bias-Induced Residual Correlations Given Number of Items in each Primary Dimensions 208

Table 32. Descriptive Statistics for Bias-Induced Residual Correlations Given Different Sample Sizes 209

Table 33. Summary Table for Two-Factorial ANOVA on the e1 Correlations 212

Table 34. Summary Table for Two-Factorial ANOVA on the e1 Correlations 214

Table 35. Summary Table for Two-Factorial ANOVA on the e1 Correlations 214

Table 36. Summary Table for Two-Factorial ANOVA on the e1 Correlations 217

Table 37. Academic Achievement Descriptors and Cut Scores for North Carolina End-of-Grade Math Test for Year 2013/2014 221

LIST OF FIGURES

Page

Figure 1. Relationships & Convergences Found in the CCSS for Mathematics, CCSS for ELA/Literacy, & the Science Framework (Lee, Quinn, & Valdes, 2013) 8

Figure 2. Taxonomy of Item Types based on Level of Constraint 15

Figure 3. The Intermediate Constraint (IC) Taxonomy for E-Learning Assessment Questions & Tasks 16

Figure 4. Cummins' (1994) Four-Quadrant Framework 19

Figure 5. Dutro & Moran's (2003) Conceptual Model from CALP to Functions, Forms, & Fluency 21

Figure 6. Three Different Item Pattern Matrices for a Test with 40 Items 143

Figure 7. Schematic Diagram of the Structure of Generated Data for 10 Items per Subtest with No Nuisance Dimension 149

Figure 8. Schematic Diagram of the Structure of Generated Data for 10 Items per Subtest with the Presence of One Nuisance Dimension 151

Figure 9. Schematic Diagram of the Structure of Generated Data for 10 Items per Subtest with the Presence of Two Nuisance Dimension 152

Figure 10. Distributions of Conditional Mean Residuals (based on Percentage Scores) Given the Amount of Nuisance Dimension 177

Figure 11. Distributions of Conditional Mean Residuals (based on Percentage Scores) Given the Strength of Correlations between Nuisance Dimensions 180

Figure 12. Distributions of Conditional Mean Residuals (based on Percentage Scores) Given the Strength of Correlations between Primary Dimensions 183

Figure 13. Distributions of Conditional Mean Residuals (based on Percentage Scores) Given Different Item Discrimination Levels on the Primary Dimensions 186

Figure 14. Distributions of Conditional Mean Residuals (based on Percentage Scores) Given Number of Items in each Primary Dimensions 188

Figure 15. Distributions of Conditional Mean Residuals (based on Percentage Scores) Given Different Sample Sizes 190

Figure 16. Profile Plots for Two-Factorial ANOVA on the Conditional Mean e2UIRT 192

Figure 17(a). Distribution of Bias-Induced Residual Correlations with the Existence of Two Nuisance Dimensions for 1,000 Examinees and for the Crossed Conditions of the Remaining Four Testing Conditions 200

Figure 17(b). Distribution of Bias-Induced Residual Correlations with the Existence of Two Nuisance Dimensions for 5,000 Examinees and for the Crossed Conditions of the Remaining Four Testing Conditions 201

Figure 18. Distribution of Bias-Induced Residual Correlations Given the Amount of Nuisance Dimension 203

Figure 19. Distribution of Bias-Induced Residual Correlations Given the Strength of Correlations between Nuisance Dimensions 205

Figure 20. Distribution of Bias-Induced Residual Correlations Given the Strength of Correlations between Primary Dimensions 206

Figure 21. Distribution of Bias-Induced Residual Correlations Given Different Item Discrimination Levels on the Primary Dimensions 207

Figure 22. Distribution of Bias-Induced Residual Correlations
Given Number of Items in each
Primary Dimensions 208

Figure 23. Distribution of Bias-Induced Residual Correlations
Given Different Sample Sizes 209

Figure 24. Profile Plots for Two-Factorial ANOVA
on the e1 Correlations 213

CHAPTER I

INTRODUCTION

Standardized tests are one of the most important measurement tools in educational assessment. Scores from such tests are useful in various decision-making processes, including school accountability and high school graduation as well as college and graduate school admissions. Over the past several decades, testing has been dramatically transformed, especially in the United States. Researchers, test users and stakeholders have demonstrated an interest in discussing available approaches for the rapid development of and employment of innovations in standardized assessments.

One area of assessment innovation is in the use of technologies and computers to deliver exams (Drasgow, 2016; Lissitz & Jiao, 2012) as computer-based testing (CBT) and automated scoring have begun to replace the paper and pencil test system with opscan test grading. When an assessment program is administered via computer, new measurement opportunities and new approaches for testing students are available. Tests can be designed to measure wider test constructs, content areas, domain skills, strands, attributes, and cognitive processes using different item and response formats (Masters, Famularo, & King, 2015; Parshall & Harmes, 2009) and different scoring procedures (Bukhari,

Boughton, & Kim, 2016; Stark, Chernyshenko, & Drasgow, 2002) beyond simple correct-incorrect scoring.

While the various testing features offered by such innovations have been considered to be advantages, testing practices have become more complex, challenging, demanding, and more risky. For instance, the test development process (Downing & Haladyna, 2006) has become more complicated with more elaborate conceptions of the constructs, the requirements from test specifications in terms of test content and skills, item types and scoring, test lengths, and other statistical characteristics (Schmeiser & Welch, 2006). van der Linden (2005) suggested that computerized test assembly procedures often require hundreds of constraints that must be met during the item selection process for a given test.

In addition to the assessment innovations, issues such as fairness and accountability have begun to receive more attention due to the transformation of testing practices, especially with the No Child Left Behind (NCLB) legislation of 2001. The NCLB was an Act passed by the US Congress which reauthorized the 1965 Elementary and Secondary Education Act and which was itself replaced by the-Every Student Succeeds Act (ESSA) in 2015. However, the impact of NCLB has been long lasting. The intent of the NCLB was the improvement of individual outcomes in education. Under the NCLB, every state was required to develop an accountability assessment system to measure statewide progress and evaluate school performance. NCLB contained a further requirement for academic assessments to be fair, equal, and provide significant opportunity for all children (including students

with disadvantages and students with limited English proficiency) to reach proficiency on challenging academic achievement standards and state academic assessments (NCLB, 2001a: Public Law, 107-110, Title I, January, 2002; NCLB, 2001b: Public Law, 107-110, Title III, January, 2002).

Concept of Validity: A Brief Description

The transformation of standardized testing is indeed due to the innovations in assessment, increased levels of academic achievement standards, and the presence of diverse subpopulations of test takers. It is critical to ensure that a given test, with such complexity, is meeting its intended purposes, uses, and interpretations, hence is valid.

Messick, in his seminal article on validity (1989) stated that "[v]alidity is an integrated evaluative judgment of the degree to which empirical evidence & theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment" (p. 13). He declared that construct validity is a combination of the study of a construct and its relationships to other constructs and observables, also referred as a nomological network (Cronbach & Meehl, 1955; Embretson, 1983). Thus, the concept of construct validity is a fundamentally unified or unitary framework that within itself includes three types of validity: criterion-related, content, and construct. In other words, construct validity is not just the study of the construct in isolation (Messick, 1989). Others have stated this differently:

> [In criterion-oriented validation,] the investigator is primarily interested in some criterion which he wishes to predict. … If the criterion is obtained some time after the test is given, he is studying predictive validity. If the test score and criterion score are determined essentially the same time, he is studying concurrent validity… Content validity is established by showing that the test items are a sample of a universe in which the investigator is interested. Content validity is ordinarily to be established deductively, by defining a universe of items and sampling systematically within this universe to establish the test. Crobach & Meehl (1955, p. 282)

This distinction is also stated as follows:

> Construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are a means of confirming or disconfirming the claim. (Cronbach & Meehl, 1955, p. 290)

One criticism of the broad framework of validity as a nomological network is that it does not illustrate how to assess the construct validity in practical terms (e.g., Kane, 2004, 2006; Lissitz & Samuelson, 2007a, 2007b). Kane (2004), acknowledging that the difficulty of applying validity theory to testing programs is "exacerbated by the proliferation of many different kinds of validity evidence and by the lack of criteria for prioritizing different kinds of evidence" (p. 136), introduced an argument-based approach to validity: "[a] methodology for evaluating the validity of proposed interpretations and uses of test scores" (p. 166). According to Kane (2006) (also see Kane, 2013), validation employs two kinds of arguments: (1) the development of an interpretive argument that determines the proposed interpretations and uses of test results by identifying the inferences and

assumptions; and, (2) the validity argument that provides an evaluation of the interpretive argument which claims that a proposed interpretation is valid by affirming that the interpretive argument is clear and coherent, the inferences are logical, and the assumptions are plausible.

Lissitz and Samuelson (2007a, 2007b) suggested a systematic structural view of test evaluation that is categorized into internal and external aspects. They emphasized the importance to prioritize the internal aspects of test evaluation that focus on practical content, theoretical latent process, and reliability, before moving on to evaluate the external aspects which are concerned with on the nomological network, practical utility, and impact. They believed that it is of paramount importance to first focus on the content elements of the assessment, their relationships, and the student behavior and cognitions that relate to those elements as they are being processed (i.e., cognitive theories of cognitive processes). Lissitz and Samuelson's (2007a) presentation of validity has received mixed responses from validity scholars (Chalhoub-Deville, 2009; Embretson 2007; Gorin, 2007; Kane, 2009; Mislevy, 2007; Moss, 2007; Sireci, 2007, 2009). Although the scholars agreed that the concept of content validity stressed by Lissitz and Samuelson (2007a) is promising (Moss, 2007), easier to describe and understand (Gorin, 2007, Sireci, 2007, 2009), establishes test meaning (Embretson, 2007), and is useful and critical in assessment design and in enhancing quality of test scores (Mislevy, 2007; Chalhoub-Deville, 2009), some feel that Lissitz and Samuelson's (2007a) conceptualization of validity is moving backward (Gorin, 2007) to traditional

cognitively grounded testing practices (Chalhoub-Deville, 2009) and is ignoring the socio-cognitive aspects of testing (Chalhoub-Deville, 2009; Mislevy, 2007). Researchers also have argued that focusing solely on content validity is insufficient and oversimplified (Embretson, 2007; Kane, 2009) and moves against the mainstream conceptions of validity that are already well-established (Sireci, 2007, 2009).

> At first, validity was viewed as *a characteristic of the test*. It was then recognized that a test might be put to multiple uses and that a given test might be valid for some uses but not for others. That is, validity came to be understood as *a characteristic of the interpretation and use of test scores*, and not of the test itself, because the very same test (e.g., reading test) could be used to predict academic performance, estimate the level of an individual's proficiency, and diagnose problems. Today, validity theory *incorporates both* test interpretation and use (e.g., intended and unintended social consequences) (The National Research Council, 2002, p. 35, emphasis added).

Several established professional testing standards that are internationally recognized, such as the Standards for Educational and Psychological Testing (hereafter *Standards*, the American Educational Research Association (AERA), the American Psychological Association (APA), and theNational Council on Measurement in Education (NCME), 1999, 2014) and the International Test Commission (ITC) (ITC, 2005a, 2005b, 2013a, 2013b, 2014, 2015) are available to ensure best testing practices. These professional standards contain sets of statements, recommendations, guides, and guidelines that are carefully constructed to provide guidance for the development and evaluation of best testing practices

and to suggest criteria for assessing the validity of interpretations of test scores for the intended test uses (see also Kane, 2013). The 2014 *Standards* (AERA, APA, & NCME, 2014) consists of three major parts: Foundations, Operations, and Testing Application. The first chapter in the Foundations part is about validity, where the five sources of validity evidence framework are delineated. The five sources are: (1) evidence based on test content, (2) evidence based on response processes, (3) evidence based on internal structure, (4) evidence based on relations to other variables, and (5) evidence for validity and consequences of testing.

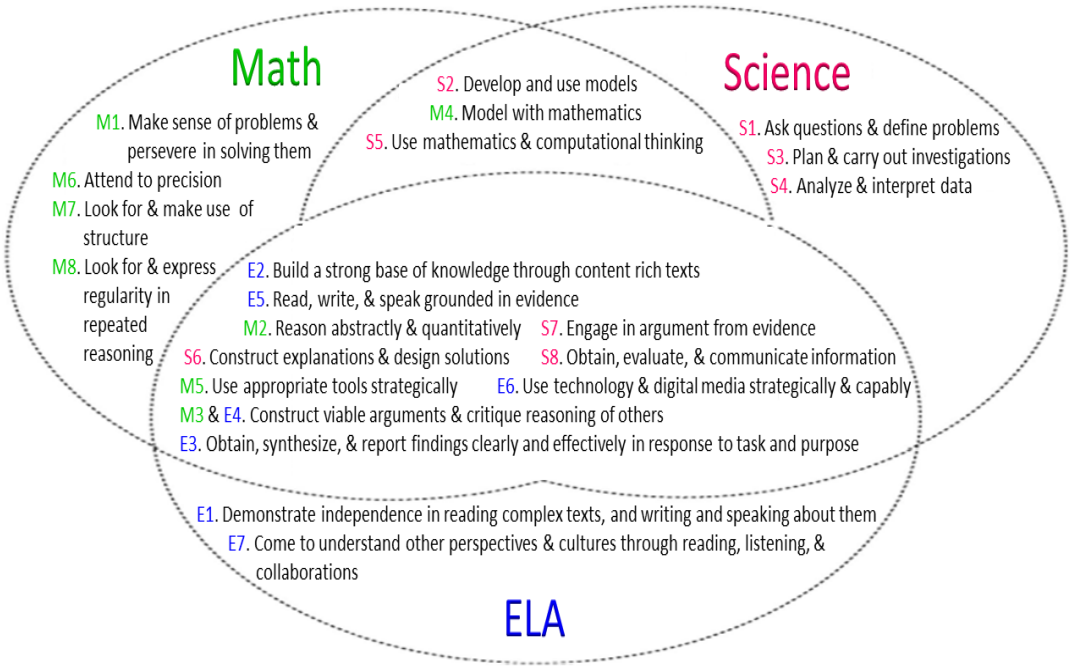
These five sources from the *Standards* (AERA et al., 1999, 2014) integrate closely with the unitary framework of construct validity (Cronbach & Meehl, 1955; Embretson, 1983; & Messick, 1989) and are in line with Kane's (2004, 2006, 2013) argument-based validation framework to support the interpretations and uses of test scores. On the other hand, Lissitz & Samuelson's (2007a) call to prioritize the internal aspects of test evaluation partially meets (Sireci, 2007, 2009) the validity evidence by the *Standards* (AERA et al., 1999, 2014) in that it only relates to the first three sources of validity evidence from the test content, examinees' response processes, and the internal structure of the test, respectively.

The Next Generation Assessments

As mentioned previously, there has been a rapid increase in the implementation of CBT across the US. Not surprisingly, the popularity of CBT will result in its use as the primary testing mode in the future (Drasgow, 2016; Lissitz &

Jiao, 2012). This is especially true with the implementation of the Common Core State Standards (CCSS) for the K-12 ELA/Literacy and math assessments (National Governors Association Center for Best Practices, Council of Chief State School Officers (CCSSO), 2010a, 2010b) and the Next Generation Science Standards (NGSS) for K-12 science assessment (NGSS Lead States, 2013). The CCSS in ELA also defines literacy standards for history/social studies, science, and technical subjects at the secondary level. Figure 1 illustrates the relationships and convergences found in the CCSS for Mathematics, CCSS for ELA/Literacy, and the Science Framework (Lee, Quinn, & Valdes, 2013).

Figure 1. Relationships & Convergences Found in the CCSS for Mathematics, CCSS for ELA/Literacy, & the Science Framework (Lee, Quinn, & Valdes, 2013)



The purpose of the CCSS is to prepare children for success in college and the workplace through the use of College and Career Readiness (CCR) assessments which detect and measure students' proficiencies in high level analytic practices of thinking and acting on knowledge. In other words, the assessments probe deeper into what students are learning in subject domains and how they are learning it. These next generation CCR assessment systems (aligned with CCSS) are currently being developed by the two multistate assessment consortia in the US: the Partnership for Assessment of Readiness for College and Careers Consortium (PARCC) and the Smarter Balanced Assessment Consortium (SBAC). Test developers from the two consortia employ CBT to assess students using more rigorous assessments that combine objective testing and assessment on complex performance tasks.

Different Item Formats in Assessments

Objective testing (e.g., Flanagan, 1939; Hambleton & Swaminathan, 1985, 2010; Lindquist, 1951; Lord and Novick, 1968; Thorndike, 1971) is often fairly straightforward and has become the mainstream in educational assessments since the 1930s (see Stufflebeam, 2001) due to its efficiency and simplicity. It is based on the standardized, norm-referenced testing programs which employ the conventional selected-response (SR) item formats that require examinees to select one best answer from a list of several possible answers. Objective tests are practical with large number of examinees, and are cost efficient (Wainer & Thissen, 1993) in

terms of development, administration, and scoring, but tend to provide only indirect, partial indicators of educational outcomes (Downing, 2006b; Kane, Crooks, & Cohen, 1999).

At the opposite end of the testing continuum, performance assessment (PA) (e.g., Bachman & Palmer, 1996, 2010; Linn, 1993; Linn & Burton, 1994; Messick, 1994; Resnick & Resnick, 1992) seems to have more to offer. PAs enable test takers to "demonstrate the skills the test intended to measure by doing tasks that require those skills" (*Standards*, AERA et al., 2014, p. 221). Several examples of PA include essay composition in writing assessment, science experiments and observations, and derivations of mathematical proofs and arguments. Nonetheless, the performance tasks being assessed are often too complex and highly contextualized (Bachman, 2002; Bachman & Palmer, 1996, 2010; Chalhoub-Deville, 2001), resulting in low generalizability and reliability of the scores (Brennan & Johnson, 1995; Kane, Crooks, & Cohen, 1999; Linn & Burton, 1994; Shavelson, Baxter, & Gao, 1993). Also, such lengthy tasks often require longer test administration, are costly (Wainer & Thissen, 1993; Wainer & Feinberg, 2015), and are difficult to score and standardize (Kane, Crooks, & Cohen, 1999; Lane & Stone, 2006).

Innovation in CBT has empowered the development of various technology-enhanced (TE) item formats that are perceived as an integration (Millman & Greene, 1989; Scalise, 2012; Schmeiser & Welch, 2006) of objective tests and PAs. TEs are computer-delivered test items that require students to engage in specialized interactions to record their responses. Eminent testing programs (Masters et al.,

2015; Poggio & McJunkin, 2012; Zenisky & Sireci, 2001) have been developing different formats of TE items (Clauser, Margolis, & Clauser, 2016; Scalise, 2012) for different operational and field testing purposes (Wan & Henley, 2012) and subject domains (Bukhari et al., 2016; Poggio & McJunkin, 2012) across different populations of examinees (Stone, Laitusis, & Cook, 2016) to better align with the CCSS.

The innovative item format is enhanced by technology in certain ways for the purpose of a given test. SBAC has developed two types of items which capitalize on technology: technology-enabled items and TE items. The differences between the two item types are elaborated in the consortium's item design training modules for ELA/Literacy and math (SBAC, 2016b). Technology-enabled items use digital media (audio, video, and/or animation) as the item stimulus but only require students to interact as is commonly done with SR or PA items. Students only select one best answer from a list of options provided in an SR item or construct short/extended responses to answer a PA task. For ELA assessments, most technology-enabled items will be part of PAs that use non-text stimuli and part of items for Claim 3: listening and speaking (see four major claims for SBAC in assessments of the CCSS for ELA/Literacy (SBAC, 2015)). On the other hand, TE items are computer delivered items that may include digital media as stimulus and require students to perform specialized interactions to produce their responses (see also Jodoin, 2003; Lorie, 2014; Wan & Henley, 2012). Students' responses to TE items are beyond those they normally perform in SR and PA items. In other words,

TE items allow the manipulation of information in ways that are not possible with the traditional item formats. Like SR items, TE items have defined responses that can be scored in an automated manner. Also, the students' complex interactions are intended to replicate the fidelity, authenticity, and directness of PAs (Downing, 2006b; Kane, Crooks, & Cohen, 1999; Lane & Stone, 2006; Shepard & Bleim, 1995). As a result, TE item formats are more difficult and demanding (Bukhari et al., 2016; Jodoin, 2003; Lorie, 2014; Parshall, Harmes, Davey, & Pashley, 2010; Sireci & Zenisky, 2006; Zenisky & Sireci, 2001, 2002) compared to the traditional SR and PA item formats, while still preserving the benefits of both items. Such potentials are deemed imperative and efficient in assessing students' readiness and predicting successful achievement in real world situations such as in college and the job market. Table 1 summarizes some of the interactions and the resulting item formats, the names of which are based on the mode of interactions required. Appendix A to D illustrate different item formats from different assessment programs.

Table 1. Sample of Technology-Enhanced (TE) Item Formats based on Examinees' Interactions

	Interaction	Formats
1	Examinees answer two selected response items. To answer the second selected response item, examinees show evidence from reading text that supports the answer they provided to the first selected response item[1]	Evidence-Based Selected Response (EBSR)
2	Examinees drag and drop objects to targets	Drag & Drop (Select-and-Order)
3	Examinees select multiple answer options	Multiple Correct Responses (Complex Selected Responses)
4	Examinee sequence events/element/info	Reordering (Create-a-Tree)
5	Examinees insert/drag and drop text	Text/ Equation-and-Expression Entry
6	Student places a mark on a graphic indicating a specified location	Hot Spot
7	Student select text within item stem or passage	Hot Text (Select-Text)
8	Student matches or classifies information/elements into specific theme/groups	Matching
9	Student is provided with the tools to create/modify a graph (e.g., a line graph, bar graph, line/curve plotter, or circle graph)	Graphing

[1] Different automated scoring procedures of EBSR items qualify EBSR as a TE item format.

From the assessment perspective, Scalise (2012, 2009) and Scalise and Gifford (2006) introduce a taxonomy or categorization of 28 innovative item types useful in CBT. The taxonomy describes "intermediate constraint (IC)" items in which items are organized by the degree of constraint and complexity placed on the test takers' options for answering or interacting with the assessment item or task. This degree of constraint and complexity is determined based on both horizontal and vertical continua of the taxonomy (Figures 2 & 3). On the horizontal plane, items are classified as fully constrained response (e.g., conventional SR item) to fully constructed response (CR) (e.g., essay). On the vertical plane, items range from the least complex (e.g., True/False) to the most complex (e.g., discussion/interview) responses. Figures 2 and 3 illustrate the exact same IC taxonomy. While Figure 2 (Scalise & Gifford, 2006) uses texts to describe the items and their corresponding references, Figure 3 (Scalise, 2009) attempts to provide the examples for most of the item formats in graphical forms and describe the details of the items in an interactive manner (see the link from the source provided).

Figure 2. Taxonomy of Item Types based on Level of Constraint

Most Constrained → **Least Constrained**

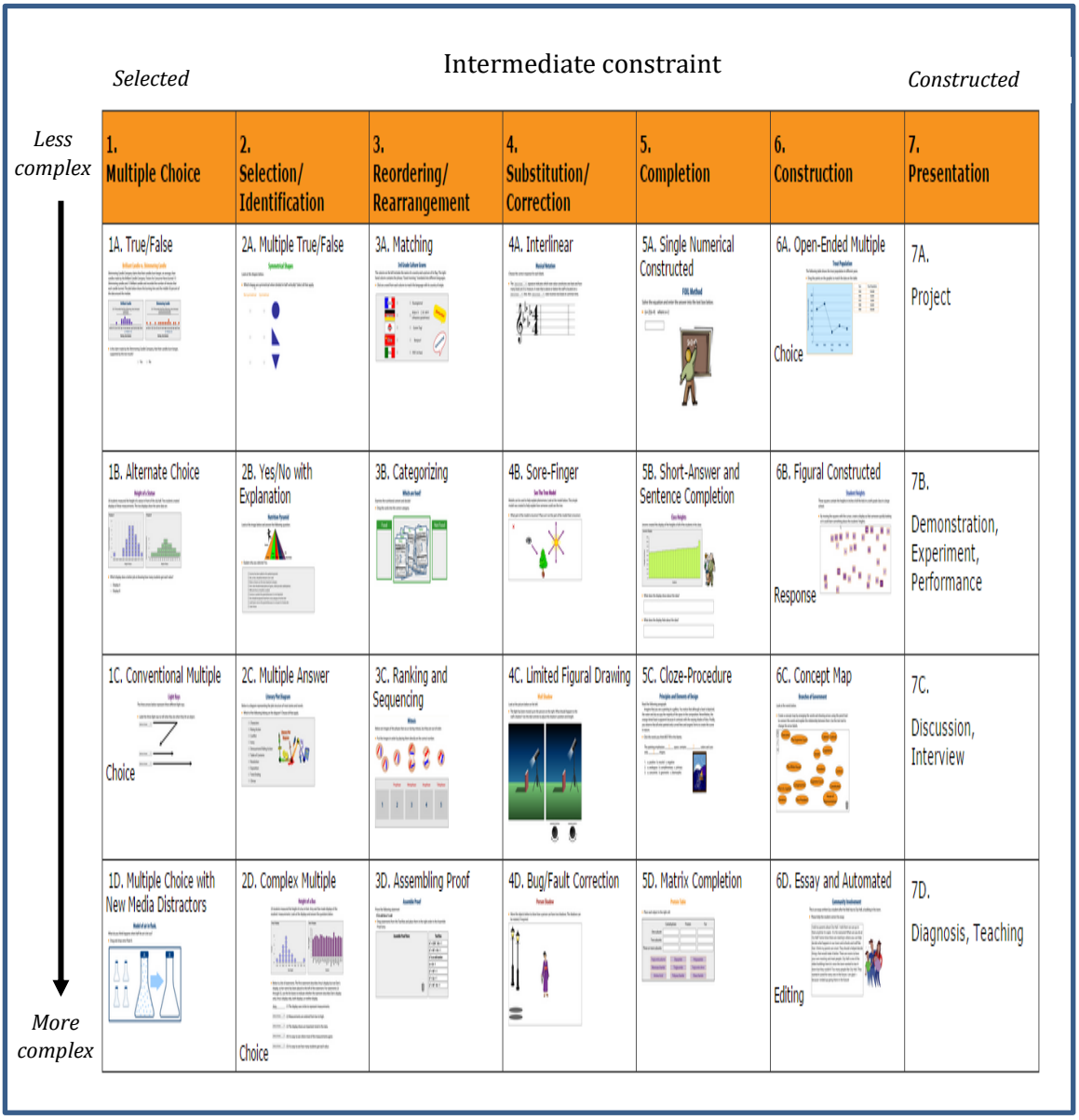
Fully Selected — *Intermediate Constraint Item Types* — *Fully Constructed*

Less Complex ↓ **More Complex**

1. Multiple Choice	2. Selection/ Identification	3. Reordering/ Re-arrangement	4. Substitution/ Correction	5. Completion	6. Construction	7. Presentation/ Portfolio
1A. True/False (Haladyna 1994c, p.54)	2A. Multiple True/False (Haladyna 1994c, p.58)	3A. Matching (Osterlind, 1998, p. 234; Haladyna, 1994c, p.50)	4A. Interlinear (Haladyna, 1994c, p.65)	5A. Single Numerical Constructed (Parshall et al., 2002, p.87)	6A. Open-Ended Multiple Choice (Haladyna, 1994c, p.49)	7A. Project (Bennett, 1993, p.4)
1B. Alternate Choice (Haladyna, 1994, p.53)	2B. Yes/No With Explanation (McDonald, 2002, p.110)	3B. Categorizing (Bennett 1993, p.44)	4B. Sore-Finger (Haladyna, 1994c, p.67)	5B. Short-Answer &Sentence Completion (Osterlind 1998, p.237)	6B. Figural Constructed Response (Parshall et al., 2002, p.87)	7B. Demonstration Experiment Performance (Bennett 1993, p.45)
1C. Conventional or Standard Multiple Choice (Haladyna, 1994c, p.47)	2C. Multiple Answer (Parshall et al., 2002, p.2; Haladyna, 1994c, p.60)	3C. Ranking Sequencing (Parshall et al., 2002, p.2)	4C. Limited Figural Drawing (Bennett, 1993, p.44)	5C. Chaze-Procedure (Osterlind, 1998, p.242)	6C. Concept Map (Shavelson, R. J., 2001; Chang & Baker,1997)	7C. Discussion, Interview (Bennett, 1993, p.4)
1D. Multiple Choice with New Media Distractors (Parshall et al., 2002, p.87)	2D. Complex Multiple Choice (Haladyna 1994c, p.57)	3D. Assembling Proof (Bennett, 1993, p.44)	4D. Bug/Fault Correction (Bennett, 1993, p.44)	5D. Matrix Completion (Embretson, 2002, p.225)	6D. Essay (Page et al., 1995, pp.561-565) & Automated Editing (Berland et al., 2001, pp.1-64)	7D. Diagnosis, Teaching (Bennett, 1993, p.4)

Reproduced from Scalise & Gifford (2006, p. 9)

Figure 3. The Intermediate Constraint (IC) Taxonomy for E-Learning Assessment Questions & Tasks



Source: Scalise (2009) http://pages.uoregon.edu/kscalise/taxonomy/taxonomy.html

Standards 12.3 and12.6 from Chapter 12: Educational Testing and Assessment of the *Standards* (AERA, et al., 2014) mandate careful test designs and development, as well as comprehensive documentation of supporting evidence on the feasibility of CBT (see Popp, Tuzinski, & Fetzer, 2016; Zenisky & Sireci, 2001) to gather information about the construct, to avoid construct-irrelevant variance (CIV), and to uphold accessibility for all examinees. CIV is one of the major threats to a fair and valid interpretation of test scores (AERA, et al., 2014; Haladyna & Downing, 2004; ITC, 2005a; Messick, 1989, 1994). Construct-irrelevance refers to the degree to which the measurement of examinees' characteristics is affected by factors irrelevant to the construct being measured. Examples of CIV that may arise with the implementation of computerized testing (Haladyna & Downing, 2004; Huff & Sireci, 2001; Zenisky & Sireci, 2006) are: test anxiety; test- "wiseness" and guessing related to SR items; test formats; and examinees' familiarity with technology that may be associated with socio-economic status (Chen, 2010; Taylor et al., 1999). Although the implementation of computer-based tests is promising, there is limited research on the possibility that such tests might introduce CIV (Haladyna & Downing, 2004, Huff & Sireci, 2001; Lakin, 2014).

Introducing new or unfamiliar computerized item formats to examinees creates particular challenges for test developers because examinees need to quickly and accurately understand what the test items require (Haladyna & Downing, 2004) as well as to understand the differences that may exist across formats (Pearson Educational Measurement, 2005). The critical challenge is how best to introduce a

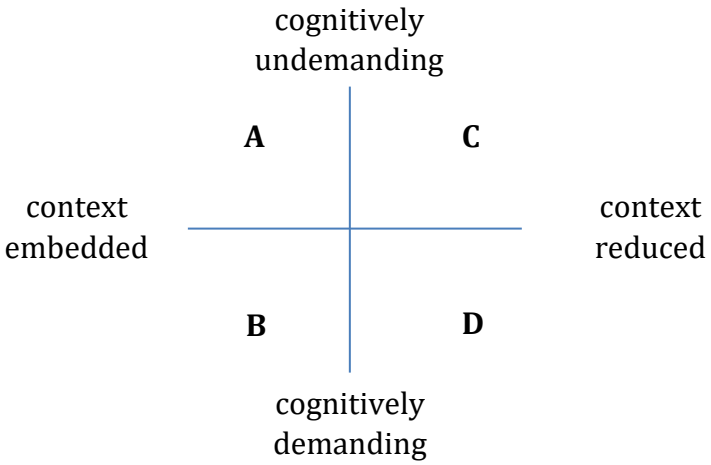
task so that all examinees are able to respond to the format as intended by the test developers. However, research to evaluate the adverse impact of the use of technology and most emerging TE items (Zenisky & Sireci, 2002) on test scores for different subgroups of examinee populations (Rabinowitz & Brandt, 2001; Sireci & Zenisky, 2006) remains incomplete.

Academic Language Proficiency

The concept of academic language (also referred to as academic English and more recently as English language proficiency (ELP)) has developed substantially since Cummins (1979, 1981, & 1994) introduced the distinction between basic interpersonal communication skills (BICS) and cognitive/academic language proficiency (CALP). Figure 4 illustrates Cummins' BICS and CALP framework, which is also known as a quadrant framework. It consists of two intersecting continua related to context and cognitive demands. On the horizontal level, context is developed as a continuum from context-embedded language (often associated with face-to-face interaction wherein facial expression, gestures, and negotiation of meaning provide context) to context-reduced language (usually written language with no physical elements of context thus successful interpretation of the message depends heavily on knowledge of the language itself). On the vertical level, the continuum extends from cognitively undemanding language (conversation on informal social topics) to cognitively demanding language (oral and written communication on the more abstract topics of academic subjects).

> Thus, conversational abilities (quadrant A) often develop relatively quickly among language minority students because these forms of communication are supported by interpersonal and contextual cues and make relatively few cognitive demands on the individual. Mastery of the academic functions of language (quadrant D), on the other hand, is a more formidable task because such uses require high levels of cognitive involvement and are only minimally supported by contextual or interpersonal cues. (Cummins, 1994, p. 11)

Figure 4. Cummins' (1994) Four-Quadrant Framework

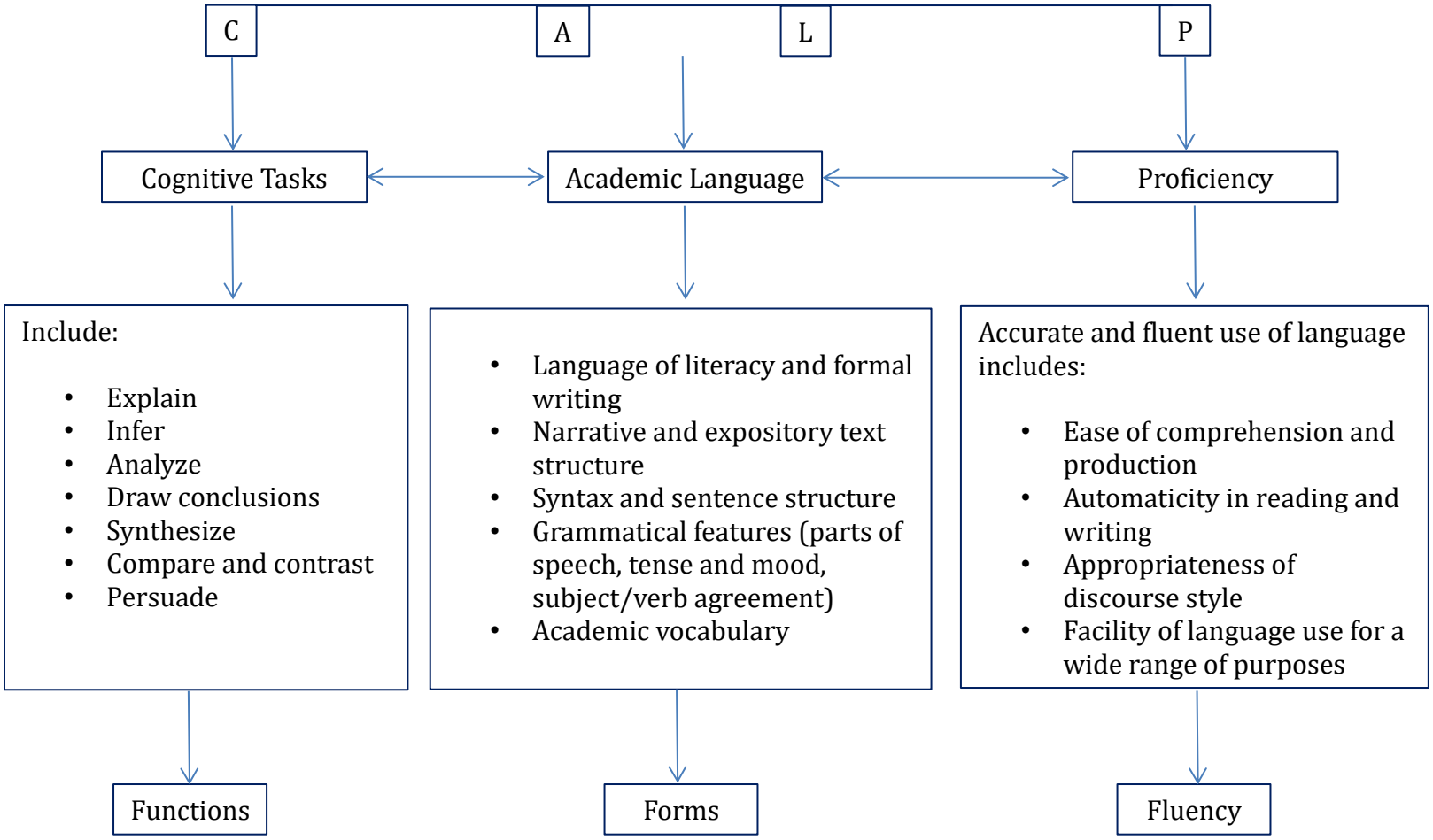


Using the BICS and CALP terms, Cummins proposes that immigrant students from non-English speaking backgrounds can more quickly (i.e., about two years) gain fluency in language used in situations outside formal learning contexts (such as BICS) than in the language needed to perform more cognitively demanding and abstract tasks in academic contexts such as CALP (i.e., about five to seven years), resulting in a lower academic achievement (Chiappe, Siegel, & Wade-Woolley, 2002; Hakuta, Butler, & Witt, 2000; Linquanti & George, 2007). The *Standards* (AERA et al., 2014) further reminds us that "[n]on-native English speakers who give the impression of being fluent in conversational English may be slower or not

completely competent in taking tests that require English comprehension and literacy skills" (p. 55).

After the authorization of the NCLB act, Dutro and Moran (2003) expanded Cummin's CALP concept, as shown in Figure 5, to include functions (e.g., explain, infer, analyze), forms (e.g., text structure, grammar, and vocabulary), and fluency (e.g., automaticity and appropriateness).

Figure 5. Dutro & Moran's (2003) Conceptual Model from CALP to Functions, Forms, & Fluency



The change in the academic language conceptualization continues in which the previously dichotomized BICS and CALP are now deemed inseparable, based on the situative/socio-cognitive perspective on academic language (Mislevy & Duran, 2014; Snow, 2008, 2010). Snow (2008, 2010) asserts that academic language and social (conversational) language can be situated at either end of a continuum without a clear boundary. This is supported by other researchers:

> … face-to-face, multimodal interaction in complex instruction involving all four modalities can support acquisition of complex analytic and academic-language skills, but it may do so in a face-to face mode relying on conversational, idiomatic forms of expression and communication that would not be acceptable as formal stand-alone written or expository language—despite representing critical and individually optimal experiences to help [non-native speaker] students develop the full range of resources that are the targets of learning. Ethnographic and discourse analytic studies of non–English-background students, for example, reveal that [such] students may use informal idiomatic peer-to-peer talk to analyze complex formal expository language in text and speech as part of academic assignments (Duran & Szymanski, 1995; Gutierrez, 2008). (Mislevy and Duran, 2014, p. 568)

Alternatively, still other researchers have categorized academic language into two types: general academic language and discipline-specific/technical language (e.g., Anstrom, DiCerbo, Butler, Katz, Millet, & Rivera, 2010; Romhild, Kenyon, & MacGregor, 2011; Wolf & Faulkner-Bond, 2016). General academic language refers to linguistic features that appear across multiple content areas, while discipline-specific/technical language appears only within specific content areas such as the language used in math and science subject domains.

With the new generation assessments that are based on the CCSS and the NGSS, students' competency in the English language of instruction is deeply implicitly assumed. A common theme that has emerged in the literature on the English language and literacy skills contained in the standards is that the language demands of various tasks instigated in the standards become greater as the rigor of performance expectations in the standards is raised through more challenging items, tasks, and texts (Abedi & Linquanti, 2012; Bailey & Wolf, 2012; Bunch, Kibler, & Pimental, 2012; Fillmore & Fillmore, 2012; Lee, Quinn, & Valdes, 2013; Moschkovich, 2012; Turner & Danridge, 2014; Wolf, Wang, Blood, & Huang, 2014).

The role of English competence in ELA/Literacy is grounded in high level analytic practices (CCSSO, 2010a) that include, for instance, the ability to recognize and synthesize complex relationships among ideas presented in informative texts and the ability to present and analyze complete established arguments based on claims made from texts. Examples in math (CCSSO, 2010b) require the ability to recognize how the verbal statements of math problems map onto the language of mathematical expressions and their conceptual meanings. The assessment also seeks to understand how examinees linguistically and symbolically present the structure of mathematical proofs, derivations, and findings. Examples for the science subject area (NGSS Lead State, 2013) include assessing the examinees' ability to verbalize, compose, and comprehend written, visual, and dynamic explanations of scientific facts, models, and principles; to provide argumentation

based on evidence; and to communicate, analyze, and validate the logic of scientific investigations (see also Lee, Quinn, & Valdes, 2013).

Developing competence in the practices mentioned above requires all four academic language modalities (listening, reading, speaking, and writing) and their integration with thinking, comprehending, and communication processes. Essentially, it is not easy to understand students' intertwined subject domain ability and language ability (see the model for interaction of communicative competence components by Celce-Murcia, Dornyei, & Thurrell (1995); the communicative language ability (CLA) framework by Bachman (1990), Bachman & Palmer (1996, 2010); and, challenges in aligning language proficiency assessments to the CCSS by Bailey & Wolf (2012)). This is also true for the native speaker students (Abedi & Lord, 2001; Erickson, 2004). The challenges are even greater when a diverse population of English language learners (ELL) is to be included in assessment systems (e.g., Abedi, 2006; Mislevy & Duran, 2014; Turner et al., 2014; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006; Wolf, et al., 2014) as initially mandated by the NCLB (NCLB, 2001b: Public Law, 107-110, Title III, January, 2002) (see also Abedi & Gandara, 2006; Bunch, 2011; Chalhoub-Deville & Deville, 2008)

This is equally true with students with disability (SWD) groups (NCLB, 2001a: Public Law, 107-110, Title I, January, 2002). Even with early intervention, educational institutions historically have struggled to provide SWD with opportunities for academic success (Harris & Bamford, 2001; Mutua & Elhoweris, 2002; Traxler, 2000). Part of the struggle has been in literacy development

(Cawthon, 2007, 2011; Lollis & LaSasso, 2009; Mitchell, 2008; Shaftel, et. al, 2006), which is often delayed.

ELL students are non-native speakers of English who have limited English proficiency. They are one of the fastest growing subgroups of K-12 students in US classrooms (National Center for Education Statistics (NCES), May, 2016). With the implementation of CCSS and NGSS, this subgroup—also referred to as emergent bilingual (EB) to recognize their bilingualism (Garcia, Kleifgen, & Farki, 2008; Valdes, Menken, & Castro, 2015)—must access academic content in the curriculum and, at the same time, develop their English proficiency. Students' content knowledge in areas such as math, science, or history/social studies may not be truly represented if they cannot understand the vocabulary and linguistic structures used in the tests.

> Research literature suggests that ELLs may not possess language capabilities sufficient to demonstrate the content knowledge in areas such as math and science when assessed in English. Thus the level of impact of language factors on assessment of ELL students is greater in test items with higher level of language demand. (Abedi, 2006, p. 377)

Findings from several studies have indicated the impact of English language proficiency on assessments in which ELLs are generally disadvantaged and perform at lower levels than the non-ELL students in reading (Abedi, Leon, & Mirocha, 2003; Chiappe, 2002; Geva, Yaghoub-Zadeh, & Schuster, 2000), math (Abedi, et al., 2003; Abedi, Lord, & Hofstetter, 1998; Abedi et al., 1997; Martiniello, 2008, 2009; Sato, Rabinowitz, Gallagher, & Huang, 2010; Shaftel et al., 2006), and science (e.g., Abedi,

Lord, Kim-Boscardin, & Miyoshi, 2000; Abedi, et al., 2003). These findings suggest that unnecessary linguistic complexity may hinder ELL students' ability to express their knowledge of the construct being measured.

> The unnecessary linguistic complexity of test items may introduce a new dimension that may not be highly correlated with the content being assessed. It may also create a restriction of range problem by lowering achievement of outcomes for ELL students that itself may lower internal consistency of test performance. (Abedi, 2006, p. 382)

As mentioned previously, this variability of assessment outcomes due to unnecessary factors such as linguistic complexity is known as CIV. I will present detailed reviews of the concept of linguistic complexity as potential CIV and of relevant studies in Chapter Two.

Multidimensionality of the Intended Assessment Construct

In the new CCR assessments, subscores are reported based on assessment claims and strand levels (e.g., PARCC, 2016c; Ohio Department of Education (DOE), 2016; SBAC, 2016a), the CCSS anchor standards (e.g., North Carolina Department of Public Instruction (NCDPI), 2016; Ohio DOE, 2016), and the NGSS domains (e.g., Florida DOE, 2016).

For example, SBAC (2016a), in general, reports three subscores for the math test domain based on four assessment claims (in which the second and fourth claims are combined into one subscore): (1) concepts and procedures, (2) problem solving, (3) communicating reasoning, and (4) modeling and data analysis. PARCC (2016)

generally provides four subscores based on similar claims: (1) major content, (2) expressing mathematical reasoning, (3) additional and supporting content, and (4) modeling and application. Other US state departments of education (see NCDPI, 2016; Ohio DOE, 2016) report subscores using the anchor standards by CCSS based on grade levels. For the grade eight math test, five subscores based on five anchor standards are reported: (1) the number system, (2) expressions and equations, (3) functions, (4) geometry, and (5) statistics and probability.

Other noteworthy examples are taken from the ELA/Literacy test domain. SBAC (2016a) reports subscores based on four assessment claims (also referred to as strands in the CCSS): (1) reading, (2) writing, (3) speaking and listening, and (4) research and inquiry. PARCC (2016c) and several departments of education (see NCDPI, 2016; Ohio DOE, 2016) only report two out of four strands (reading and writing). The ELA reading strand provides three subscores: (1) literary text, (2) informational text, and (3) vocabulary. The ELA writing strand reports two subscores: (1) writing expression, and (2) knowledge and use of language conventions. In addition to reporting based on the assessment claims/strands, some state departments of education also include the anchor standards in each of the ELA/Literacy CCSS strand to report subscores. In the ELA reading strand, subscores are reported based on four anchor standards: (1) key ideas and details, (2) craft and structure, (3) integration of knowledge and ideas, and (4) range of reading and level of text complexity. The ELA writing strand also consists of four anchor standards: (1) text types and purpose, (2) production and distribution of writing, (3) research

to build and present knowledge, and (4) range of writing. For the speaking and listening strand, two anchor standards are often used as subscores: (1) comprehension and collaboration, and (2) presentation of knowledge and ideas. The ELA language strand includes three anchor standards that can be used as subscores: (1) conventions of standard English, (2) knowledge of language and, (3) vocabulary acquisition and use.

Last but not least, the NGSS disciplinary core ideas (DCI) for science and engineering discipline highlight four major subdomains: (1) physical science, (2) life science, (3) earth and space science, and (4) engineering. These subdomains are adopted and adapted according to students' grade levels (Florida DOE, 2016). For instance, the grade five science test domain reports four subscores based on four NGSS subdomains: (1) the nature of science, (2) earth and space science, (3) life science, and (4) physical science; the grade ten biology test (based on the life science NGSS subdomain) reports three subscores: (1) molecular and cellular biology; (2) classification, heredity, and evolution; and (3) organisms, populations, and ecosystems.

These subscores will determine whether students meet or exceed expectations (mastery/exemplary/proficient), approach expectations (satisfactory/approaching proficiency), or do not yet meet or partially meet expectations (below satisfactory/inadequate/not proficient), to move to the next grade and eventually to enter college and the job market.

Universal Design Principle

Universal design is a concept that originated in the field of architecture (Center of Universal Design, 1997), but was later expanded into "environmental initiatives, recreation, the arts, health care, and now education [(Center for Applied Special Technologies, CAST, 2017)]" (Thompson, Johnstone, & Thurlow, 2002, p.2, citation added). Universally designed assessments are designed and developed to allow participation of "the widest possible range of students" (p.2) to provide valid inferences about performance on grade-level standards for all students who participate in the assessment (Thompson, Thurlow, & Malouf, 2004) for the sake of fairness.

> There is a tremendous push to expand national and state testing, and at the same time to require that assessment systems include all students — including those with disabilities and those with limited English proficiency— many of whom have not been included in these systems in the past. Rather than having to retrofit existing assessments to include these students (through the use of large numbers of accommodations or a variety of alternative assessments), new assessments can be designed and developed from the beginning to allow participation of the widest possible range of students, in a way that results in valid inferences about performance for all students who participate in the assessment. (Thompson, Johnstone, & Thurlow, 2002, p. 2)

The seven critical elements of universal design for educational assessments (Thompson, Johnstone, & Thurlow, 2002; Thompson, Thurlow, & Malouf, 2004) are: (1) an inclusive assessment population; (2) a precisely defined construct; (3) accessible, non-biased items; (4) amendable to accommodations; (5) simple, clear,

and intuitive instruction; (6) maximum readability and comprehensibility; and (7) maximum legibility.

Given the legislative emphasis (Individuals with Disabilities Education (IDEA), 2004; NCLB, 2001a, 2001c) on the use of universally designed assessments, test publishers and developers are responding to calls from the industry to incorporate universal design principles in novel test designs to ensure fairness. A lack of well-defined test development specifications for universally designed tests has led to a range of conceptualizations of how to best support students with special needs in assessment systems. Ketterlin-Geller (2008) presents a model of assessment development integrating student characteristics with the conceptualization, design, and implementation of standardized achievement tests. She integrates the universal design principle with the special needs of students using the twelve specific steps in test design and development specified by Downing (2006a). This effort is later expanded by Stone et al. (2016) in the context of "accessibility of assessments through CBT, including assistive technologies that can be integrated into an accessible testing environment, and the adaptive testing mode that allows for tailoring test content to individuals" (p. 220). Again, the principle of universal design for computerized assessments is emphasized.

The CAST (2017) has trademarked their principles for universal design for learning, focusing primarily on three principles: (1) multiple means of representation; (2) multiple means of action and expression; and (3) multiple means of engagement. Fortunately, the concept of TE item formats is tailored closely

to all three principles of CAST (2017). TE item formats represent different ways to engage and enable students to demonstrate what they know and can do based on their own capacity and learning style. Again, one challenge for this testing approach is related to the comparability of the difficulty level across different TE item formats and response modes. Moreover, different item formats, response modes, and computerized features of assessment and accommodations (e.g., linguistic modification, customized English glosses and dictionary, language translator) will tend to result in new, extraneous constructs or dimensions for different student populations (Abedi, 2006; Chapelle & Douglas, 2006; Popp, et. al, 2016; Zenisky & Sireci, 2001).

Description of Problem

The new educational assessments in general are apparently more demanding and challenging than students are currently prepared to face (Bukhari et al., 2016; Smarter Balanced News, May/June 2014; Dessoff, 2012; Wan & Henley, 2012). This is especially true when more critical thinking and problem solving questions with a high level language demand are presented through the incorporation of various item formats. The academic language demands in the assessments have also increased through the addition of more challenging items, tasks, and texts, instigated by the standards. As a result, the formative information retrieved from the test scores is twice as important as the traditional assessments.

Two types of factors may contribute to the test scores: (1) factors or dimensions that are intended, relevant, and of primary interest to the construct or test domain; and, (2) factors or dimensions that are "nuisance" or irrelevant to the construct, causing residual covariance that may impede the assessment of psychometric characteristics. Different item formats such as new TE items, computerized PA, and other item formats as well as the linguistic complexity of the items' stems and stimuli, in most cases, may improperly influence the response data and the psychometric characteristics of the test. Conscientious distinctions between the primary dimensionality (the intended test construct) versus the nuisance dimensionality that might contribute method variance resulting from the testing features and were not meant to be measured by the test should be made to ensure best testing practices and the validity of test scores (i.e., evidence based on internal structure (*Standards*, AERA et al., 1999, 2014)) and to support their interpretations given the intended uses of the test (AERA et al., 1999, 2014; ITC, 2013a; Kane, 2013).

In the context of the CCR assessments instigated by the CCSS and NGSS, test scores are used to determine the readiness of individual students to perform in college and the workplace as well as to make decisions about schools or states with the implementation of test-based accountability systems. Describing the CIV in the context of item or response formats and linguistic complexity is imperative in the effort especially to uphold fairness in testing (AERA et al., 2014; IDEA, 2004; ITC 2013a; NCLB, 2001a: Public Law, 107-110, Title I, January, 2002; NCLB, 2001b:

Public Law, 107-110, Title III, January, 2002) and to embrace the universal design principle in educational assessments (CAST, 2017; Ketterlin-Geller, 2008; Stone et al., 2016; Thompson et al., 2002; Thompson et al., 2004).

I have previously explicated the concept of validity, the conceptualization and characteristics of the next generation assessments, and the discussion of how the features of such assessments might be restraining the performance of certain examinees from various subpopulations and students deemed at-risk and disadvantaged, who previously were not included in the testing system. The purpose of such an elaborate introduction is very much needed and is a critical first step so that the reader may acquire an initial understanding and to provide some important context. Furthermore, I also describe the principle of universal design in general and specifically in terms of educational test development and design for best testing practices.

As a student of and a researcher in the educational measurement field with some interest and training background in innovative and language assessments, I will count it a privilege if I am able to gather and analyze large-scale, real testing data from the next generation assessments which include innovative features and which cater to all student populations (including the ELLs and the SWDs) since there are insufficient reported examinations of the effects of the relationship of different construct-irrelevant factors on psychometric constructs. Nevertheless, in a real world, such an intention might be difficult to accomplish.

Alternatively, simulation studies could be conducted to allow researchers to answer specific questions about data analysis, statistical power, and the best practices for obtaining accurate results in empirical research. Such studies also enable any number of experimental conditions that may not be readily observable in real testing situations to be tested and carefully controlled. Moreover, simulation enables researchers to replicate study conditions easily and consistently that would be very expensive when conducted with live subjects. Although a simulation of educational testing situations will never accurately feature the true complexity and inherent context of real data (Luecht & Ackerman, 2017) and therefore does not permit for conclusive conclusions, simulations are useful for framing general patterns and trends of a limited selection of phenomena of interest. I therefore prefer to attempt to frame my study using the context specific to my interests to help me create more realistic conditions and thus a better simulation—i.e., closer to a "simulation study-in context" (cf. Bachman & Palmer, 1996; Chalhoub-Deville 2003; Chalhoub-Deville & Deville, 2006; Luecht & Ackerman, 2017; Snow, 1994).

To date, researchers investigating item response theory (IRT)-based simulations in educational measurement have not generated simulated observed data which mirrors the complexity of real testing data due to two fundamental limitations (Luecht & Ackerman, 2017): (1) the difficulty of separating different types of errors from different sources, and (2) comparability issues across different psychometric models, estimators, and scaling choices. A simulation study of the various testing configurations of the new generation assessments and of the impact

of nuisance dimensions on residuals and residual covariance (an indication of local dependency) structures is needed to understand the consequences of these underlying unintended dimensions on the psychometric characteristics of test items and the scale scores (AERA et al., 1999; 2014; ITC, 2013b; Lissitz & Samuelson, 2007a, 2007b) as well as on the interpretations and uses of the scores (AERA et al., 1999, 2014; ITC 2013a; Kane 2013).

Purposes of Research

The primary purpose of this research is to explore the statistical complications encountered when potential nuisance dimensions exist explicitly in models in addition to the primary dimensions of interest in the context of the next generation K-12 assessments. Specifically, I first explore the effects that two calibration procedures (i.e., a unidimensional model and a confirmatory, compensatory multidimensional model) have on the structure of residuals of such complex assessments when nuisance dimensions are not explicitly modeled during the calibration processes and when tests differ in testing configurations. In other words, my first purpose is to examine whether unidimensional models could adequately recover the predominant construct of interest and to explore the consequences of analyzing multidimensional tests—with dimensions that vary in purpose and associations—using a unidimensional model. The two calibration models are a unidimensional item response theory (UIRT) model and a multidimensional item response theory (MIRT) model. Again, both models only

include the four primary dimensions of interest. Second, I also want to examine the residual covariance structures when the six test configurations vary. The residual covariance in this case indicates statistical dependencies due to unintended dimensionality.

To examine the residuals and residual covariance structures of the items in the context of next generation assessments, I will incorporate a new technique developed by Luecht and Ackerman (2017) that employs the expected response function (ERF) approach—which is based on the expected raw scores (ERS)—which can be used to compare the different components of residuals and errors in the test. More importantly, this approach is metric-neutral in that it allows for direct comparison between the unidimensional and multidimensional scales. I will elaborate this ERF-based approach (Luecht & Ackerman, 2017) in detail in Chapter Three.

I will conduct a simulation study in which I will generate item response data under a variety of realistic test configurations (e.g., sample sizes of examinees, correlations between primary dimensions, number of items per dimension, and discrimination levels of the primary dimensions) and include at least one nuisance dimension, termed as (1) item/response format or/and (2) linguistic complexity. When two nuisance dimensions are present, I will also vary their correlations with each other. Specifically, I will address the following research questions.

Research Questions

1. How much ERS-based residual covariance do different, more parsimonious IRT calibration models produce when the generated ("true") model represents a more complex reality with nuisance dimensions such as in the next generation, mixed-method, online assessments. Which calibration method performs best:
    a. When the nuisance dimension(s) is(are) present?
    b. When correlations between nuisance dimensions vary?
    c. When correlations between primary dimensions vary?
    d. When item discrimination levels on primary dimensions vary?
    e. When the number of item in each primary dimension varies?
    f. Over various sample sizes?
2. In what ways is the amount of modeled residual covariance impacted by:
    a. The presence of a nuisance dimension?
    b. The number of nuisance dimensions?
    c. The strength of correlations between nuisance dimensions?
    d. The strength of correlations between primary dimensions?
    e. Changes in discrimination ratios on the primary dimensions?
    f. The number of items in each primary dimension?
    g. Changes in the ratio of dichotomous items to polytomous items?
    h. Changes in sample size?

Organization of the Study

To answer the research questions, I will provide a review of the relevant literature in Chapter Two. I will begin Chapter Two with a discussion of modern IRT by describing the dichotomous and polytomous unidimensional IRT (UIRT) models, and their underlying assumptions, as well as item and test information. I will also provide description of both dichotomous and polytomous multidimensional IRT (MIRT) models, and their items and test statistics. I will then discuss one of the UIRT assumptions of local independence and the consequences of violating the assumption. I will then synthesize research on the dimensionality of mixed-format tests as well as the procedures to calibrate and score such tests. I will review potential CIV for the next generation assessments based on two sources: (1) different item and response formats that employ technology; and, (2) unnecessary linguistic complexity for different subgroups of students such as the ELL and the SWD. In the following section, I will synthesize studies of the relationships between academic English language and content performance for K-12 ELL students. Altogether, with this literature review, I aim to synthesize significant trends and identify potential room for research regarding issues on dimensionality and local dependency, especially in the context of K-12 next generation assessments. More importantly, I hope to better understand and create realistic conditions relevant to the next generation assessments, which I will later justify in Chapter Three.

I will describe the detailed methodologies that I will employ in this study in Chapter Three. In this chapter, I will outline the simulation design for the studies by

delineating the constant and conditions of the simulation and their corresponding rationale for selection. To answer the two research questions, I will present the description of the IRT calibration models and the scoring methods that I will employ along with the intended outcomes of the analysis. Finally, I will introduce the expected response function (ERF)-based residuals approach by Luecht and Ackerman (2017) which I will use to answer my two research questions.

A Note on Terminology

I use interchangeably throughout this document the terms `traits', `factors', 'constructs', 'domains', 'dimensions', 'proficiency', and 'ability'. *Estimation of latent trait/ability* I also refer to as scoring. *Estimation of item parameters* will also be described as calibration. I sometimes refer to *concurrent calibration* also as *simultaneous calibration*, in which dichotomous and polytomous items from SR, TE, and computerized PA formats are calibrated together in a given commercial software to produce one estimate of ability based on responses to those item/response formats for dichotomous and polytomous items.

The term *item/response format* refers primarily to the different TE items and computerized PAs regardless of whether they are dichotomously or polytomously scored. *Item type* specifically refers to dichotomous and polytomous items. The terms 'dichotomous' and 'polytomous' also refers to different scoring procedures.

CHAPTER II

LITERATURE REVIEW

In the context of educational and psychological assessment, measurement is defined as the systematic process by which numbers are assigned to individuals, objects, or events according to rules to represent their properties or characteristics (Bock & Jones, 1968; Lord & Novick, 1968; Stevens, 1951). The processes of scoring and scaling are critical at the operational stage of measurement. Thissen and Wainer (2001) defined test scoring as the process of "combining the coded outcomes on individual test items into a numerical summary of the evidence the test provides about the examinee's performance" (p. x).

> Scaling is the process of associating numbers or other ordered indicators with the performance of examinees on [a given] test. These numbers and ordered indicators are intended to reflect increasing levels of achievement or proficiency. The process of scaling produces a score scale, and the scores that are used to reflect examinee performance are referred to as scale scores. (Kolen, 2006, p. 155)

The scale score is a summary of the evidence contained in an examinee's responses to the test items related to the construct or a set of constructs being measured. The type of summary desired and the extent to which that summary can be generalized beyond the examinees' specific responses rely heavily on "the theoretical orientation of the test scorer" (Thissen & Wainer, 2001, p.1).

Thissen and Wainer (2001) classified the theoretical orientation into two schools of thought. One perspective is completely empirical. This perspective (known as the traditional test theory) views the scale score as a summary of responses to the items on the test, makes no further generalization of the responses, and is based on the concept of the true score (see Gulliksen, 1950, 1987; Lord & Novick, 1968). Researchers from the other perspective on scale score view the item responses as indicators of the examinee's level with respect to some underlying trait or traits. In this sense, it is appropriate to draw inferences from the observed responses to make an estimate of the examinee's level of the underlying trait or traits. This latter perspective is in agreement with a long tradition of psychological scaling (Binet & Simon, 1905; Thurstone, 1925) and has developed into modern IRT.

This chapter begins with a delineation of the concepts from modern IRT perspectives. My review of modern IRT is based on a discussion of the assumptions of the UIRT models and the description of the UIRT models for both dichotomous and polytomous test items. What follows next is a description of the MIRT models which are an extension of most of the UIRT models along with the items and test statistics associated with the multidimensional models. Following the discussion of the IRT models, I will specifically focus on the assumption of local independence of items that is critical in most measurement models, especially in the UIRT models. I will then synthesize the research conducted on the dimensionality of mixed-format tests and review studies that described the procedures to calibrate and score such tests. My review of potential CIV for the next generation assessments is based on

two potential CIV sources deemed related to such assessments: (1) different item and response formats that employ technology; and, (2) unnecessary linguistic complexity for different subgroups of students such as the ELL and the SWD. In the following section, I will synthesize studies on the relationships between academic English language and content performance for K-12 ELL students. Chapter Two ends with an overall summary of the previous sections.

Item Response Theory

Item response theory (IRT) is perhaps the most important technical innovations in educational and psychological measurement for almost a century (Thurstone, 1925; Lord & Novick, 1968). It has been modernized ever since (e.g., Birnbaum, 1968; Finney, 1952; Haley, 1952; Rasch, 1960) and is widely used in educational and psychological measurement research. IRT provides an advanced statistical framework for modeling how examinees respond to test items in isolation or in components (Thissen & Wainer, 2001; Yen & Fitzpatrick, 2006). The family of statistical models in IRT provide powerful ways to model individual examinee response patterns by specifying how the underlying trait or traits of examinee(s) interact with the item characteristics (i.e., item difficulty, item discrimination) to produce an expected probability of the response pattern (de Ayala, 2009; Embretson & Reise, 2000; Hambleton & Swaminathan, 2010; Reckase, 2009; Thissen & Wainer, 2001; Yen & Fitzpatrick, 2006). Thus, a major purpose of IRT is to separate the characteristics of the sampled population of examinees and the

characteristics of item parameters from a given test (i.e., observed response data) in order to understand and study the examinees and items separately. This parameter separation often requires advanced numerical analysis techniques for effective estimation (i.e., parameters estimation methods) (Baker & Kim, 2004; Bock, 1983; Bock & Aitkin, 1981; Bock & Mislevy, 1982; Cai, 2010b, 2010c; Kim & Bolt, 2007; Lord, 1980; Patz & Junker, 1999; Samejima, 1980; Warm, 1989). Using a selected parameter estimation method, the test items are placed on a common measurement scale as the examinees' latent ability (i.e., item calibration), enabling the interpretations of both item and test characteristics to specific points or regions of the underlying proficiency scale (Lord & Novick, p. 86, as cited in Thissen & Orlando, 2001; Thurstone, 1925, p. 437, as cited in Thissen & Orlando, 2001). As such, IRT offers a flexible model-based approach that is often deemed more meaningful than the traditional test theory (Embretson and Reise, 2000, pp. 14-39 provided a detailed comparison of the traditional test theory and IRT in the context of the old and the new measurement rules.). "When used appropriately, IRT can increase the efficiency of the testing process, enhance the information provided by that process, and make detailed predictions about unobserved testing situations" (Yen & Fitzpatrick, 2006, p. 111). Essentially, IRT provides numerous desirable properties for quantifying item properties, evaluating item quality, understanding measurement precision, developing assessments, and evaluating the properties of scores generated by assessments (de Ayala, 2009; Embretson & Reise, 2000; Hambleton & Swaminathan, 2010; Reckase, 2009).

The simplest and mostly used IRT models are the models that specify a single/unidimensional latent ability. UIRT models are easy to understand and employ parameter estimation methods that are, to some extent, computationally friendly. On the other hand, many educational (e.g., College Board, 2015; CCSSO, 2010a, 2010b; NGSS Lead States, 2013) and psychological (e.g., Criteria Corp., 2017; ETS, 2016) assessments are multidimensional in nature. With more recent advances in IRT research and computational power of personal computers for parameter estimation, the development and use of MIRT models is becoming more rapid and common.

Unidimensional Item Response Theory Models

UIRT encompasses a set of models that specify the interactions of examinees and items (i.e., item response theories). These models posit that only one hypothetical construct primarily influences the examinee(s) performance on test items. UIRT models use mathematical expressions, each containing a single parameter (i.e., unidimensional) describing the characteristics of the examinee(s). The basic representation of a UIRT model is:

$$P(U = u \mid \theta) = f(\theta, \xi) \tag{1}$$

In equation (1), θ represents the unidimensional parameter that describes the characteristics of the person, ξ represents a vector of parameters that describe the characteristics of the test items, U represents the score on the test item for a particular examinee, u denotes a possible value for the score, and f denotes a function that describes the relationship between the parameters and the probability of the response, $P(U = u)$.

Assumptions of UIRT models

UIRT models have several assumptions. The first and strong assumption of the models is the assumption of a single person parameter, θ, for a given UIRT model. This is commonly known as assumption of unidimensionality. The assumption indicates that despite the complexities of the data (e.g., other cognitive ability, personality, level of motivation, test-taking strategy factors, ability to work quickly, familiarity with the use of answering sheets, tools, and item formats), "only one ability or trait is necessary to 'explain' or 'account' for examinee test performance" (Hambleton & Swaminathan, 2010, p. 16). For example, it can be assumed that scores on a math test are primarily influenced only by the students' latent math ability that is intended to be measured. A failure to completely define the latent ability space in the case of UIRT will lead to violation of the assumption of unidimensionality.

The second assumption is the functional form of item characteristic curves (ICCs) assumption (de Ayala, 2009; Embretson & Reise, 2000; Hambleton &

Swaminathan, 2010). "This assumption states that the data follow the function specified by the model" (de Ayala, 2009, p. 21). In other words, it is assumed that the chosen UIRT model fits the data. An ICC is a mathematical function that models how changes in ability level relate to changes in the probability of a specified response (Embretson & Reise, 2000; Hambleton & Swaminathan, 2010). "It is the nonlinear regression function of item score on the trait or ability measured by the test" (Hambleton & Swaminathan, 2010, p.25). For a correct response on a dichotomous item, the ICC regresses the probability of item success on trait level. For a polytomous item, the ICC regresses the probability of responses in each category on trait level. According to Yen (1993), if an appropriate model is used, it typically can accurately describe the observed ICCs regardless of whether or not item scores are locally dependent (i.e., a concept of local item dependence (LID): this will be discussed in detail in later). She later found, however, when LID was extreme in a PA task of math subject, there was a great effect of LID on the accuracy of the ICCs predictions (Yen, 1993). In the measurement literature, ICC is also referred as item characteristic function, item response function (e.g., Penfield, 2014), item category response function (e.g., Muraki, 1992, 1993; Samejima, 1969), operating characteristic curve (Samejima, 1969), and trace line (e.g., Thissen & Steinberg, 1986; Thissen & Orlando, 2001; Yen, 1993).

The third assumption of UIRT models is the monotonicity assumption (Reckase, 2009). Most UIRT models assume that the probability of selecting or producing the correct response to a test item increases as the examinees ability, θ ,

increases. This assumption is closely related to the aforementioned second assumption on the functional form of ICC and the assumption of local independence (LInd) (Rosenbaum, 1984).

One implicit assumption of most UIRT models that is seldom stated is the speededness assumption (Hambleton & Swamintahan, 2010). It is assumed that the tests to which the models are fit are not administered under speeded conditions in which examinees fail to answer test items because of limited ability and not because they failed to reach the test items. Oshima (1994) found that test speededness had a substantial effect on the item parameter estimates and a minimal effect on the estimated ability parameters. Assumption of speededness is often not explicitly mentioned as a separate assumption of UIRT because it is often subsumed (Yen, 1993) under the fifth yet one of the most critical assumptions for UIRT models: the local independence (LInd) assumption.

The assumption of LInd (also referred to as the assumption of conditional independence or conditional non-association) is one facet of a model-data fit (see Ames & Penfield, 2015) investigation. In the UIRT models, LInd is an important assumption to indicate that the success on one item on a given test is not influenced by the success on another item from the same test. Yen (1993) argued, if the only goal of a given assessment is a one-time measurement of a latent trait or construct using a set of items, then the LInd assumption "is an unimportant psychometric nicety that can be ignored" (p. 190). Nevertheless, independent items in educational assessments are essential to provide scores that can distinguish a student's relative

achievement and ability on educational outcomes. Such independent items are desirable to produce scores that are sufficiently reliable and that can be validated to support their interpretations for intended uses of tests (AERA et al., 1999, 2014). The LInd assumption is deemed to be "equivalent" or "directly linked" to the assumption of unidimensionality (McDonald, 1981 1982, as cited in Hambleton & Swaminathan, 2010, p. 25 & Yen & Fitzpatrick, 2006, p. 123) although "they are unequivocally distinct mathematical entities" (Ip, 2010, p. 396). Edwards & Cai (2011) argued that LInd relates to the correct specification of the amount of common factors rather than dimensionality issues. If there were two common factors, but they were modeled correctly, the item responses would be locally independent. A common example of such a case is when a set of items that share common stimulus or that are context-dependent, thus are deemed locally dependent (e.g., testlet/item bundle), is treated as an independent unit from another set of items that shared similar context or stimulus (e.g., Haladyna, 1992; Rosenbaum, 1988; Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wainer & Lewis, 1990). Given that the concept of LInd is closely related to the purposes of the current study, I will provide a detailed discussion of the LInd assumption and the implication of its violations after I discuss the unidimensional and multidimensional models of IRT.

UIRT Logistic Model for Dichotomous Items

The 3PL model (Birnbaum, 1968) is a general IRT model that is appropriate for dichotomously scored items. Dichotomous items are test items with two score categories of correct (score of 1: $u = 1$) or incorrect (score of 0: $u = 0$), such as the SR item formats. It is characterized by the following mathematical function:

$$P(U_i = 1 | \theta_j, a_i, b_i, c_i) = P_{ij}(\theta)_{3\text{PL}} = c_i + (1 - c_i) \frac{\exp\left[Da_i(\theta_j - b_i)\right]}{1 + \exp\left[Da_i(\theta_j - b_i)\right]} \quad (2)$$

where $P_{ij}(\theta)$ is the probability of getting item *i* correct for an examinee j having proficiency scores denoted as θ. D is a scaling constant often set to 1.702 to approximate a cumulative normal probability function, where θ is distributed with a mean of zero and variance of one. The item parameters, a_i, b_i, and c_i, determine the shape of a particular response function across the θ scale. The a parameter determines the steepness of the ICC slope. It reflects the item discrimination and is equivalent to the biserial/point-biserial correlation (Reckase, 2009; Urry, 1974) or item-total correlation (Penfield, 2010) index in the traditional test theory. The b parameter represents the location of the IRT parameter, which is also known as the item difficulty index. It is similar to the mean of scores (Penfield, 2010) or proportion of correct scores on a given item (Reckase, 2009; Urry, 1974) in the traditional test theory. The c parameter denotes the lower asymptote of the response function and is associated with noise in the response patterns at the

lowest proficiency levels (sometimes known as the pseudo guessing parameter). It does not have a direct counterpart in the traditional test theory. The two-parameter logistic (2PL) (Birnbaum, 1968) and the one-parameter logistic (1PL)/Rasch (Rasch, 1960) models are also shown in equations (3) and (4) below accordingly:

$$P(U_i = 1 \mid \theta_j, a_i, b_i) = P_{ij}(\theta)_{2\text{PL}} = \frac{\exp\left[Da_i(\theta_j - b_i)\right]}{1 + \exp\left[Da_i(\theta_j - b_i)\right]} \tag{3}$$

$$P(U_i = 1 \mid \theta_j, b_i) = P_{ij}(\theta)_{1\text{PL}} = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)} \tag{4}$$

UIRT Logistic Model for Polytomous Items

In addition to the UIRT models for dichotomous items, models for polytomous items are also available (see also Embretson & Reise, 2000; Mellenbergh, 1995; Nering & Ostini, 2010; Penfield, 2014; Thissen & Steinberg, 1986). Polytomous items are test items with more than two score categories with possible score values of, for example, $u = (0, 1, 2, \ldots, M)$. Thus, there are a total of $M + 1$ score categories. In this section I will describe four polytomous logistic models (Andrich, 1978a, 1978b; Masters, 1982; Muraki, 1992; Samejima, 1969) using a common notation. These models are extensions of the dichotomous models particularly from either the 1PL/Rasch or the 2PL models and are widely used in the operational testing environment.

The first model is the rating scale model (RSM) introduced by Andrich (1978a, 1978b). A common example that is used to describe the use of the RSM

model is the five-level Likert item (Likert, 1932) in which all items share a common set of level descriptors for the response categories: strongly disagree, disagree, neutral, agree, and strongly agree. The RSM is chosen to model the data for Likert items when it is assumed that each level is equidistant with each other and that all Likert items in a given scale are assumed to have the same underlying equidistant levels. For example, the distance of disagree level and neutral level is equal to the distance of neutral level to agree level; and that this equal "affective intensity" (Penfield, 2014, p.43) is constant across all Likert items. The mathematical function of the RSM can be written as:

$$P(U_{ikj} = u \mid \theta_j, b_i, d_k) = P_{ik}(\theta_j)_{\text{RSM}} = \frac{\exp \sum_{c=0}^{k} (\theta_j - b_i - d_c)}{\sum_{k=0}^{m} \exp \sum_{c=0}^{k} (\theta_j - b_i - d_c)} \quad (5)$$

where $P_{ijk}(\theta)$ is the probability of responding in category k ($k = 0, 1, \ldots, m$) of item i for examinee j, u is the score on item i, b_i denotes the item location, and d_k is the step threshold parameter.

Masters (1982) developed the partial credit model (PCM) that is deemed as an extension of Andrich's RSM (1978a, 1978b). PCM is often used to analyze polytomous test items with multiple steps where it is important to assign partial credit/score for completing steps such as in PAs or even for the Likert items with level descriptors that are assumed to vary. Instead of having the equidistant descriptor levels within an individual item and across all items, PCM models the

variation of the level descriptors or item steps within a particular item or across the items. The mathematical expression for PCM given by equation (6) below is similar to the mathematical function of RSM in equation (5) except that the threshold parameter in equation (5) is now become the item step threshold that can vary across items (d_{ik} in equation (6)).

$$P(U_{ikj} = u \mid \theta_j, b_i, d_{ik}) = P_{ik}(\theta_j)_{\text{PCM}} = \frac{\exp \sum_{c=0}^{k} (\theta_j - b_i - d_{ic})}{\sum_{k=0}^{m} \exp \sum_{c=0}^{k} (\theta_j - b_i - d_{ic})} \quad (6)$$

Both RSM (Andrich, 1978a, 1978b) and PCM (Masters, 1982) are also known as "Rasch polytomous models" (Muraki, 1992, p.160) given that they are formulated using the 1PL/Rasch model in equation (4). Another model considered as an extension of the 1PL/Rasch (Rash, 1960), hence included in the Rasch model family, is the generalized partial credit model (GPCM) by Muraki (1992). GPCM is a generalization of PCM (Masters, 1982) in which it allows item discrimination parameter (a_i) to differ across items within a given score scale. Therefore, GPCM can also be employed to analyze polytomous test items such as the ones analyzed by PCM. However, the items are assumed to have different discrimination levels. The mathematical function of GPCM is:

$$P(U_{ikj} = u \mid \theta_j, a_i, b_i, d_{ik}) = P_{ik}(\theta_j)_{\text{GPCM}} = \frac{\exp \sum_{c=0}^{k} Da_i(\theta_j - b_i - d_{ic})}{\sum_{k=0}^{m} \exp \sum_{c=0}^{k} Da_i(\theta_j - b_i - d_{ic})} \quad (7)$$

where D is a scaling constant often set to 1.702.

With the ability to model different discrimination of items, GPCM is indeed a polytomous version of the 2PL (Birnbaum, 1968) model. The equation (7) is formulated from the 2PL's mathematical function in equation (3). The 2PL model has also been explicitly extended to other polytomous UIRT models such as the graded response model (GRM) of Samejima (1969). However, the model does not belong in the Rasch-type model family. GRM is formulated for test items that have somewhat different requirements than the polytomous Rasch models that have been discussed (i.e., RSM, PCM, and GPCM). These models consider the items to have a number of independent parts and the score determines how many parts were successfully answered or accomplished. Thus, an issue related to different ordering of the item step threshold parameters may occur when using RSM, PCM, and GPCM (e.g., Reckase, 2009, pp. 33-35).

In contrast, GRM considers a test item to require a number of steps but the successful performance of one step requires the successful performance of the previous steps. If item step k is accomplished, then previous steps are also assumed to be accomplished. Here, the parameterization of GRM considers the lowest score on item i to be 0 and the highest score to be m. The probability of accomplishing k

or more steps of an item is represented by the 2PL model in equation (3). The probability of receiving a specific score, k, is the difference between the probability of responding to the task for k or more steps and the probability of responding to the task for $k+1$or more steps. If the probability of performing the task including step k at a particular level of θ is $P*(U_{ij}=k \mid \theta_j)$ then the probability that an examinee j will receive a score of k is

$$P(U_{ij}=k \mid \theta_j) = P*(U_{ij}=k \mid \theta_j) - P*(U_{ij}=k+1 \mid \theta_j) \quad (8)$$

where $P*(U_{ij}=0 \mid \theta_j)=1$, because performing the task for step 0 or more is a certainty for all examinees and $P*(U_{ij}=m+1 \mid \theta_j)=0$ because it is impossible to accomplish a task representing more than category m. The two latter probabilities are defined so that the probability of each score can be determined from equation (8). Samejima (1969) referred to the terms on the right side of equation (8) as the cumulative category response functions and the ones on the left side of the equation as the category response function. In the polytomous UIRT literature, several researchers also refer to GRM as a cumulative model (Mellenbergh, 1995; Penfield, 2014). To better illustrate GRM, consider a math item that requires derivations of mathematical proofs. To receive a full score, an examinee is required to show four steps ($k=0,1,2,3$) of derivations for his or her work. Thus, the probability that the examinee will receive a specific score at each k step is:

$$P_{i0}(\theta_j) = 1 - P_{i1}^{*}(\theta)$$
$$P_{i1}(\theta_j) = P_{i1}^{*}(\theta) - P_{i2}^{*}(\theta)$$
$$P_{i2}(\theta_j) = P_{i2}^{*}(\theta) - P_{i3}^{*}(\theta)$$
$$P_{i3}(\theta_j) = P_{i3}^{*}(\theta) - 0$$

Replacing the probability of accomplishing k or more steps of an item ($P_{ik}^{*}(\theta)$) with the 2PL model in equation (3) produces the following:

$$P_{i0}(\theta_j) = 1 - \frac{\exp\left[Da_i(\theta_j - b_{i1})\right]}{1 + \exp\left[Da_i(\theta_j - b_{i1})\right]}$$
$$P_{i1}(\theta_j) = \frac{\exp\left[Da_i(\theta_j - b_{i1})\right]}{1 + \exp\left[Da_i(\theta_j - b_{i1})\right]} - \frac{\exp\left[Da_i(\theta_j - b_{i2})\right]}{1 + \exp\left[Da_i(\theta_j - b_{i2})\right]}$$
$$P_{i2}(\theta_j) = \frac{\exp\left[Da_i(\theta_j - b_{i2})\right]}{1 + \exp\left[Da_i(\theta_j - b_{i2})\right]} - \frac{\exp\left[Da_i(\theta_j - b_{i3})\right]}{1 + \exp\left[Da_i(\theta_j - b_{i3})\right]}$$
$$P_{i3}(\theta_j) = \frac{\exp\left[Da_i(\theta_j - b_{i3})\right]}{1 + \exp\left[Da_i(\theta_j - b_{i3})\right]}$$

Altogether, the complete mathematical function of GRM is expressed by Reckase (2009) as:

$$P(U_{ikj} = u \mid \theta_j, a_i, b_{ik}) = P_{ik}(\theta_j)_{\text{GRM}} = \frac{\exp\left[Da_i(\theta_j - b_{ik})\right] - \exp\left[Da_i(\theta_j - b_{i,k+1})\right]}{\left[1 + \exp\left[Da_i(\theta_j - b_{ik})\right]\right]\left[1 + \exp\left[Da_i(\theta_j - b_{i,k+1})\right]\right]} \quad (9)$$

Information

The measurement precision in an IRT system can be characterized as a function of θ. Thus, precision does not have to be represented by a single overall reliability, as in the traditional test theory. Precision in an IRT system is often described in terms of the information function ($I(\theta)$), the conditional error variance (σ_e^2), or the standard error (σ_e) which also vary as functions of θ. The standard error of measurement of the trait estimates ($\sigma_e(\hat{\theta})$) is the reciprocal of the square root of the test information:

$$\sigma_e(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}} \tag{10}$$

The test information for trait estimates is computed by summing the information of the items contributing to the test score (Lord, 1980):

$$I(\hat{\theta}) = \sum_{i=1}^{n} I_i(\hat{\theta}) \tag{11}$$

Therefore, the item information function in IRT indicates the contribution of each item to score precision within particular regions of the θ scale. Nonetheless, Green, Bock, Humphreys, Linn, and Reckase (1984) (as cited in Sireci et al., 1992, p. 240; Thissen & Orlando, 2001, p. 119) suggested marginal reliability in case it is desirable to present a single number that summarizes test precision for tests constructed using IRT. The calculation of marginal reliability is analogous to that for

average reliability in the traditional test theory. First, the average (marginal) measurement error variance for a population with proficiency density $g(\theta)$ is computed. The formula is:

$$\overline{\sigma}_e^2(\theta) = \int \sigma_e^2(\theta)\, g(\theta)\, d(\theta) \tag{12}$$

where $\sigma_e^2(\theta)$ is the expected value of the error variance associated with the latent ability estimate at θ. The formula for marginal reliability in general and for standardized θ in particular is shown in equations (13) and (14) respectively.

$$\overline{\rho} = \frac{\sigma_e^2(\theta) - \overline{\sigma}_e^2(\theta)}{\sigma_e^2(\theta)} \tag{13}$$

$$\overline{\rho} = 1 - \overline{\sigma}_e^2(\theta) \tag{14}$$

The item information functions for the 3PL, 2PL, and 1PL/Rasch models are expressed in equations (15), (16), and (17) respectively.

$$I_i(\theta)_{3\text{PL}} = \left[\frac{D^2 a_i^2 (1 - P_i(\theta)_{3\text{PL}})}{P_i(\theta)_{3\text{PL}}}\right]\left[\frac{(P_i(\theta)_{3\text{PL}} - c_i)^2}{(1 - c_j)^2}\right] \tag{15}$$

When $c_i = 0$, equation (15) is equivalent to the information for 2PL model below,

$$I_i(\theta)_{2\text{PL}} = Da_i[P_i(\theta)_{2\text{PL}}][1 - P_i(\theta)_{2\text{PL}}] \tag{16}$$

and in addition, if $a_i = 1$, then the equation (16) simplifies to the information function for the 1PL/Rasch model, as shown in equation (17) below:

$$I_i(\theta)_{1\text{PL}} = [P_i(\theta)_{1\text{PL}}][1 - P_i(\theta)_{1\text{PL}}] \tag{17}$$

Given that PCM is the generalization of RSM, only the information function for the former will be described. The item information functions for the PCM and GPCM are given in equations (18) and (19) respectively.

$$I_i(\theta)_{\text{PCM}} = \sum_{k=0}^{m} \left[U_{ik} - E(U_{ik} \mid \theta)\right]^2 P_{ik}(\theta)_{\text{PCM}} \tag{18}$$

$$I_i(\theta)_{\text{GPCM}} = D^2 a_i^2 \sum_{k=0}^{m} \left[U_{ik} - E(U_{ik} \mid \theta)\right]^2 P_{ik}(\theta)_{\text{GPCM}} \tag{19}$$

where $E(U_{ik} \mid \theta)$ is the expected score range from 0 to M as a function of θ.

To simplify the notation for the information function for GRM, the following simplified notations are used: $P_{ik}^*(\theta_j) = P*(U_{ij} = k \mid \theta_j)$ and $Q_{ik}^*(\theta_j) = 1 - P*(U_{ij} = k \mid \theta_j)$. The mathematical expression for the GRM information provided by a test item is shown by Reckase (2009) as:

$$I_i(\theta)_{\text{GRM}} = \sum_{k=1}^{m+1} \frac{\left[Da_i^2 P_{i,k-1}^*(\theta_j) Q_{i,k-1}^*(\theta_j) - Da_i^2 P_{ik}^*(\theta_j) Q_{ik}^*(\theta_j)\right]^2}{P_{i,k-1}^*(\theta_j) - P_{ik}^*(\theta_j)} \tag{20}$$

Multidimensional Item Response Theory Models

In practice, examinees response data seldom meet the rigorous assumptions of the UIRT models. The nature of educational tests especially the ones instigated by the CCSS and the NGSS are inherently complex and often not unidimensional. Thus, it is usually not appropriate to fully define the latent ability space with only one latent factor. If such a claim is made, other IRT models that allow for more than one latent factor of dimension could be deployed. In addition to UIRT models, there are also a collection of mathematical models that have been formulated and are useful to describe the complex interactions between examinees and test items (i.e., item response theories). These models differ from the UIRT models in that they postulate that multiple hypothetical constructs influence the performance on test items instead of only one hypothetical construct (Reckase, 2009).

The most commonly used models are for items scored dichotomously or using two score categories although MIRT models for items with more than two score categories (i.e., polytomous items) are also gaining popularity in operational settings (e.g., Adams, Wilson & Wang, 1997; Muraki & Carlson, 1995; Yao & Schwarz, 2006). The basic form of the models considered here is given by Reckase (2009) as

$$P(U = u \mid \boldsymbol{\theta}) = f(\boldsymbol{\theta}, \boldsymbol{\gamma}) \tag{21}$$

In equation (21), $\boldsymbol{\theta}$ represents a vector of parameters denoting the location of the examinee(s) in the multidimensional space, $\boldsymbol{\gamma}$ is a vector of parameters describing the characteristics of test items. U is the score on the test item for a particular

examinee, u denotes a possible value for the score, and f indicates a function that describes the relationship between the parameters and the probability of the response, $P(U = u)$.

MIRT scholars (e.g., Ackerman, 1989; 1996; Reckase, 2009) have often classified MIRT models into two categories. The first category is commonly known as compensatory models. The model from this category is based on a linear combination of coordinates of θ. The linear combination is used to specify the probability of a response. The linear combination of θ-coordinates can produce the same sum with various combinations of θ-values. If one θ-coordinate is low, the sum will be the same if another θ-coordinate is sufficiently high.

The second category of model is often called noncompensatory. This type of model separates the cognitive tasks in a test item into parts and uses a UIRT model for each part. The probability of correct response for the item is the product of the probabilities of each part. The fact that the probability of correct response cannot exceed the highest of the probabilities in the product reduces the compensation of a high θ-coordinate for a low θ-coordinate. Reckase (2009) preferred to refer to this category of model as partially compensatory "because a high θ-coordinate on one dimension does actually yield a higher probability of response than a low value on that dimension" (p. 79), resulting in some compensation.

In this review, I will only describe the first category of the MIRT models—the compensatory models—for both dichotomous and polytomous items. The models

that are described here are the ones that most commonly appear in the research literature and are the extensions of the UIRT models described in the previous section.

MIRT Logistic Model for Dichotomous Items

A fairly straightforward extension of the 3PL UIRT model produces the multidimensional version (M3PL). The model (Reckase, 2009, 1985) is mathematically given by the following equation:

$$P(U_i = 1 \mid \boldsymbol{\theta}_\mathbf{j}, \mathbf{a_i}, c_i, d_i) = P_{ij}(\boldsymbol{\theta})_{\text{M3PL}} = c_i + (1 - c_i)\left[\frac{\exp[D(\mathbf{a_i}\boldsymbol{\theta}_j + d_i)]}{1 + \exp[D(\mathbf{a_i}\boldsymbol{\theta}_\mathbf{j} + d_i)]}\right] \qquad (22)$$

where $\boldsymbol{\theta}$ is a $1 \times p$ vector of person coordinates (person abilities or traits) with p indicating the number of dimensions or latent factors in the coordinate space. Suppose that there are p latent factors, $\boldsymbol{\theta}_\mathbf{j} = \theta_{\mathbf{j}1}, \ldots, \theta_{\mathbf{j}p}$; $\mathbf{a}_\mathbf{i}$ is a $1 \times p$ vector of item discrimination parameters or item factor loadings, $\mathbf{a_i} = a_{\mathbf{i}1}, \ldots, a_{\mathbf{i}p}$; d_i is the item intercept parameter also known as a location parameter. The $\mathbf{a}_\mathbf{i}$ and the d_i could not be compared directly to the unidimensional item discrimination and item difficulty from the UIRT model (the statistical descriptions of item and test functioning for the multidimensional case will be provided after the polytomous MIRT model section), c_i is a single lower asymptote or pseudo-guessing parameter to specify the probability of correct response for examinees with very low values in

$\boldsymbol{\theta}$, and D is a scaling adjustment (usually 1.702) used to make the logistic metric more closely correspond to the traditional normal ogive metric (Reckase, 2009).

The multidimensional 2PL (M2PL) model (McKinley & Reckase, 1983; Reckase, 2009; Reckase & McKinley, 1991) follows directly from the M3PL model above but with the absence of the lower asymptote parameter, c_i. The M2PL model can be written as:

$$P(U_i = 1 | \boldsymbol{\theta}_\mathbf{j}, \mathbf{a}_\mathbf{i}, d_i) = P_{ij}(\theta)_{\text{M2PL}} = \left[\frac{\exp[D(\mathbf{a}_\mathbf{i}\boldsymbol{\theta}_\mathbf{j} + d_i)]}{1 + \exp[D(\mathbf{a}_\mathbf{i}\boldsymbol{\theta}_\mathbf{j} + d_i)]} \right] \tag{23}$$

The multidimensional 1PL or Rasch (M1PL) (Reckase, 2009) is given by the following equation:

$$P(U_i = 1 | \boldsymbol{\theta}_\mathbf{j}, \mathbf{a}_\mathbf{i}, d_i) = P_{ij}(\theta)_{\text{M1PL}} = \left[\frac{\exp(\mathbf{a}_\mathbf{i}\boldsymbol{\theta}_\mathbf{j} + d_i)}{1 + \exp(\mathbf{a}_\mathbf{i}\boldsymbol{\theta}_\mathbf{j} + d_i)} \right] \tag{24}$$

where $\mathbf{a}_\mathbf{i}$ is a vector with elements that indicate the dimension or dimensions that are required to obtain the correct score on item i and d_i is a scalar.

Without the scaling adjustment, D, in the M2PL model, the mathematical expression of the M1PL model appears to be identical to the one for the M2PL model. However, the difference between the two is the way that the $\mathbf{a}_\mathbf{i}$ vector is specified (Reckase, 2009). In M2PL, $\mathbf{a}_\mathbf{i}$ is a characteristic of item i that is estimated from the data. In M1PL, $\mathbf{a}_\mathbf{i}$ is a characteristic of item i that is specified by the test

developer. In the case of the M2PL model, statistical estimation procedures are used to determine the elements of $\mathbf{a}_i$ that will maximize some criterion for model/data fit. For the M1PL model, the values are specified by the analyst. According to Reckase (2009), the elements of $\mathbf{a}_i$ in M2PL can take on any values (except for the usual monotocity constraint that requires the values of the $\mathbf{a}_i$ elements to be positive) while the elements of $\mathbf{a}_i$ in M1PL typically take on integer values.

MIRT Logistic Model for Polytomous Items

The multidimensional extension of the GPCM (MGPCM) is formulated to model the interaction of persons with items that are scored with more than two categories. As previously mentioned in the UIRT section, the score assigned to an examinee on the item is represented by $u = (0, 1, 2, \ldots, M)$ in which there are a total of $M + 1$ score categories and $k = 0, 1, \ldots, m$. The mathematical expression of the MGPCM is given by Yao & Schwarz (2006, p.471) and is reparameterized by Reckase (2009, p.103) in the following equation:

$$P(U_{ikj} = u \mid \boldsymbol{\theta}_j, \mathbf{a}_i, \beta_{ik}) = P_{ik}(\boldsymbol{\theta}_j)_{\text{MGPCM}} = \frac{\exp k\mathbf{a}_i\boldsymbol{\theta}_j \sum_{c=0}^{k} \beta_{ic}}{\sum_{k=0}^{m} \exp k\mathbf{a}_i\boldsymbol{\theta}_j - \sum_{c=0}^{k} \beta_{ic}} \tag{25}$$

where β_{ik} is the threshold parameter for score category k, and all other symbols follow their previously defined meaning. There are two important differences between the equation for the MGPCM and that for the UIRT GPCM given in equation

(29). First, the model is parameterized such that difficulty and threshold parameters are no longer separated. Second, since $\boldsymbol{\theta}$ is a vector and the β s are scalars, it is not possible to subtract the threshold parameter from $\boldsymbol{\theta}$ (Reckase, 2009).

There are a number of simplifications of the multidimensional version of the GPCM that have the special properties of the 1PL/Rasch model in which they have observable sufficient statistics for the item and person parameters. Adams et al. (1997) presented one form of the multidimensional extension of PCM (MPCM). The model is presented below, with consistent notations from the previous models.

$$P(U_{ikj} = u \mid \boldsymbol{\theta}_{\mathbf{j}}, \mathbf{a}_{\mathbf{ik}}, d_{ik}) = P_{ik}(\boldsymbol{\theta}_{\mathbf{j}})_{\text{MPCM}} = \frac{\exp(\mathbf{a}_{\mathbf{ik}}\boldsymbol{\theta} + d_{ik})}{\sum_{k=0}^{m} \exp(\mathbf{a}_{\mathbf{ik}}\boldsymbol{\theta} + d_{ik})} \tag{26}$$

in which all $\mathbf{a}_{\mathbf{ik}}$ are constraint to be equal.

For Samejima's (1969) multidimensional GRM (MGRM) (see also Muraki & Carlson, 1995), suppose again that there are unique k steps ($M+1$) for item i, with intercepts $d_i = d_{i1}, \ldots, d_{i(k-1)}$. Thus, the boundary of response probabilities can be defined as

$$P_{i0}(\boldsymbol{\theta}_{\mathbf{j}}) = 1 - \frac{\exp\left[D\mathbf{a}_{\mathbf{i}}(\boldsymbol{\theta}_{\mathbf{j}} + d_{i1})\right]}{1 + \exp\left[D\mathbf{a}_{\mathbf{i}}(\boldsymbol{\theta}_{\mathbf{j}} + d_{i1})\right]}$$

$$P_{i1}(\boldsymbol{\theta}_{\mathbf{j}}) = \frac{\exp\left[D\mathbf{a}_{\mathbf{i}}(\boldsymbol{\theta}_{\mathbf{j}} + d_{i1})\right]}{1 + \exp\left[D\mathbf{a}_{\mathbf{i}}(\boldsymbol{\theta}_{\mathbf{j}} + d_{i1})\right]} - \frac{\exp\left[D\mathbf{a}_{\mathbf{i}}(\boldsymbol{\theta}_{\mathbf{j}} + d_{i2})\right]}{1 + \exp\left[D\mathbf{a}_{\mathbf{i}}(\boldsymbol{\theta}_{\mathbf{j}} + d_{i2})\right]}$$

$$\vdots$$

$$P_{ik-1}(\boldsymbol{\theta}_{j}) = \frac{\exp\left[D\mathbf{a}_{\mathbf{i}}(\boldsymbol{\theta}_{\mathbf{j}} + d_{ik-1})\right]}{1 + \exp\left[D\mathbf{a}_{\mathbf{i}}(\boldsymbol{\theta}_{\mathbf{j}} + d_{ik-1})\right]} - \frac{\exp\left[D\mathbf{a}_{\mathbf{i}}(\boldsymbol{\theta}_{\mathbf{j}} + d_{ik})\right]}{1 + \exp\left[D\mathbf{a}_{\mathbf{i}}(\boldsymbol{\theta}_{\mathbf{j}} + d_{ik})\right]}$$

$$P_{ik}(\boldsymbol{\theta}_{\mathbf{j}}) = \frac{\exp\left[D\mathbf{a}_{\mathbf{i}}(\boldsymbol{\theta}_{\mathbf{j}} + d_{ik})\right]}{1 + \exp\left[D\mathbf{a}_{\mathbf{i}}(\boldsymbol{\theta}_{\mathbf{j}} + d_{ik})\right]} - 0$$

As in the UIRT representation, these boundaries lead to the probability that an examinee j will receive a score of k, which is:

$$P(U_{ij} = k \mid \boldsymbol{\theta}_{\mathbf{j}}, \mathbf{a}_{\mathbf{i}}, d_i) = P^*(U_{ij} = k \mid \boldsymbol{\theta}_{\mathbf{j}}, \mathbf{a}_{\mathbf{i}}, d_i) - P^*(U_{ij} = k+1 \mid \boldsymbol{\theta}_{\mathbf{j}}, \mathbf{a}_{\mathbf{i}}, d_i) \qquad (27)$$

Altogether, the complete mathematical function of MGRM can be written as:

$$P_{ik}(\boldsymbol{\theta}_{\mathbf{j}})_{\text{MGRM}} = \frac{\exp\left[D\mathbf{a}_{\mathbf{i}}(\boldsymbol{\theta}_{\mathbf{j}} + d_{ik})\right] - \exp\left[D\mathbf{a}_{\mathbf{i}}(\boldsymbol{\theta}_{\mathbf{j}} + d_{i,k+1})\right]}{\left[1 + \exp\left[D\mathbf{a}_{\mathbf{i}}(\boldsymbol{\theta}_{\mathbf{j}} + d_{ik})\right]\right]\left[1 + \exp\left[D\mathbf{a}_{\mathbf{i}}(\boldsymbol{\theta}_{\mathbf{j}} + d_{i,k+1})\right]\right]} \qquad (28)$$

Items and Test Statistics for MIRT Model

The MIRT models that have been described provide mathematical descriptions of the interactions of persons and test items. While the parameters of the models summarize the characteristics of the items, the vectors of item parameters could not be directly compared to the corresponding item parameters

from the UIRT model and thus lack intuitive meaning. This section contains a description of items and test characteristics for the MIRT models.

Multidimensional Item Discrimination (MDISC). The UIRT discrimination parameters are compared with a multidimensional scalar discrimination index, MDISC (Reckase, 1985; Reckase & McKinley, 1991). MDISC is the norm of the vector of the MIRT discrimination parameter estimates and represents an item's maximum discrimination in a particular direction of the factor space. MDISC has the same relationship to multidimensional item difficulty as the item discrimination parameter (a_i) has to the item difficulty parameter (b_i) for the UIRT model. MDISC is a measure of an item's capacity to distinguish between examinees that have different locations in the factor space. If an item has a high value of MDISC, then it will provide a relatively large amount of information somewhere in the factor/trait space. MDISC for each item i is defined as the following:

$$\text{MDISC}_i = \sqrt{\sum_{f=1}^{p} \hat{a}_{if}^2} \tag{29}$$

where P represents a dimension/latent factor and $\hat{a}_i$ represents an estimate of item discrimination for a given dimension.

Multidimensional Item Difficulty (MDIFF). Because the item intercept/location parameter (d_i) does not correctly represent a difficulty parameter, MDIFF (Reckase, 1985) or the signed distance is used as the comparative difficulty or location

parameter estimate corresponding to the UIRT (b_i) to compensate for the confounding of direction and location present in the multidimensional model parameter d_i.

MDIFF for the M2PL model (McKinley & Reckase, 1983)) represents the distance and direction from the origin in the θ -space to the point of the steepest slope. The MDIFF formula for a dichotomous item is

$$\text{MDIFF}_i = \frac{-\hat{d}_i}{\text{MDISC}_i} \tag{30}$$

where $\hat{d}_i$ is an estimate of the item intercept/location parameter.

Reckase (2009) noted that the description of test items using the concepts of MDIFF, MDISC , and direction of steepest slope in the multidimensional space can also be used with polytomous items. Muraki and Carlson (1995) derived the statistics for the MGRM (Samejima, 1969). The MDIFF for the step difficulty for an item can be written as:

$$\text{MDIFF}_{ik} = \frac{-\hat{d}_{ik}}{\text{MDISC}_i} \tag{31}$$

where MDIFF_{ik} is the step difficulty for the step k of the GRM item and $\hat{d}_{ik}$ is the estimate of the step parameter for item i.

Yao & Schwarz (2006, p. 479) derived the MDIFF_i for the MGPCM. Following Reckase's (2009, p.103) mathematical notations, the MDIFF for the step difficulty for an item can be written as:

$$\mathrm{MDIFF}_{ik} = \frac{\sum_{c=0}^{k} \beta_{ic}}{\mathrm{MDISC}_i} \tag{32}$$

The concept of item information employed in UIRT can also be generalized to the multidimensional case. The multidimensional item information (MINF) was first defined by Reckase & Mckinley (1991) when they introduced the M2PL model in which item discrimination is employed in more than one dimension, MDISC . They noted the relation of MINF to the MDISC in which an item with a high value of MDISC will have a large amount of information in the latent ability space. MINF however differs from MDISC since it describes "the capability of the item to discriminate at each point in the space, rather than just at the steepest point of the item response surface" (p. 356).

The Angle Measure. Another statistic associated with both item discrimination and difficulty is the angle measure. Reckase (1985) proposed describing multidimensional difficulty by both the MDIFF and the angle measure or direction cosines. The use of direction cosines removes any confounding of the item location parameters with the discriminations and provides an angular measure of the direction of maximum discriminating power of each item with respect to the

latent abilities axes. The direction of greatest/steepest slope, in degrees, from the origin with dimension p for item i is given by:

$$\hat{\alpha}_{ip} = \cos^{-1}\left(\frac{\hat{a}_{ip}}{\text{MDISC}}\right) \tag{33}$$

where $p \geq 1$ dimension(s). This reference angle represents the composite of the latent ability space ($\boldsymbol{\theta}_j$) that item i best measures (Reckase, 1985, 2009; Ackerman, 1994a, 1994b).

Multidimensional Item Information (MINF). The item information in the multidimensional case is commonly known as MINF. MINF is also used to provide measurement precision of a given item in which the reciprocal of the information function is the asymptotic variance of the ability estimate (Ackerman, 2005). This relationship indicates that higher information function will reduce the asymptotic variance, thus increasing the measurement precision. MINF is computed similarly to the computation of item information for its UIRT counterpart except that the direction of the information must also be considered.

The MINF formula was originally introduced by Reckase and McKinley (1991, p. 365). Reckase (2009, p. 121) provided the generalization of MINF as:

$$\text{MINF} = I_{\alpha i}(\boldsymbol{\theta}) = \frac{\left[\nabla_{\alpha} P(\boldsymbol{\theta})\right]^2}{P(\boldsymbol{\theta})Q(\boldsymbol{\theta})} \tag{34}$$

where $\boldsymbol{\alpha}$ is the vector of angles with the coordinate axes that defined the direction taken from the $\boldsymbol{\theta}$ –point, ∇_{α} is the directional derivative or gradient, in the direction α, $P(\boldsymbol{\theta})$ is the probability of correct response for $\boldsymbol{\theta}$ skills , and $Q(\boldsymbol{\theta})$ is the probability of incorrect response which can also be rewritten as $1-P(\boldsymbol{\theta})$. Complete derivations of the item information function are given by Reckase (2009, pp. 121-123). Test information is simply the sum of item information values:

$$I_T(\boldsymbol{\theta}) = \sum_{i=1}^{n} I_{\boldsymbol{\alpha} i}(\boldsymbol{\theta}) \tag{35}$$

Reckase and McKinley (1991, p. 367) derived MINF for the M2PL model as

$$\text{MINF} = P(\boldsymbol{\theta})[1-P(\boldsymbol{\theta})]\left[\sum_{f=1}^{p} \alpha_{if} \cos \alpha_f\right]^2 \tag{36}$$

where $\boldsymbol{\alpha}_{\mathbf{ip}}$ represents the angle between the vector representing item i and the θ_1 axis for dimension p. MINF provides a measure of information at any $\boldsymbol{\theta}$ value on the latent ability space (i.e., measurement precision relative to the composite).

If the direction of greatest/steepest slope from equation (33) is substituted in equation (36) for MINF of the M2PL model, the result is the MINF in the direction of maximum slope. It is given mathematically by Reckase (2009, p. 123) as:

$$\text{MINF} = I_{\alpha i \max} = P(\boldsymbol{\theta})[1-P(\boldsymbol{\theta})]\sum_{f=1}^{p} \alpha_{ip}^2 = P(\boldsymbol{\theta})Q(\boldsymbol{\theta})\text{MDISC}^2 \tag{37}$$

Yao & Schwarz (2007) provided the formulas and corresponding derivations for both M3PL model and MGPCM respectively.

Sources of Local Item Dependence

Scholars have discussed various potential factors of LID that can violate the assumption of LInd. These factors could result from either the examinees or the nature of the test and of the test items or from the interaction of both (Chen & Thissen, 1997; Haladyna, 1992; Sick, 2010; Yen, 1984, 1993). Yen (1993) listed several examinee effects on test items, which are often uncontrolled, that could result in LID. Some of these include: external assistance or interference in the test taking process; the effect of fatigue or lowered motivation in lengthy test settings; and, the different effect of test practice and test-taking strategy. Items that measure unique content for examinees with different background knowledge, proficiency, and opportunity to learn can also exhibit LID. Usually, these items display differential item functioning (DIF) (Clauser & Mazor, 1998; Penfield & Lam, 2000).

Chen & Thissen (1997) categorized the factors that can violate the LInd assumption into two different types: underlying LID and surface LID. The underlying LID model "assumes that there is a separate trait that is common to each set of locally dependent items but it is not common to the rest of the items" (p. 271). The surface LID model is based on the premise that "a pair of items are so similar (in content or in location in the test) that the test taker responds identically to the second item without the underlying processing implied by the IRT model (Thissen,

Bender, Chen, Hayashi, & Wiesen, 1992)" (Chen & Thissen, 1997, p. 272). Examples of surface LID are test speededness, omission of items at the end of the test due to lengthy test, relative position of items in the test, success due to guessing in matching item format where there is an equal number of stems and choices that make it easy to guess the hardest item correctly (Sick, 2010), and redundant survey items (e.g., summary item, negative restatement of another item).

Others (Goodman, 2008; Goodman, Luecht, & Zhang, 2009) have classified the sources of LID into three categories: contextual, scoring, and dimensional. Items that share contextual information may be related to one another in a manner that the primary ability of interest cannot be explained. Many studies on LID have focused on passage-dependent items (Sireci, et al., 1991; Thissen et al., 1989; Wainer & Kiely, 1987; Wainer & Lewis, 1990; Zenisky, Hambleton, & Sireci, 2002) and items that are built based on an associated set of items. The latter items provide context for future items such as in cloze tests (Baghaei & Ravand, 2016; Sick, 2010) or require multi-step solutions (Ferrara, Huynh, and Michaels, 1999; Yen, 1984) and an explanation of the reasoning or process behind the answer (Ferrara, Huynh, and Baghi, 1997; Yen, 1993). A similar concept of contextual dependence can be extended to items that share a common setting, theme, stimulus, distractors, set of directions and scenarios, or set of resources (Haladyna, 1992; Rosenbaum, 1988; Wang, Cheng, & Wilson, 2005). Ferrara et al. (1997) and Yen (1993) observed substantial amounts of LID in sets of math problems that are linked to a common theme or stimuli. Yan (1997) and Ferrara et al. (1999) demonstrated that science

assessments associated with a common experiment, graphic, table, or general topic tend to display LID that is due in part to contextual attributes.

Choices in item-level scoring procedure, especially on TE (Bukhari et al., 2016; Lorie, 2014) and computerized PA (Goodman, 2008; Goodman et al., 2009; Stark, Chernyshenko, Drasgow, 2002) item formats, may also lead to LID between items. Scoring procedures that share objects (e.g., stimulus, instructions, scenarios) by awarding credit in more than one place for a correct response on a particular item can also lead to dependencies due to scoring. Similarly, an item that requires explanation of the previous answer or multiple problem-solving steps (Yen, 1984, 1993) that are each graded separately (i.e., dichotomous (Bukhari et al., 2016; Goodman, 2008; Goodman et al., 2009; Lorie, 2014; Stark et al., 2002) and componential (Lorie, 2014) scoring rules) may also result in scoring dependency. Bukhari et al. (2016) extended Stark et al. (2002) and Lorie's (2014) studies by comparing the amount of UIRT information from the evidence-based selected response (EBSR) item format (see Table 1) using three different scoring procedures: (1) a polytomous TE scoring rule with penalty for guessing where students would get a score of zero for an incorrect response to the first item, even if the second item was correct; (2) traditional polytomous scoring where students will receive partial credit if they answer at least one item in the EBSR pair correctly and full credit if they answer both items correctly; and (3) a dichotomous SR scoring rule in which the EBSR item is treated as two separate SR items. The EBSR item format combines two SR items in which (in the second SR item) students are asked to show evidence

from the text that supports the answer they provided to the first SR item. Scoring dependency may occur when EBSR is scored separately using the dichotomous scoring rule.

A test can be considered to have some degree of multidimensionality when items require more than one skill, knowledge, and abilities to successfully explain an examinee's response. A test that employs different item or response formats to assess its construct can also display multidimensionality. As Chen and Thissen (1997) described, (underlying) LID is an indicator that multiple proficiency traits may be underlying the collective response patterns for a set of items which are uncommon to the rest of the items in a test. If the relative magnitude of the multidimensionality is large, the residual covariance cannot be ignored and regarded as a result of nuisance dimensions. Once subject-matter experts determine that the constructs are essential to the purposes of the test, additional score scales may be required.

Measuring Local Item Dependence

In general, LID measures perform by examining for departures from what would be expected if there was no LID. The indices/statistics differ given particular aspect of the model they examine. Non-exhaustive lists of methods for assessing LID have been developed in the IRT literature. These include Yen's (1984) Q_2 and Q_3, Stout's (1987) DIMTEST procedure, Chen and Thissen's (1997) use of Pearson's χ^2,

the likelihood ratio G^2 statistic, the standardized ϕ coefficient difference, the standardized log-odds ratio difference(τ), comparison of reliability estimates of testlet-unit and independent items (Wainer & Thissen, 2001, 1996), Tsai and Hsu's (2005) absolute value of mutual information difference (AMID), a suggestion to use the Mantel–Haenszel test with multiple testing corrections (Ip, 2001), and Gessaroli and De Champlain's (1996) NOHARM-based χ^2 approximation. Some of the most successful methods (Chen & Thissen, 1997; Houts & Edwards, 2013; Kim, DeAyala, Ferdous, & Nering, 2011) for assessing LID that are used in practice will be reviewed in this section.

Chen & Thissen (1997) LID Statistics

The first two methods, Pearson's χ^2 and the likelihood ratio G^2 test, are described using the observed and expected frequencies of score patterns for pairs of items in contingency tables to assess LID (Chen & Thissen, 1997, p.268). To detect LID, both statistics test whether the observed frequencies conform to the expected frequencies under the null hypothesis of LInd. Following Chen & Thissen (1997, p. 268), for each item pair with dichotomous responses, the following (marginal) Table 2 can be constructed for the observed frequencies:

Table 2. Two by Two Table for Observed Frequencies

Item i \ Item h	0	1
0	O_{11}	O_{12}
1	O_{21}	O_{22}

In this table, O_{pq} is the observed frequency, where 1 and 0 represent the correct and incorrect responses, respectively. For example, the response vectors for 47 examinees having the same latent ability θ, to two test items are as follows:

Item 1: 01110101010111110000010010100001010101010100101

Item 2: 00010101010100000111101000001000011110001101001

(Marginal) Table 3 for the observed frequency for Items 1 and 2 will be:

Table 3. Example of a Two by Two Table for Observed Frequencies

Item 1 \ Item 2	0	1
0	15	10
1	13	9

The same structure applies to the expected frequencies as shown in Table 4 below:

Table 4. Two by Two Table for Expected Frequencies

		Item h	
		0	1
Item i	0	E_{11}	E_{12}
	1	E_{21}	E_{22}

In this table, E_{pq} is the expected frequency that is predicted by the IRT model:

$$E_{pq} = N \int_{-\infty}^{\infty} P_i(\theta)^p P_h(\theta)^q [1 - P_i(\theta)]^{(1-p)} [1 - P_h(\theta)]^{(1-q)} f(\theta)\, d\theta \qquad (38)$$

where N is the number of examinees, $f(\theta)$ is the population distribution for examinee locations (typically assumed to be $N[0,1]$) and $P_i(\theta)$ and $P_h(\theta)$ are the probability of a correct response on (or the ICCs for) items i and h respectively, according to an IRT model. The integral is approximated numerically. Both statistics are formulated by Bishop, Fienberg, & Holland (1957, p.57, as cited in Chen & Thissen, 1997, pp. 269-260) and are distributed as χ^2 with ($K-1$) degrees of freedom, where K is the number of score categories.

Chen & Thissen (1997) applied the Pearson's χ^2 as the index for standardized LID χ^2 value to dichotomously scored data that was calibrated using UIRT models. The formula is given as:

$$\chi^2 = \sum_{h=1}^{2}\sum_{i=1}^{2}\frac{(O_{hi} - E_{hi})^2}{E_{hi}} \tag{39}$$

where O_{hi} is the observed correlation between item pair i and $\boldsymbol{h}$, and E_{hi} is the model-implied expected response frequencies for each item pair. For this test of independence for dichotomous data in 2 X 2 tables, with $K = 2$, the degree of freedom is one, $df = 1$. Lin, Kim and Cohen (2006, as cited in Goodman, 2008, p.37) extended the formula for application to polytomous data in which it is written as:

$$\chi^2 = \sum_{h=1}^{K}\sum_{i=1}^{K}\frac{(O_{hi} - E_{hi})^2}{E_{hi}} \tag{40}$$

where K is the maximum number of score categories, and O_{hi} and E_{hi} are the observed and model-derived expected values for the cells in the $K \times K$ table.

Similar to the Pearson's χ^2, the likelihood ratio G^2 test (Chen & Thissen, 1997) is designed to detect differences between observed and expected frequencies of score patterns. The formula for the likelihood ratio G^2 for both dichotomous and polytomous data are given respectively in equation (41) and (42)

$$G^2 = -2\sum_{h=1}^{2}\sum_{i=1}^{2} O_{hi} \ln\left(\frac{E_{hi}}{O_{hi}}\right) \tag{41}$$

$$G^2 = -2\sum_{h=1}^{K}\sum_{i=1}^{K} O_{hi} \ln\left(\frac{E_{hi}}{O_{hi}}\right) \tag{42}$$

where the elements of this equation are defined in the same manner as in the Pearson's χ^2 statistic. Following Chen and Thissen (1997), if an observed cell is empty (e.g., $O_{hi} = 0$), the contribution to G^2 from that cell, which from the formula would be undefined, is set to zero.

Significant Pearson's χ^2 and G^2 statistics indicate that items h and i are locally dependent. Chen (1996, as cited in Thompson & Pommerich, 1996, p. 5) has recommended that item pairs with values greater than 10.0 to be flagged for potential LID. Both methods are effective in detecting dependent item pairs, but are limited to detecting the presence and not the direction of LID. G^2 has been shown to be slightly more powerful (i.e., power and Type I error rate) than χ^2 in detecting LID (Chen & Thissen, 1997).

When the two statistics were first introduced (Chen & Thissen, 1997), they were computed using the IRT_LD computer program developed by Chen (1993). A free FORTRAN-based program written by Kim, Cohen, and Lin (2006) computes the two LID indices for dichotomous and polytomous data and is available on request. A 'mirt' package (Chalmers, 2012) in R program (R Core Team, 2016) and flexMIRT®3.0RC (Vector Psychometric Group, 2017) can also output both indices using specified arguments.

The Jackknife Slope Index (JSI)

A recently developed index for LID detection, the Jackknife Slope Index (JSI), introduced by Edwards and Cai (2008, 2011), is based on the observation that locally dependent items often exhibit inflated slopes. Using this phenomenon as a basis for a jackknife-type procedure, they suggested obtaining item parameter estimates for a full data set including all items and, in the subsequent steps, removing one item to obtain revised item parameter estimates. Edwards & Cai (2011, p. 13) explained in their own words: "We calculate all the slope parameters and then, one at a time, omit an item and re-analyze the remaining items. For each item we take a difference between the "full set" slope and the "minus one" slope and divide it by the standard error of the "minus one" slope."

A single value of the JSI for item h when item i is removed from the scale is then calculated as:

$$JSI_{h(i)} = \frac{a_h - a_{h(i)}}{se[a_{h(i)}]} \tag{43}$$

where a_h is the full IRT data slope estimate, h indexes the item impacted, i indexes the removed item, and $se[a_{h(i)}]$ is the standard error of the item-removed slope parameter (or in other words $se[a_h]$ is the standard error of the slope parameter when estimated with all items included). For each pair of items, a JSI value is calculated for the slope change in the first item induced by removing the second item, as well as the slope change in the second item induced by removing the first

item. Each item receives a vector of $n-1$ diagnostics, one calculated with the removal of each other item. The resulting n item by n item matrix, with empty diagonals, is inspected by the user and item pairs with JSI values substantially larger than the other values in the matrix indicates an item pair that should be noted as possibly exhibiting LID (Houts & Cai, 2015). "If a set of items is unidimensional, removing any individual item should have virtually no impact on the slopes of the remaining items. On the other hand, if the item removed is locally dependent, then the user "might expect to see a fairly significant change in the slope of the remaining offender" (Edwards & Cai, 2011, p.11). flexMIRT®3.0RC (Vector Psychometric Group, 2017) can calculate and output the JSI index using specified arguments.

Comparison of Reliability Estimates

The presence of LID can also be tested by comparing separate reliability estimates of the same tests (Sireci & et al., 1991; Wainer & Thissen, 2001, 1996; Zenisky et al., 2002). The first reliability estimate assumes that all items are locally independent (i.e., item-level reliability). The second estimate models the reliability after forming testlets (i.e., testlet-unit reliability) for context-dependent sets of items. If the testlet-unit reliability is substantially lower than the item-level estimate, LID is present for some or all of the items in the testlets. Sireci et al. (1991) reported that testlet-unit reliability estimate was about 10-15% lower than item-level reliability estimates for two reading comprehension tests with four passages and five to twelve SR items connected to each passage. Wainer & Thissen (1996),

using several forms of a state accountability SR reading test, found that the difference between the reliability estimates for testlet-unit and item-level was smaller but always in the direction that the testlet-unit estimates were lower. Their further analysis over several admission tests revealed that the more items are linked to each testlet, the greater is the reduction in testlet reliability estimate relative to item-level reliability estimates. In conclusion, when testlet items are locally dependent, testlet-unit reliability should be used (Sireci et al., 1991; Wainer & Thissen, 1996, 2001). The reliability estimate can be computed using the standardized α coefficient for traditional test theory or using marginal reliability as shown in equations (34) to (36) when IRT is employed.

Methods to Assess LID due to Test Dimensionality

If the source of the LID is presumed to be due to dimensionality, several methods are effective to either explore or confirm the dimensionality structure of a set of items. In fact, there is an extensive number of approaches to analyze test dimensionality (e.g., Brown, 2006; Fraser & McDonald, 1988; Gorsuch, 1983; Stout, 1987; Nandakumar & Stout, 1993; Stout et al., 1996; Reckase, 2009).

> Test dimensionality assessment methods can generally be organized according to a two-by-two classification scheme. First, the methods can be categorized as either parametric or nonparametric. …. Secondly, the methods either attempt full dimensionality estimation (number of dimensions and which items measure which dimensions) or merely attempt to estimate or detect the lack of unidimensionality (whether or not the test is unidimensional). (Roussos, Stout, & Marden, 1998, p. 3)

Several researchers (Burge, 2007; DeChamplain & Gessaroli, 1998; Gonulates, 2004) have provided detailed summaries and delineation of methods for assessing test dimensionality.

The parametric model is a convenient conceptual mechanism to characterize various knowledge or skill dimensions. It aims to provide a parsimonious and quantitative description of data structure. Parametric methods include several approaches: classical factor analytic, item factor analytic, IRT, or some combination of item factor analytic and IRT. The classical factor analytic approaches (e.g., Gorsuch, 1983) refer to the traditional, linear factor analysis of correlation matrices, such as in exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) (e.g., Brown, 2006). Structural equation modeling (SEM) (e.g., Kline, 2011), also subsumed under the classical factor analytic, is used to confirm a proposed dimensional structure or to compare competing dimensional structures. SEM provides a battery of fit statistics (modification indices) and residual matrices for assessing the degree to which the data fits a proposed multidimensional model (e.g., Gessaroli & De Champlain, 1996). The item factor analytic method (e.g., Fraser & McDonald, 1988) is an extension of classical factor analysis. It uses a nonlinear relationship between the probability of a correct examinee response and one or more examinee latent factors or abilities. In this regard, the item factor analysis models are equivalent (McDonald & Mok, 1995) to MIRT models (Bock, Gibbons, & Muraki, 1988; Reckase, 2009).

Nonparametric approaches to measure test dimensionality were motivated by several factors: the failure of parametric IRT models in certain circumstances; the utility of nonparametric methods with small number of items and examinees (Tate, 2003); enabling more efficient data analysis because the approaches are not as computationally intensive as those of parametric methods; and avoidance of strong parametric modeling assumptions while still adhering to the fundamental principles of IRT. Nonparametric methods only assume that the ICC is monotonic thus they are not restricted to the highly-prescriptive assumed models used in parametric approaches; in other words, nonparametric models do not use IRT models and hence do not require the estimation model parameters or do not have to be constrained by model specificity. The use of a nonparametric method does not confound lack of model fit by a particular unidimensional parametric family of models when working with potentially multidimensional data (Stout, 2002). Three nonparametric methods that are commonly used in practice to assess dimensionality are: (1) the test of essential unidimensionality and LInd for dichotomous (DIMTEST) and polytomous (POLY-DIMTEST) items by Stout (1987), Nandakumar & Stout (1993), and Stout et al. (1996); (2) the test of multidimensionality (DETECT) by Kim (1994) and Zhang & Stout (1999a, 1999b); and (3) hierarchical cluster analysis and conditional covariance proximation matrix (HCA/CCPROX) by Roussos (1992) and Roussos, Stout, and Marden (1998). These methods are based on the conceptualization of LID with nonparametric computation of conditional item covariances. The three methods treat

dimensionality as a whole. HCA/CCPROX searches for clusters of homogenous items using cluster analysis. DIMTEST is sensitive to the methods used to generate compensatory multidimensional data (c.f., Hattie, Krakowski, Rogers, & Swaminathan, 1996). DIMTEST tests whether the test data is essentially unidimensional and if it is not, DETECT will be used to calculate the extent of multidimensionality in the test data using the DETECT index. Multidimensionality structure is maximized when the correct number of dimensions is used in partitioning a test (Zhang & Stout, 1999b).

According to Stout et al. (1996), each of the three nonparametric approaches mentioned above addresses a different aspect of test structure but "together they provide an almost complete summary of the test's dimensional characteristics" (p. 351). Gessaroli & De Champlain (1996) proposed an approximate χ^2 test (a parametric approach) to improve the interpretability of the residual item covariances produced by nonlinear factor analysis, and compared it to DIMTEST. They concluded that "the approximate χ^2 was at least as good as Stout's T statistic in all conditions and was better than T with smaller sample sizes and shorter tests" (p. 157).

Yen's Q_3 LID Index

Yen (1984), building on Kingston and Dorans work (1982, as cited in Yen, 1984, p. 127), proposed Q_3 statistics, a correlation of residuals between item pairs from the IRT model after accounting for some measure of performance, θ^*. In this

sense Q_3 is a standardized residual covariance structure for all item pairs. If the assumption of LInd holds and if θ^* adequately represents the latent space, the item pair-correlations should be zero or, after accounting for θ^*, any residuals constitute random measurement error. One distinct advantage of Q_3 is that it takes the form of a correlation. This simplifies the interpretation of the magnitude of LID present and also allows the direction of the residual covariance to be assessed.

Using $\hat{\theta}_j$ and the item parameter estimates for a UIRT model, the examinee's expected performance on each item is computed. The expected score for items h and i for examinee j are

$$E(u_{hj}) = (U_h \mid \hat{\theta}_j) = P_{hk}(\hat{\theta}_j, \xi_h) \tag{44}$$

$$E(u_{ij}) = (U_i \mid \hat{\theta}_j) = P_{ik}(\hat{\theta}_j, \xi_i) \tag{45}$$

where $\hat{\theta}$ is an estimated latent ability of the examinee and ξ is a vector of item parameters for a given item in the UIRT model. The deviations of the scores for item h and item i for examinee j are shown in the respective equations (46) and (47) below

$$d_{hj} = u_{hj} - E(u_{hj}) \tag{46}$$

$$d_{ij} = u_{ij} - E(u_{ij}) \tag{47}$$

where u_{hj} and u_{ij} are the score of an examinee j on items h and i, $\hat{\theta}$ is the point estimate for each examinee. Yen's Q_3 pairwise index of item dependence (e.g., items h and i) then can be computed using the correlation of the residuals of the two items.

$$Q_{3_{h,i}} = r_{d_h, d_i} \tag{48}$$

When a polytomous IRT model is used, Q_3 index is computed by simply redefining the expected score function for item h and item i to be

$$E_{hj} = (U_h \mid \hat{\theta}_j) = \sum_{C=1}^{M} (C-1) P_{hC}(\hat{\theta}_j, \xi_h) \tag{49}$$

$$E_{ij} = (U_i \mid \hat{\theta}_j) = \sum_{C=1}^{M} (C-1) P_{iC}(\hat{\theta}_j, \xi_i) \tag{50}$$

where $P_{\bullet C}(\hat{\theta}_j)$ is the probability of the given item falling into category C.

Because the item responses used in calculating the correlations are also used in estimating the person's location (for example, for item i, Q_3 includes an item score explicitly in u_{ij} and implicitly in E_{ij} through the use of $\hat{\theta}_j$), Q3 is expected to be slightly negatively biased (Yen, 1984) due to part-whole contamination. When LInd is true (i.e., LInd assumption holds for all item pairs), the expected value of Q_3 (Yen, 1993, p. 198) is approximately:

$$E(Q_3) = -\frac{1}{(n-1)} \tag{51}$$

where n is the total number of items used to estimate the latent score θ^*.

Concluding Remarks on LID Measures

Critical values for flagging the existence of LID with Q_3 do not exist. Therefore, in practice, a cut point for Q_3 of $.20$ has been used for identifying items that are exhibiting LID (Yen, 1984, 1993). However, Chen and Thissen (1997) suggested that using $.20$ as the cut point for Q_3 would result in very low power for Q_3. Instead, they suggest that simulated data under LInd should be conducted to empirically determine the optimal cut points of Q_3 for a given sample size and test length. Several modifications of the Q_3 statistics such as the Fisher's r -to- z transformed Q_3 (Yen, 1984; Chen & Thissen, 1997) and the Q_3 for non-monotonic (generalized graded unfolding model (GGUM) of Roberts, Donoghue, and Laughlin (2000)) item response model (Habing, Finch, & Roberts, 2005) are also available and have been used in practice (e.g., Goodman et al., 2009).

Q_3 is an effective way to describe the presence and magnitude of LID and has been demonstrated to outperform other LID indices (Chen & Thissen, 1997; Kim et al., 2011; Zenisky, Hambleton, & Sireci, 2002). It has also been suggested that Q_3 can be generalized to address models outside of UIRT (Goodman et al., 2009). Goodman and colleagues (2009) proposed that the definition of the conditioning variable θ^*

be expanded beyond a single IRT latent trait estimate to represent any combination of variables that best represent the latent space. In this context, θ^* could be a composite trait or a vector of traits produced from several separate UIRT calibrations, from a MIRT model, or from an alternative model such as a bi-factor model or model for testlets.

In a comparative analysis of the performance based on ten indices to measure LID, Kim et al., (2011) concluded that Yen's (1984) Q_3 statistic is one of the effective indices that offers a reasonable compromise between maximum power and minimized false positive rates. The index has been widely used and may still be considered in practice due to its simplicity (Kim et al., 2011). On the other hand, they also noted that the false positive rate for G^2 statistic (Chen & Thissen, 1997) was better than that of Yen's (1984) Q_3 and that the G^2's Type I error rate was close to the significance level. However, G^2's power to detect LID was comparable with that of Q_3 only for tests with 20 items and for weak LID level, regardless of the LID percentage. As the LID level increased or as the instrument length increased, G^2's power was consistently less than Q_3. Furthermore, the researchers found that the G^2 index was particularly adversely affected by estimation problems especially for high parameter models such as the 3PL (Birnbaum, 1968) model for dichotomous items with sample sizes less than 3,000. This finding is also supported by Finch & Habing (2007) in which they found that the IRT model used may affect the performance of LID indices, although they investigated different indices (i.e.,

covariance-structure-based indices) of LID. Kim and colleagues (2011) nevertheless concluded that the G^2 statistic may also be deemed as a feasible compromise between maximum power and minimum false-positive rate, especially if 1PL/Rasch (Rasch, 1960) and 2PL (Birnbaum, 1968) models are used in item calibration (Houts & Edwards, 2013).

Another group of researchers (Houts & Edwards, 2013) also conducted a comparative analysis of the performance (i.e., power and Type I error) of Q_3 (Yen, 1984) and G^2 (Chen & Thissen, 1997) statistics in addition to the JSI (Edwards & Cai, 2008, 2011) and several other LID indices in the context of psychological assessments. The researchers explicitly concluded that, when using 2PL model (Birnbaum, 1968) and GRM (Samejima, 1969), the JSI and G^2 displayed adequate-to-good performance in most simulation conditions (i.e., scale lengths, sample sizes, number of locally dependent pairs, number of response categories for polytomous model, types of LID (Chen & Thissen, 1997), and within-LID conditions). Overall inspection of the results indicated that Q_3 performance was acceptable, its use was still appropriate and it did not cause any serious, noticeable damage.

Measurement Implications of Ignoring Local Item Dependence

Ignoring the presence of LID when using UIRT models or when the common factors of interest are not correctly specified (Edwards & Cai, 2011) will affect the psychometric properties of the test, hence jeopardizing the validity of test scores

(Lissitz & Samuelson, 2007a, 2007b), their interpretations and uses (AERA et al., 1999, 2014; Kane, 2013).

When the assumption of LInd is violated in a UIRT model, the test information and reliability are overestimated while the standard errors of the ability estimates are underestimated (Chen & Thissen, 1997; Sireci et al., 1991; Thissen et al., 1989; Wainer & Thissen, 1996; Yen, 1993). Apparently, LID is known to affect the estimation and accuracy of item parameters (e.g., Ackerman, 1987; Edwards & Cai, 2010; Oshima, 1994; Reese, 1995; Tuerlinckx & De Boeck, 2001; Wainer & Wang, 2000; Yen, 1993). Ackerman (1987) and Edwards & Cai (2010), in separate studies, found that item discriminations were overestimated (i.e., "inflated slopes" (Edwards & Cai, 2010, p. 9)) when a set of items were locally dependent. According to Edwards & Cai (2010), when one item from a pair of items that exhibit LID is removed from the analysis, the slope on the other remaining item will usually decrease slightly. Wainer and Wang (2000) found that lower asymptotes were overestimated when dependencies were ignored between testlets.

Reese (1995) observed that LID caused low scores to be underestimated and high scores to be overestimated, especially in sets of items that exhibit high LID. This effect caused the score distribution to spread out at the tails and flatten in the middle. Zenisky et al., (2002) found that the presence of LID impacted the estimation of an examinee's proficiency on a large-scale, high-stakes admission test to medical colleges. They noted that the impact was particularly noticeable on the items measuring verbal reasoning, where LIDs were most evident.

With the use of item banks in automated test assembly or computer adaptive testing, inaccurate item parameter estimates can threaten test fairness (Thompson & Pommerich, 1996). Test scaling and equating practice that depend on accurate parameter estimates can be adversely impacted when LID is detected (De Champlain, 1996; Reese & Pashley, 1999). Finally, if residual covariances differ for various population subgroups, such as the ELL and SWD, DIF results may be impacted. Methods for addressing the practical effects of LID are worthy of more investigation, for on any test (especially with context and scoring dependencies) associated item dependencies can seriously impact both the statistics used to inform test design practices and the scores that are ultimately reported to examinees.

Managing Local Item Dependence

In situations where LID is present, or likely to be present, due to contextual and/or scoring dependencies, certain courses of action are advisable to reduce the effects and magnitude of LID. The most common solution in practice is to form testlet-units (Goodman, 2008; Goodman et al., 2009; Sireci et al., 1991; Wainer & Thissen, 1996, 2001; Zenisky et al., 2002; Yen, 2002) from the related items and create one or more "super" polytomous items from the cluster by summing the individual scored objects. The resulting testlet-unit super item can then be scaled using an IRT model for polytomous data. Polytomous scoring of testlets has been demonstrated as effective in reducing LID (Goodman, 2008; Sireci et al., 1991; Stark, Chernyshenko, & Drasgow, 2002; Yen, 1993; Zenisky et al., 2002). If a test has

several related sets of items (i.e. several reading passages with related clusters), then this method is most effective if the created polytomous items can be created so that LInd is maintained across all the newly created polytomous items. However, creating polytomous items from unrelated subsets of items has been shown to decrease reliability and test information (Yen, 1993). Although dichotomous scoring may overestimate test information, polytomous scoring may underestimate test information and results in inappropriate examinee classification when pass/fail decisions are made (Keller, Swaminatahan, & Sireci, 2003).

> One potential caveat to the use of polytomous IRT models could be a trade-off in information (Thissen, et al., 1997; Yen, 1993). By summing item scores within a testlet to compute testlet scores, information regarding the specific items examinees answered correctly is lost. In addition, fewer parameters are used to model the test compared to discrete-item scoring. For example, if a 60-item test comprising ten six-item testlets were scored dichotomously using the three-parameter IRT model, 180 item parameters would be estimated. In contrast, if the test were calibrated using a polytomous model to account for the testlet structure (e.g., Samejima's (1969) graded response model), only one discrimination parameter and six threshold parameters would be estimated for each testlet (a total of 70 parameters). Thus, some measurement information may be lost when collapsing items into testlets. (Zenisky et al., 2002, p. 5)

The desirable course of action is less clear when a test and test items exhibit multidimensionality. The first strategy—deemed as the simplest and perhaps most common practice—is to continue to assume that the mixture of multiple dimension forms an essentially unidimensional measure, i.e., using UIRT as an "approximation model" (Ip, 2010, p. 397) for item responses that are considered not strictly unidimensional. A substantial number of simulation studies (Ackerman, 1989;

Ansley & Forsyth, 1989; Folk & Green, 1989; Ip, 2010; Kim, 1994; Kirisci, Hsu, & Yu, 2001; Luecht & Miller, 1992; Reckase, 1979; Spencer, 2004; Way, Ansley, & Forsyth, 1988; Yen, 1984) have been conducted to investigate the consequences in taking such an avenue. Two important findings have emerged from this body of literature. First, if there is a predominant general factor in the response data and if the other dimensions in addition to the predominant factor are small and unrelated to specific features of the items or content of the test, the manifestation of multidimensionality has little effect on item parameter estimates and the associated ability estimates. Second, if the underlying multidimensionality of the data includes strong factors in addition to the first factor, unidimensional parameterization produces item and ability parameter estimates that are "pulled" towards the strongest factor in the set of item responses in which this tendency is improved to some extent when factors are highly correlated.

For example, when the unidimensional 2PL model was used to calibrate two-dimensional item response data generated using a compensatory MIRT model, Way et al. (1988) decided that the item discrimination estimate ($\hat{a}$) values were best considered as the sum of the true values of the item discrimination for the two dimensions ($a_1 + a_2$) and that the item difficulty estimate ($\hat{b}$) values were best considered as averages of the true values of the item difficulty for the two dimensions ($\frac{b_1 + b_2}{2}$). Ansley and Forsyth (1988), on the other hand, decided that the $\hat{a}$ values from the calibration using unidimensional 3PL model were more

comparable to the average of the a_1 and a_2, $\frac{a_1 + a_2}{2}$, when fitting a non-compensatory MIRT model (Sympson, 1978) and that the $\hat{b}$ values were best considered as an overestimate of the item difficulty for the first dimension (b_1). Using both compensatory and non-compensatory (Sympson, 1978) MIRT models, Ackerman (1989), in general, found that, when the two latent ability (θ_1 and θ_2) were not highly correlated, the relationship between $\hat{\theta}$ and θ_1 was found to be stronger (c.f., Folk & Green, 1989). The researchers (Ackerman, 1989; Ansley & Forsyth, 1989; Way et al., 1988) unanimously decided that the latent ability estimates ($\hat{\theta}$) were highly related to the average of the true ability for each dimension ($\frac{\theta_1 + \theta_2}{2}$).

The resulting total-test ability estimate ($\hat{\theta}$) is considered to represent a weighted composite of the measures (Luecht & Miller, 1992) from each individual dimension. The composite ability estimate is effectively weighted according to the relative number of items linked to each trait and the average information exhibited by those items. Nevertheless, as the magnitude of multidimensionality increases, the projection of any ancillary dimensions onto a single reference composite can alter the nature of the total-test composite in unexpected ways (e.g., it might not adequately account for relationship among dimensions (Reckase & McKinley, 1983); scores tended to be related to one or the other ability instead of to a

composite (Folk & Green, 1989)) hence jeopardizing the validity of the test score and of its interpretations for the intended uses.

The second strategy often consists of two stages (see also Luecht & Miller, 1992). The first step is to determine the dimension of a test—whether empirically (e.g., Kim, 1994; Nandakumar & Stout, 1993; Roussos, Stout, & Marden, 1998; Stout, 1987; Zhang & Stout, 1999a, 1999b) or by relying on a subject matter area expert's knowledge (Crocker, Miller, & Franks, 1989; McDonald, 1981; Sireci & Geisinger, 1995)—and to thoughtfully select an appropriate MIRT model for fitting the item response data. This strategy relates to a more recently used approach: scaling sets of items assumed to represent different traits, constructs, or skill sets separately. This approach allows separate scores for each scale/dimension (i.e., subscores) to be reported (de la Torre, Song, & Hong, 2011; Haberman, 2008; Haberman & Sinharay, 2011; Sinharay, Puhan, Haberman, 2011; Yen, 1987; Wainer, Vevea, Camacho, Reeve, Rosa, Nelson, Swygert, & Thissen, 2001) and, ideally, results in ability estimates that adequately explain the responses to related sets of items. Again, this course of action is not without practical consequences (Sinharay, 2010; Sinharay, Puhan, & Haberman, 2010). Breaking the complete set of test items into separate tests *a posteriori* will result in smaller tests which in turn, produce less reliable scores. Less reliable scores may adversely affect the quality of the ability estimates. Statistical augmentation (Haberman, 2008; Wainer et al., 2001) can be used to improve the reliability of the multidimensional estimates, but not without

the risk of regression bias to the mean (Skorupski & Carvajal, 2010; Stone, Ye, Zhu, & Lane, 2010) (c.f., Sinharay, Haberman, & Wainer, 2011).

The aforementioned MIRT models can estimate multiple abilities jointly, describe the relationship between sets of traits, and allow for factorial complex structures within the test (DeMars, 2005; Md Desa, 2012; Yao, 2012; Yao & Schwarz, 2006; Yao & Boughton, 2007). These scaling methods are more computationally complex and require much larger sample sizes. Furthermore, software packages to fit these models to data tend to be limited in number and usability. Technical statistical issues such as rotational indeterminacy have also remained largely unresolved for MIRT models.

Mixed-Format Tests

The next generation CCR assessments contain a mixture of item formats that requires different item response models and scoring procedures. A summary of item formats that are used in PARCC and SBAC assessments are displayed in Tables 5 and 6 respectively. The item formats are summarized from the informational guide, high level blueprints, test specifications, and technical reports (PARCC, 2014; SBAC, 2016b) of the assessment consortia, which can be retrieved from their respective official websites. These tables can be perused together with Table 1; Figures 2 and 3; and Appendices A, B, C, and D. As can be seen in Table 5, for the math test, PARCC also categorizes the item formats according to the type of task and level of difficulty for the response mode.

Table 5. Summary of Item Formats from Partnership for Assessment of Readiness for College and Careers Consortium (PARCC) Assessments

PARCC: Math	
Task Type	Description
Type I	Conceptual understanding, fluency, and application Computer-scored only
Type II	Written arguments/justifications, Critique of reasoning, or precision in math statements Computer- & hand-scored tasks
Type III	Modeling/application in a real-world context or scenario Computer- & hand-scored tasks
Level of Response Mode	Item Format
Low	Selected Response (SR) Drag-and-Drop Hot Spot Single Numeric Entry
Moderate	Multiple Response Modes in a Single Item Graphing Tool Equation Editor Extended Responses
High	Extended Responses
PARCC: ELA	
Evidence-Based Selected Response (EBSR)	The term refers to a type of ELA/Literacy test item that asks students to show the evidence in a text that led them to a previous answer.
Prose Constructed Response (PCR)	Specific item type on the PARCC ELA/Literacy assessments in which students are required to produce written prose in response to a test prompt. These measure reading and writing claims.
Technology-Enhanced Constructed Response (TECR)	This ELA/Literacy item uses technology to capture student comprehension of texts in authentic ways that have been historically difficult to capture using current assessments. Examples include using drag and drop, cut and paste, and highlight text features.

Table 6. Summary of Item Formats from Smarter Balance Assessment Consortium (SBAC) Assessments

SBAC: Math	
Correct/Incorrect	Selected Response (SR)
Technolofy-Enhanced (TE)	Click-and-Drop, Drag-and-Drop, Equation/Numeric, Fill-in Table, Graphing, Hot Spot; Multiple Choice, Single Correct Response; Multiple Choice, Multiple Correct Responses; Short Text; Matching Tables
Extended Performance Tasks	Assessment Tasks
SBAC: ELA	
Computer Adaptive Test (CAT) (uses TE items): Machine-scored items	Multiple-Choice Single Answer (MC) Multiple-Choice Multiple Correct Answer Items (MS); Hot-Text Items (HT) also known as Select Text Items (ST) Matching Table Items Reorder Text Two-Part MC, with Evidence Responses also known as EBSR
CAT (uses TE items): Short-text items	Brief-Writes Have item-specific rubrics for scoring. (human or/and artificial intelligent (AI) scoring)
Performance Task (PT): Research items	The full-write based on 3 primary traits for Grade 8 Will be scored by subject-matter expert (SME) using a multi-trait rubric • Narrative • Explanatory • Argumentative
Performance Task (PT): Machine-scored items	Multiple-Choice Single Answer (MC) Multiple-Choice Multiple Correct Answer Items (MS) Hot-Text Items (HT)/Select Text Items (ST) Matching Table Items
Performance Task (PT): Short-text items	Brief Writes

A common hypothesis that is used concerning mixed format tests is that different item formats measure traits that are different from the traditional SR items (Traub, 1993). An item designed to measure one trait may also measure different latent traits and cognitive processes (Ackerman & Smith, 1988) and could contribute differently to item characteristics (In'nami & Koizumi, 2009; Hohensinn & Kubinger, 2011; Yen, 1984). Moreover, the choice of scale scores and score scales for such mixed-format assessments relies on meeting the needs of test users and on accomplishing certain psychometric properties of the scores, including intended score precision/reliability (Kolen & Lee, 2011; Yao & Schwarz, 2002) and score comparability with the alternate forms of a test (Kim & Kolen, 2006; Kim, Walker, & McHale, 2010; Kolen, 2006). Essentially, scores of a given assessment are important components of the test validation process (Kane, 2006) necessary to support score interpretations for the intended uses of the test (AERA et al., 1999, 2014; Kane, 2013). This section provides discussion of dimensionality issues for mixed format tests and reviews studies that deal with calibration and scoring of such tests.

Dimensionality of Mixed Format Tests

With the application of IRT methodology, it is crucial to decide whether a single dimension is sufficient to describe performance over the mixed item formats. To date, results on dimensionality concerning tests with mixed item formats are somewhat limited and not fully consistent (e.g., Ackerman & Smith, 1988; Birenbaum & Tatsuoka, 1987; Downing, Baranowski, Grosso, & Norcini, 1995;

Dudley, 2006; Haberkorn, Pohl, & Carstensen, 2016; Rodriguez, 2003; Thissen, Wainer, & Wang, 1994; Traub, 1993; Wainer & Thissen, 1993). Nonetheless, information on dimensionality is crucial, as a unidimensional scale score might lead to biased parameter estimates and incorrect inferences about examinees, when the response formats form empirically distinguishable components (Walker & Beretvas, 2003).

Studies of the dimensionality of items with different response formats have primarily been conducted for SR and CR items. Overall, there are equivocal results on the dimensionality of SR and CR item formats across the different studies. Some researchers have reported on multidimensionality in tests with SR and CR item formats (Ackerman & Smith, 1988; Birenbaum & Tatsuoka, 1987; Walker & Beretvas, 2003; Ward, Frederikson, & Carlson, 1980). Birenbaum and Tatsuoka (1987) administered SR and CR items assessing fraction arithmetic abilities for eighth grade students. A non-parametric multidimensional scaling procedure, known as smallest space analysis, was employed to examine the underlying structure for both item formats. The procedure mapped the items into points in Euclidean space and revealed considerable differences between the two formats. The underlying structure seemed more apparent in the CR where the configuration of the items in the two-dimensional space clearly indicated two clusters: one for items with equal denominators and the other for items with different denominators. The SR items, on the other hand, were dispersed, with no distinct separation between the different types of items. An EFA conducted on the inter-item

correlations, incorporating the principal factor method, also revealed similar findings in which all CR fraction items with identical denominators loaded on one factor while all CR fraction items with different denominators loaded on the second factor. The factor solution for the SR items produced a less clear distinction with half of the items not loading as expected by the researchers.

Ackerman and Smith (1988) used CFA to investigate the similarity of information provided by direct and indirect methods of writing assessment. Basing their study on the cognitive model of writing behavior, first proposed by Hayes and Flower (1980), and on the concept of PA as discussed in Chapter One of this dissertation, the researchers used CR and essay item formats to assess directly the writing construct and SR item format to assess indirectly the same construct. They concluded that scores obtained from direct and indirect methods of writing assessment provided different information. Specifically, the CFA procedures suggested that an essay task can more assess the skill of generating topic knowledge while CR items can measure the ability to organize coherent paragraphs better than SR items.

Other researchers hold opposing views, stating that SR and CR items are measuring quite the same latent traits (Bacon, 2003; Bennet, Rock, Braun, Frye, Spohrer, & Soloway, 1990; Thissen et al., 1994; Traub & Fisher, 1977; Wainer & Thissen, 1993). In a computer science test, Bennett and his colleagues (1990) found evidence of unidimensionality for three different item formats: SR, CR, and constrained CR in which CFA was used to test the fit of a three-factor model to the

item response data. The researchers concluded unidimensionality of the item formats given the highly correlated factor inter-correlations of the constrained CR items to both SR and CR items.

Traub and Fisher (1977) employed methodology that equated score scales and error variances on three item formats for verbal and quantitative measures. Two of these formats were SR and CR. Using CFA, they found little evidence of a format effect for the mathematical reasoning items and only weak evidence that the CR and SR items were measuring a different construct for verbal comprehension items.

Wainer and Thissen (1993) in their study of several weighting options to ensure efficacious reliability of mixed format test scores, argued that in many cases, the constructs measured by SR and CR items are similar enough that they can be analyzed concurrently using UIRT models. In an additional study, Thissen, Wainer, and Wang (1994) employed restricted factor analysis to examine the underlying structure of two mixed format tests from an advanced placement program. They found that the CR sections measured the same underlying proficiency as the SR sections for the majority of the test. However, they also noted a significant yet relatively small amount of LID among the CR items which resulted in a small degree of multidimensionality for each test.

Several researchers (Rodriguez, 2003; Traub, 1993) reported mixed findings on the dimensionality of SR and CR item formats. Traub (1993) reviewed a number of studies to investigate whether SR and CR items measured the same construct

across different domains. His findings suggested that the unidimensionality assumption held for SR and CR items in the test instruments assessing reading comprehension and other quantitative domains, whereas in the writing domain, the different item formats formed a multidimensional structure. Traub (1993) used a construct equivalence criterion which implies true score correlations of 1.00 to determine the dimensionality structure. In a meta-analysis, Rodriguez (2003) explored the comparability of SR and CR item formats with variations in item stem and content. Based on the definition of construct equivalence used by Traub (1993), Rodriguez concluded that the SR and CR items are measuring different constructs, although he also reported that, in certain situations, the constructs are very similar. For items that share the same stem, a high average correlation of .95 between the response formats was obtained, indicating unidimensionality. Even when the items did not share the same stem, but the content to be measured was intended to be the same, the latent correlations remained high with an average correlation of .92 .

Ward et al. (1980) compared the SR and the CR item formats in a test measuring a science subject. Even though their data were restricted and their analysis focused on correlations of the resulting scores with personality and other cognitive variables, the findings indicated that the two item formats measure different constructs. Ward (1982) concluded that, for verbal aptitude items, different item formats are essentially unidimensional in terms of both the psychometric adequacy of the resulting measures and the construct interpretations of the resulting scores.

In addition to the CR item format, studies on dimensionality of item formats have also been conducted to compare the underlying structure of SR and a special case of SR item. This item format is also known as multiple correct responses/complex SR (see Table 1) or multiple true/false (see cell 2A in Figures 2 & 3). I will refer to this item format as a complex SR (CSR) item. In many large-scale educational assessments, SR and CSR are usually scaled using unidimensional models (e.g., Organisation for Economic Co-operation and Development (OECD), 2012; Pohl & Carstensen, 2012). Although the CSR is different from SR in terms of its IC (Scalise, 2012, 2009; Scalise & Gifford, 2006) and thus might hold a different underlying structure than the SR format, several empirical studies (e.g., Downing et al., 1995; Dudley, 2006; Haberkorn, Pohl, & Carstensen, 2016) have confirmed the assumption of unidimensionality of both item formats.

Downing et al. (1995) incorporated both CSR and SR items in medical certification tests in order to examine dimensionality. Their analyses demonstrated that the scores from the two formats that were intended to assess similar construct were highly correlated, with latent correlations varying between .89 and .97. CSR appeared to primarily measure knowledge (recalling facts and basic concepts) rather than synthesis or judgment in the tests. Although the scores for CSR items were more reliable, the scores for SR items were more highly correlated to an external performance variable, supporting the criterion-related validity (AERA et al., 1999, 2014; Cronbach & Meehl, 1955). Dudley (2006) examined the concurrent validity (AERA et al., 1999, 2014; Cronbach & Meehl, 1955) of SR and CSR items in

several second language tests (including the Michigan Test of English Proficiency (University of Nevada Las Vegas, 2015)) taken by first-year undergraduate students at a Japanese university. The latent correlations between the scores from the two response formats ranged between .64 and 1.00 in vocabulary and reading, depending on the test form.

Using data from two scientific literacy tests for grades six and nine of the National Educational Panel Study (NEPS) in Germany, Haberkorn and colleagues (2016) confirmed that SR and CSR formed a unidimensional measure across content areas and studies. Results revealed that unidimensional GPCM (Muraki, 1992) fit the data better than the two-dimensional between-item model (Adams et al., 1997) for the CSR items. Moreover, the latent correlations between the two dimensions based on SR and CSR items exceeded .95. Results were cross-validated with the results from a scientific literacy test of the Program for International Student Assessment (PISA) (OECD, 2009, 2012, 2014). The researchers concluded that the assumption of unidimensionality held across all studies with SR and CSR items measuring knowledge about science and knowledge of science that require similar mental processes of recall, recognition, and evaluation.

Similarly, no item format-specifics were found by Hohensinn and Kubinger (2012) in a German language awareness achievement test administered to eight graders using three different response formats: SR, CR, and a hot text (HT) item (see Table 1). The researchers employed a special case of the Rasch model, the linear logistic test model by Fischer (1995), to examine whether different response

formats measure different latent dimensions and whether the formats could modify the difficulty of a given item. Although different response item formats did not exhibit multidimensionality, the researchers identified a distinct impact on the difficulty of the item formats. Specifically, the HT items were more difficult for the eight grade students. Hohensinn & Kubinger (2012) suspected that the results emerged due to the examinees' unfamiliarity with the new atypical format and due to the potential similarity of the solution strategies of HT and the CR as perceived by examinees.

Calibrations and Scoring of Mixed Format Test

Kolen (2006) suggested that when a test developer considers different item formats to measure different dimensions, it is possible to fit a UIRT model separately for each item format and the IRT proficiency (θ) can be calculated separately for each item format and composite formed (c.f., Luecht & Miller, 1996). Thissen, Wainer, and Wang (1994), on the other hand, suggested that, should the test developer agree that the different item formats employed in a given test are sufficiently similar and thus exhibit unidimensionality, the different item formats can be analyzed simultaneously using the UIRT models. Using appropriate software, the dichotomous SR may be estimated with a 3PL model (Birnbaum, 1968) and the polytomous items such as CR, TE and (computerized) PA can be estimated with a GPCM (Muraki, 1962). After the estimation of item parameters, estimation of proficiency can be conducted using the maximum likelihood or Bayesian methods.

Their approach was later implemented by Bukhari et al. (2016), Ercikan, Schwarz, Julian, Burket, Weber, and Link (1998), Lorie (2014), Rosa, Swygert, Nelson, and Thissen (2001), Sykes and Yen (2000), and Thissen, Nelson, & Swygert (2001).

Rosa et al. (2001) developed an alternative UIRT method for scoring the mixed format tests. In this method, a hybrid of IRT response pattern scoring and IRT summed score scoring is calculated for each item type to produce scale scores based on patterns of summed scores. IRT proficiency is estimated from these summed scores using Bayesian methods. The result of this method was that the vexing weighting problem associated with mixed item formats was implicitly solved and a new system of (implicit) optimal weights was used to score the test. Rosa et al. (2001) suggested that this procedure is preferable to typical pattern scoring "both to implement and to explain to consumers" (p. 255). In general, however, the weighting of each item format still depends on the extent that the item format discriminates near an examinee's proficiency. Sykes and Hou (2003) used various weighting schemes and then evaluated the psychometric properties using UIRT.

Thissen et al. (2001) introduced a system that approximated Rosa's (2001) patterns of summed scores scoring with weighted linear combinations of the scores on each item format component. In Thissen's et al. (2001) method, the scoring weights were visible in the solution in which, for a given mixed format test, each section of item format was given a score and a weight. Using the weights, the two scores were then combined into the total score. This approximation method combined basic IRT and some concepts of the traditional test theory.

Using a set of third grade tests in reading, language, math, and science, Ercikan and colleagues (1998) demonstrated the construction of a common score scale by combining scores from SR and CR item formats. The dichotomous SR items were calibrated using a 3PL model (Birnbaum, 1968) and the polytomous CR items (scored by raters) were calibrated using a two-parameter partial credit (2PPC) model (Yen, 1993), a model which is similar to Muraki's (1992) GPCM with slightly different parameterizations. The examination of the tests indicated that SR and CR items assessed constructs that were sufficiently similar to allow the creation of a common scale and provide a single set of scores for responses to both item formats. An examination of item information provided by concurrent calibrations and separate calibrations indicated that concurrent calibrations led to loss of information for CR items. However, Ercikan et al. (1998) noted that in most tests, the differences in information were negligible, and that all the large differences were due to LID. Results regarding the differences in difficulty, LID, and low reliabilities of short CR tests provided support for combining scores from the two item formats. The researchers concluded that increasing test length by combining the two item formats naturally increases overall measurement accuracy. In addition, combining the two item formats enhances the reliability of the test since these items, despite their different formats, tended to measure the same test construct.

On the other hand, Ercikan and colleagues (1998) reported slightly different findings from Goodman's (2008) study. Basing his study in the context of a mixed format certification/licensure exam that employed SR and computerized PA,

Goodman (2008) discovered that treating computerized PA and SR items as two separate and distinct scales was effective in controlling the amount of LID. This effectiveness lessened as the two item formats became moderately or highly associated. Thus, Goodman (2008) concluded that, when the correlations between the item formats were moderate or high, the amount of LID from simultaneous calibration is similar to the amount of LID from separate calibration. In fact, as the two formats measured more similar construct, concurrent calibration produced score estimates that are more precise.

Bukhari et al. (2016) examined the amount of IRT information provided by different TE and SR items for several interim CCR assessments in ELA and math. In this study, three combinations of IRT models were employed in which all dichotomous and polytomous items were concurrently calibrated: (1) 3PL & 2PPC; (2) 2PL & 2PPC; and (3) 1PL/Rasch & 1PPC. Their findings indicated that the 1PL/Rasch model did not fit well for the TE items utilized within the assessments. The researchers suggested that the lack of fit was most likely due to the fact that the item discriminations varied across the different TE item formats. Similarly, Lorie (2014) calibrated the items for a CCR end-of course examination in Algebra and English using a combination of a 2PL/3PL model and PCM (Masters, 1982). Although Lorie (2014) and Bukhari et al. (2016) focused more on the effect of the scoring procedures (i.e., dichotomous, componential, and polytomous) employed for different TE item formats and did not examine the effect of concurrent versus

separate calibrations, they proved that the concurrent calibration of different item formats for the new CCR assessments is widely used in practice.

In the studies of Ercikan et al. (1998) and Bukhari et al. (2016), the item parameters were estimated using a proprietary program called PARDUX (Burket, 1991, 2010). Lorie (2014) used PARSCALE (Muraki & Bock, 1991) to conduct item calibrations. Both PARDUX and PARSCALE estimate item parameters using marginal maximum likelihood procedures implemented with the EM algorithm (Bock & Aitkin, 1981).

Potential Sources of Construct-Irrelevant Variance in Scores Reporting

The Assessments Peer Review Guidance promulgated by the US Department of Education (2009) required "strong correlations of test and item scores with relevant measures of academic achievement and weak correlation with irrelevant characteristics, such as demographics" (p. 42). This requirement is in line with the professional guidelines stated by the *Standards* (AERA et al., 2014) and the ITC (ITC, 2013a, 2013b). The outcomes of assessment are often confounded with nuisance variables that are not related to the construct being measured. These extraneous variables come from many different sources. The variability of assessment outcomes due to these contaminants is referred to as construct-irrelevant variance (CIV). As mentioned previously in Chapter One, CIV is one of the major threats to fair and valid interpretation of test score (AERA, et al., 2014; Haladyna et al., 2004; ITC, 2005a; Messick, 1989, 1994). Construct-irrelevance refers to the degree to which

measurement of examinees' characteristics are affected by factors irrelevant to the construct being measured.

Several CIVs that may result from test content, test context, test response, and different learning opportunities for examinees can threaten the fairness as well as valid interpretations of test scores for the intended uses of the tests (AERA et al., 2014). For examples, CIV based on test context "may result from a lack of clarity in test instructions, from unrelated complexity or language demands in test tasks, and/or from other characteristics of test items that are unrelated to the construct but lead some individuals to respond in particular ways." (p. 55). With regard to test response, CIV "may arise because test items elicit varieties of responses other than those intended or because items can be solved in ways that were not intended. To the extent that such responses are more typical of some subgroups than others, biased score interpretations may result." (p. 56). In this study, I consider two factors that may to some extent contribute to CIV in the constructs of interests hence contaminating scores of the new generation assessments: employment of different item/response formats and unnecessary linguistic complexity.

Different Item Formats with Technology

Downing (2006a) delineated the item development process as the fourth out of twelve steps to ensure effective test development. According to Downing, the selection of item format is a major source of validity evidence for the test. Therefore, a clear rationale for item format selection is required. In practice, the selection of

item format "may quite legitimately rest largely on pragmatic reasons and issues of feasibility" (Downing, 2006b, p. 11). Popp et al. (2016) explored issues that test developers should consider when determining which computerized item format (i.e., text, video, animation) to adopt for situational judgment tests. They proposed a framework to judge the appropriateness of the various item formats based on four considerations: psychometric, applied, contextual, and logistics. The psychometric consideration concerns issues related to reliability of scores, validation efforts, impact of non-relevant constructs, and biasing of responses. The applied consideration involves: construct-format matching; the examinees' perspective, engagement, and distraction; accommodations; face validity; availability of data for scoring; content development, modifications, and expansion; and, test security. The third consideration imperative in test development and selection of item formats is based on contextual factors. These include: diversity of representation; organizational image; and examinees' expectations, access to, and familiarity with technology and target devices. Finally, Popp et al. (2016) include a logistic consideration which involves content writer experience and availability, and issues specific to the production of the technology-enabled item format that incorporates animation (e.g., securing talent, recording session, and cost of production).

Similarly, in a feasibility review of 25 computerized PA item formats that were operationally used in a certification test for accountants (i.e., the Certified Public Accountant exam), Zenisky & Sireci (2001) employed two important criteria to evaluate the item formats: psychometric and operational. Using the psychometric

criterion, an item format should represent the construct of interest, avoid CIV, meet the consequential validity requirement (see Messick, 1994), and result in adequate reliability of scores. Under the operational criterion, factors such as cost, face validity, implementation, scoring, test security, and the implications of training and tutorial for examinees should be carefully scrutinized.

Sireci (2016), in his commentary on several chapters about the use of technology to enhance assessments in Drasgow (2016), urged researchers to provide sufficient focuses on the five sources of validity evidence — test content, internal structure, response processes, consequences, and external variables — valued by the *Standards* (AERA, et al., 2014). He notes: "Unfortunately, only validity evidence based on test content was covered … A much more powerful evaluation of technological enhancements on construct representation would involve cognitive labs, dimensionality studies, and criterion-related validity studies." (p. 107).

Indeed, there is limited research on the possibility that computerized item formats might introduce CIV (Haladyna & Downing 2004, Huff & Sireci, 2001; Lakin, 2014; Sireci, 2016). Several early studies that examined the CBT's possibility to introduce CIV focused on the effect of examinees' computer experience on their test performance (Johnson & White, 1980; Lee, 1986; Mazzeo, Druesne, Raffeld, Checkettes, & Muhlstein, 1991; Powers & O'Neil, 1993). However, these studies have limitations in terms of small sample sizes, questionable measures of computer familiarity, and the utilization of only traditional SR items.

Taylor et al. (1999), in an attempt to address the limitations of previous studies, provided a comprehensive example of how a study to examine CIV in item formats could be conducted. They examined the relationship between computer familiarity and examinees performance on a set of 60 items from the computer-based Test of English as a Foreign Language (TOEFL). The item formats employed in the three sections of the TOEFL test (i.e., listening comprehension, reading comprehension, and structure and written expression) vary between drag-and-drop, multiple-correct-responses, matching, click-within-a-passage, insert-sentence-in-a-passage, and short/extended-essay.

The study was conducted in two phases. In phase one, a questionnaire was developed and administered to 90,000 examinees worldwide to assess their access to and familiarity with computers and to distinguish between examinees with high, moderate, and low computer familiarity levels. Findings "revealed small differences in computer familiarity by age and gender; however, differences were more pronounced among groups defined by native language and native region" (p. 265). A group of 1,200 examinees, identified as low-computer-familiarity and high-computer-familiarity groups from phase one, were selected for the second phase of the study. These examinees were carefully selected to ensure the comparability of their background variables with examinees from phase one. In the second phase, the 1,200 examinees' performance on the CBT TOEFL tasks was examined after they received a specially-designed computer-based tutorial. The authors used analysis of covariance (ANCOVA) for this purpose in which examinees' English ability (i.e., the

initial scores of paper-based test (PBT) version of TOEFL) was used as a covariate. Initial findings indicated that, with no adjustment for language ability, "examinees who were familiar with computers had significantly higher TOEFL test scores and CBT scores than those who were not" (p. 266). The researchers speculated that TOEFL examinees with high levels of computer familiarity in general have more opportunities for language and computer instruction and use. Nevertheless, there was no evidence of an adverse relationship between computer familiarity and computer-based TOEFL test performance due to lack of prior computer experience after administering the computer tutorial and controlling for their English ability level.

The correlation between the total scores on CBT items after the tutorial administration and the examinees' ratings on computer familiarity was .20. The correlations between scores from the three subsections—listening comprehension, reading comprehension, and structure and written expression—with the ratings on computer familiarity were .20, .16, and .15, respectively. The correlations between the PBT TOEFL total scores (the covariate) and the CBT TOEFL total scores were .84 while the correlations between their section scores were .74, .72, and .74 respectively.

Interfering Linguistic Complexity

Linguistic complexity for test item format refers to the level of language used in question stems and responses. High levels of linguistic complexity in test items generally consist of difficult vocabulary, less frequent words, multiple-meaning words, lengthy words and sentences, long question phrases, long noun clauses, subordinate clauses, comparative structures, embedded clauses, passive sentence structures, prepositional phrases, sentence and discourse structure, conditional clauses, negation, concrete versus abstract or impersonal presentations, and other features difficult for ELL students (Abedi, 2006; Abedi & Linquanti, 2012).

Researchers have investigated the relationship between specific types of linguistic features of items that contribute to linguistic complexity non-central to the construct of interests. The impact of such language factor has resulted in increased test difficulty for ELL students from different K-12 grade levels in math (e.g., Abedi & Lord, 2001; Abedi et al., 1997; Martiniello, 2009; Sato et al., 2010; Shaftel et al., 2006), science (Abedi, Courtney, Mirocha, Leon, & Goldberg, 2005; Abedi et al., 2000; Chang, 2013; Wolf & Leon, 2009) (c.f., Avenia-Tapper & Llosa, 2015; Lee, Quinn, & Valdes, 2013), and various content areas (Abedi, Leon, & Mirocha, 2003). In ELA (e.g., reading) and ELP assessments, language is so inherent to the focal construct that the concept of unnecessary linguistic complexity may not apply (see PARCC, 2016a). Nevertheless, Abedi & Linquanti (2012) stress that, even in these areas, excessive linguistic complexity can still be avoided (see Abedi et al., 2003).

Math test items can be linguistically modified to provide accommodation by reducing the complexity of the language used without altering the construct being assessed. The content task and content terminology are retained but the language is simplified to make the items more accessible to students, particularly the ELL subgroups (Abedi & Lord, 2001; Abedi et al., 1997; Martiniello, 2009; Sato et al., 2010; Shaftel et al., 2006).

Sato and colleagues (2010) examined the linguistic complexity of math items and compared the performance of seventh- and eighth grade students across levels of English proficiency (ranging from not proficient to proficient English-language users). Specifically, they were interested on the effect of linguistic modification on students' performance on two sets of math items (original and linguistically modified) across three subgroups of students: ELL students, non-ELL students who were not ELA-proficient students, and non-ELL students who were ELA-proficient. Their findings indicated a consistent trend of better performance on math content test items across all groups with lower linguistic complexity than items with higher linguistic complexity. This difference was most striking for ELL students, compared with students who were no longer considered ELL but were not yet fully proficient in ELA or students who were fully English proficient. Reduction in linguistic complexity appears to give a specific benefit to students who are not yet proficient in English but has no impact on those who are fully proficient.

Martiniello (2009) compared the grade four math performance for both ELL and non-ELL students. The math test assesses five major learning strands: (1)

number sense and operations; (2) patterns, relations, and algebra; (3) geometry; (4) measurement; and (5) data analysis, statistics, and probabilities. It consists of a mixture of SR and CR math word problem items of varying linguistic complexity, pictorial support, and schematic support. According to Martiniello,

> [p]ictorial [supports] include concrete images (Presmeg, 1986), sometimes called mental pictures (Andersen, as cited in Johnson, 1987), which depict details of objects described in the math problem (Hegarty & Kozhevnikov, 1999). Schematic [supports] are more abstract than pictorial images. They are more abstract than pictorial images. They are meaning structures representing several elements of parts (i.e., objects, people, events) and their pattern of connections and relationships (i.e., causal, part-whole, temporal sequence relationships) (Johnson, 1987). (p. 166)

The findings supported Martiniello's (2009) hypothesis in which she proposed that items with greater grammatical and lexical complexity were more difficult for ELL students compared to their non-ELL peers. However, her findings also revealed that the inclusion of items which provide nonlinguistic schematic support helped the ELL students to make meaning of the text and thus mitigated the negative effect of increased linguistic complexity in math word items.

Abedi and Lord (2001), using a large math test in the US (i.e., National Assessment of Educational Progress (NAEP)) compared ten original items from the test with items for which, similar to Sato et al. (2010), the content task and terminology were retained but the language was simplified. Their findings revealed small but significant improvements in the scores of 1,031 out of 1,174 grade eight students in low- and average- level math classes using the linguistically modified

test items. Among the linguistic features that contributed to the discrepancies were passive voice verb and low-frequency vocabulary. Data from reading aloud interview identified a student who changed the difficult-to-process passive voice form (*would be expected*) into its active verb form (*would you expect to find*) which was more familiar to that student.

General academic vocabularies are those words that are not among the 2,000 most common words in a language (Coxhead, 2000). "[These] words (e.g., *substitute*, *underlie*, *establish*, *inherent*) are not highly salient in academic texts as they are supportive of but not central to the topics of the texts in which they occur" (p. 214). On the other hand, words that are encountered more frequently (i.e., the New General Service List by Browne, Culligan, & Phillips, 2016) are likely to be familiar to most readers. "Readers who encounter a familiar word are likely to interpret it quickly and correctly, spending less cognitive energy analyzing its phonological component (Adams, 1990; Chall, Jacobs, & Baldwin, 1990; Gathercole & Baddeley, 1993)" (Abedi, 2006, p. 385). A more recent mixed-method study by Chang (2013) confirmed this by revealing that students, especially ELL students, found difficult vocabulary as one of the greatest challenges to comprehending science passages because ELL students' knowledge of academic vocabulary was significantly lower compared to the fluent-English speaking students.

Modifying sentence length can also make a difference in students' abilities to comprehend content in math tests. When the items from the large scale math test used by Abedi et al. (1997) were grouped into long and short items, Abedi and Lord

(2001) found that the eighth-grade ELL students performed significantly lower on the longer test items regardless of the items' level of content difficulty. Item length was measured as number of sentence lines in the stem and answer choices. Short item consisted of one line and long item was an item with two or more lines. Abedi's results further suggested that ELL students had higher proportions of omitted/not-reached items and had more difficulty with the items that were identified by content and language experts to be linguistically complex.

Abedi and colleagues (2000) compared performance of 422 eighth grade students on the NAEP science test with test accommodations. Students answered 20 science items in three test formats: (1) one booklet in original form (no accommodation), (2) one booklet with English glosses and Spanish translations in the margins, and (3) one booklet with customized English dictionary (contained only "noncontent" (see Abedi, 2006, p. 382) words that appeared in the test items). Findings revealed that ELL students scored the highest when accommodated with customized dictionary. That is, when their language needs are addressed.

English Language Proficiency and Content Performance for K-12 English Language Leaner Students

Given the previous discussion of the ways in which different linguistic features could contribute to unnecessary linguistic complexity in content assessments and how different accommodations and linguistic modifications have helped reduce the achievement gap (especially for ELL students), it is of paramount

importance to have a better understanding of the relationship of students' ELP and content assessment performance, particularly for the ELL subgroups. Through a representative sample of studies conducted between years 1952 to 1968, Aiken (1972) summarized the partial correlation between language and math content scores after controlling for general intelligence for K-12 students, which ranged from .45 to .55. Secada's (1992, as cited in Chen, 2010, p. 15) summary showed the correlation between language and math content ranged from .20 to .50. All of the correlations were positive and statistically significant. The results of these studies confirmed that there is a non-negligible relationship between language and math achievement.

Reports from the literature that language proficiency and content performance are related are further substantiated with the conceptualization of NCLB. Using the assessment from an educational consortium of the US state departments of education, the World-Class Instructional Design and Assessment (WIDA) consortium, Parker, Louie, and O'Dwyer (2009) employed hierarchical linear modeling (HLM) and samples of ELL students from grades five and eight classes nested within three US states to examine the relationships between language domain scores (in reading, writing, listening and speaking) and academic content scores (in reading, writing and math). They found that, after accounting for the variance explained by student and school level covariates, the four English domain scores explained between 14% and 21% of the variance in content scores for the eighth-grade sample, and between 21% and 30% of the variance in the fifth-grade

sample. At the English domain level, Parker and colleagues also found that written language scores (reading and writing) were significant predictors of all three types of content assessment performance for both samples in all three states. Oral language scores (listening and speaking) predicted content outcomes for some grade levels and content areas, but in all cases these relationships were significantly less strong than those observed for written language.

Similarly, Kim and Herman (2008) used HLM to model the relationships between language (reading, writing, listening, and speaking) and content domains (ELA, math, and science) for fourth to eighth grade students. They found strong, significant relationships between ELP domain scores and content assessment scores for all grades in three US states, with no significant variance in these relationships across schools within these states. In other words, a positive and strong relationship remained fairly consistent in different settings. Essentially, the study findings also provided empirical evidence for a strong relationship between ELP and content performance for the ELL students. Furthermore, the researchers discovered a quadratic, instead of linear, relationship between the ELP scores and content assessment scores suggesting that ELL students at higher ELP tended to have an increased likelihood of higher content scores.

Using a quantile regression analysis, Chen (2010) explored two significant areas of research. First, she examined a changing relationship between students' ELP and math achievement and found that language proficiency (operationalized by reading ability) positively affected math achievement differently across all quantiles

(at different math ability levels) at all grades. At grade one, the correlation between students' reading ability and math scores was .65, but increased to .73 for eighth grade students. Second, she modeled math growth longitudinally, after accounting for language proficiency, using students' data at four different points to detect the long term math achievement gap between three subgroups: (1) ELL students, (2) former ELL students, and (3) non-ELL students. Her results suggested that language demand in tests may have contributed to the large achievement gap between the ELL and non-ELL students.

While prior studies generally support a strong, significant relationship between ELL students' ELP and content assessment performances, they did not address the construct being measured in ELP assessments. To better understand these studies' findings related to the relationship between ELP and content assessments, it is essential to understand the constructs of ELP assessments. Wolf and Faulkner-Bond (2016) recently conducted a validation study of the test content and of three large scale ELP assessments for ELL students across three different US states. In addition to test content, the researchers investigated the relation of the ELP assessment scores to scores on the states' content assessments.

Specifically, the content validation, conducted by trained ELP content analysis raters, examined the types and degree of academic language items in the sampled ELP assessments. The raters identified three specific constructs across the four language domains from the items in the three ELP assessments: (1) social language, (2) general academic language, and (3) technical academic language. They

also observed "variation across the three states' ELP assessments with respect to the analytic rating of academic language characteristics such as academic words and syntactic complexity." (p. 12). The complex syntactic features included occurrences of passive voice, relative clauses, and/or conditional structures. The number of academic vocabulary words and syntactic features determined the linguistic complexity of a given item.

Next, the researchers employed HLM analysis to provide "empirical evidence of the importance of academic language proficiency by examining its prediction patterns for content assessments delivered in English." (p. 7). Their findings supported their hypotheses that highly technical academic language would have a stronger relationship with technical subject matter such as math than general academic proficiency. The partial correlations among content scores and language domain scores from their study are presented in Table 7, and represent the relationship between the row and column variables after controlling for all other variables in the table.

Table 7. Partial Correlations among Language Domain & Content Scores from Wolf & Faulkner-Bond (2016) Study

	State A *df*=5503					State B *df*=2668					State C *df*=3553				
	Social L	General AL	Technical AL	ELA	Math	Social L	General AL	Technical AL	ELA	Math	Social L	General AL	Technical AL	ELA	Math
Social L	1.00					1.00					1.00				
General AL	.51	1.00				.17	1.00				.27	1.00			
Technical AL	.21	.22	1.00			.59	.15	1.00			.39	.20	1.00		
ELA	.37	.12	.07	1.00		.15	.14	.02	1.00		.12	.40	.10	1.00	
Math	.07	.07	.18	.36	1.00	.06	.04	.23	.47	1.00	-.10	.14	.23	.30	1.00

Note. df: degree of freedom, L: language, AL: academic language, ELA: English Language Arts, Math: mathematics

Summary in the Context of Current Research

Testing and measurement in education inherently requires a group of items to operationalize and quantify constructs of interest that are often intricate but sufficiently unambiguous and fungible on their own. Traditional test theory has limitations regarding multicomponent and complex test designs as well as item formats. As a result, IRT has become the contemporary tool of choice for measurement, and, to a certain extent, for explanation in educational testing. Due to its simplicity (i.e., parsimony) and practical mathematical models, UIRT has been predominantly used across educational (and psychological) research.

Nonetheless, the various assumptions of UIRT have made its application to multicomponent and complex test designs and formats somewhat limited. UIRT models assume that each item within a test measures the same unidimensional construct and that item responses, given the latent construct, are locally independent. With the rapid development of the new generation educational assessments stipulated by the CCSS and the NGSS that are administered online (i.e., CBT) using mixed format items, a test that is purely unidimensional is no longer feasible and attractive. Thus, one could argue that analysis using traditional UIRT models is similarly no longer feasible or practical.

Relatedly, the LInd assumption has also been found to be too stringent in many testing situations (Chen & Thissen, 1997; Yen, 1984, 1993). Motivated by the inherent complexity of testing and by the advancement of computational psychometric software, IRT has been expanded to a multidimensional context in

which methods for fitting MIRT models to item response data have become better developed (Reckase, 2009).

LID complicates analyses and inferences given the confounding effect of the interactions between examinees from the diverse subpopulations and (not-so-) obvious heterogeneous characteristics of the assessment items and tasks. The capability to detect LID, especially with the multidimensionality and complexity of the new assessments, often surpasses our ability to understand it. Simple statistical methods can help test for the presence of LID. However, conclusively determining the source or sources of LID—whether from contextual factors, scoring procedures, or the underlying dimensionality—is not straightforward.

A critical first step is to make a clear distinction between "valid", intended, primary multidimensionality and potential "nuisance" multidimensionality. Valid dimensionality supports the test purpose, where content and item design properly align items with the intended factors. Nuisance dimensionality could result from unintended traits such as linguistic complexity or from method variance sources such as item/response formats. Linguistic complexity, new TE items, and other item formats may therefore support the valid dimensionality or may improperly influence the response data.

It is becoming more common to include dichotomous and polytomous complex PA and TE items on many types of tests, especially in the next generation educational assessments used to assess examinees' readiness to enter and perform in college and workplace. Empirical research is limited in regards to whether these

skills align with the assessment purpose, content, and intended uses. Similarly, the use of different technologically enhanced item formats for different subgroups of examinee populations has not been adequately studied.

With the effort to include all students in the educational assessment system, it is important to ensure that linguistic complexity, when present, does not interfere with the construct of interest and contaminate test scores of examinees and does not result in improper use of and negative consequences due to the test scores. Nonetheless, the linguistic complexity in the texts and the tasks necessitated by the standards of the assessments have often been found to also impede the performance of examinees especially the ELLs and SWDs.

Table 8 provides the summary of the potential issues relevant to the CIV concept in educational assessment based on what I have reviewed thus far. Table 9 provides the summary of the relevant literature as context for my "simulation study-in-context".

Table 8. Summary of Potential Issues in Educational Assessments (Not Limited to Next Generation Assessments)

Topic		Summary	References
Validity Issues on Technology-Enhanced (TE) Item Formats	Potential & Threats of Computer Based-Testing (CBT) on Validity	• Test anxiety. • Test-wiseness and guessing related to SR items. • Different test formats. • Examinees' familiarity with technology that maybe associated with social-class differences	Haladyna & Downing, 2004; Huff & Sireci, 2001; Lakin, 2014; Rabinowitz & Brandt, 2001; Randall, Sireci, Li, & Kaira, 2012; Sireci & Zenisky, 2006; Zenisky & Sireci, 2002; Zenisky & Sireci, 2006
Construct-Irrelevant Variance (CIV)	Computer Familiarity as a Construct Irrelevant Factor	Effects of examinees' computer experience on CBT performance	Johnson & White, 1980; Lee, 1986; Mazzeo, Druesne, Raffeld, Checkettes, & Muhlstein, 1991; Powers & O'Neil, 1993; Taylor, Kirsch, Eignor, & Jamieson, 1999
Impact of other Non-Relevant Construct	Linguistic Complexity	The impact of linguistic complexity has resulted in increased test difficulty for English Language Learner (ELL) students from different K-12 grade levels in math, science, and various content areas	Abedi, 2006; Abedi & Lord, 2001; Abedi, Lord, & Plummer, 1997; Martiniello, 2009; Sato, Rabinowitz, Gallagher, & Huang, 2010; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006; Abedi, Courtney, Mirocha, Leon, & Goldberg, 2005; Abedi, Courtney, & Goldberg, 2000; Chang, 2013; Wolf & Leon, 2009; Abedi, Leon, & Mirocha, 2003

Table 9. Summary of the Relevant Literature to Provide Context for Simulation Study

Topic	Summary	References
Dimensionality of Mixed-Format Assessments	• Limited research on dimensionality of most item formats • Many researchers concluded that Selected Response (SR) and Constructed Response (CR) items are unidimensional when use to assess the same construct • Same goes to SR and Complex SR (CSR) • Different item format introduced different levels of difficulty especially the ones that are atypical	Bacon, 2003; Bennet, Rock, Braun, Frye, Spohrer, & Soloway, 1990; Thissen, Wainer, & Wang, 1994; Traub & Fisher, 1977; Wainer & Thissen, 1993; Downing, Baranowski, Grosso, & Norcini, 1995; Dudley, 2006; Haberkorn, Pohl, & Carstensen, 2016; Hohensinn & Kubinger, 2012 (see also Taylor et al., 1999)
Calibration of TE Items	• If different item formats measure different dimensions, it is possible to fit a UIRT model separately for each item format • If item formats measure the same construct, simultaneous calibration can be used	Kolen, 2006; Bukhari et al., 2016; Ercikan, Schwarz, Julian, Burket, Weber, and Link, 1998; Lorie, 2014; Rosa, Swygert, Nelson, and Thissen, 2001; Sykes and Yen, 2000; Thissen, Nelson, & Swygert, 2001
Sources of Local Item Dependence (LID)	• Underlying & surface LID • Incorrect specification of the amount of common factors • Contextual, dimensional, scoring	Chen & Thissen, 1997; Edwards & Cai, 2011; Goodman, 2008; Goodman, Luecht, & Zhang, 2009; Haladyna, 1992; Yen, 1984, 1993

Measurement Implications of Ignoring (LID)	• LID overestimates
	  • test information
	  • reliability
	  • high scores
	• LID underestimates
	  • the standard errors of the ability estimates
	  • low scores
	• LID affects the estimation and accuracy of item parameters
		Chen & Thissen, 1997; Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wainer & Thissen, 1996; Yen, 1993; Ackerman, 1987; Edwards & Cai, 2010; Oshima, 1994; Reese, 1995; Tuerlinckx & De Boeck, 2001; Wainer & Wang, 2000; Yen, 1993; Reese, 1995

Up to this point, I have described the research problem in the context of next generation assessment, provided the gap in (IRT) model-based, computer simulation study, and presented my research questions for my current simulation study. At the same time, I have substantively reviewed the literature to explicate the IRT models along with the assumption of LInd and LID, discuss potential complexities and issues in the educational measurement, the educational assessments in general, and the next generation K-12 assessments in particular.

In my first research question, I want to compare the performance of the UIRT and the MIRT models used to calibrate complex test items for which I will generate the simulated, examinees' item response data for the items with six different test configurations. Specifically, I will generate tests that vary based on sample sizes of examinees, correlations between four primary dimensions, number of items per dimension, and discrimination levels of the primary dimensions. I will also manipulate two of the configurations so that there will be some contaminating factor(s) in addition to the factors of interest. To re-state this, I decided to introduce some local dependency among the items with the presence of construct-irrelevant factor(s) that will be affecting the items. In my second research question, I am interested in examining the structure of the residual covariance (given my six test configurations) with LID explicitly modeled by the nuisance factor(s). As I stated previously, I have situated my research on nuisance factors within the context of the next generation K-12 assessments which include ELL and SWD populations and which incorporate the innovative and computerized item/response formats. This is

my modest attempt to conduct what I describe as a "simulation study-in context" (cf. Bachman & Palmer, 1996; Chalhoub-Deville 2003; Chalhoub-Deville & Deville, 2006; Luecht & Ackerman, 2017; Snow, 1994).

A remaining question is the method by which I will compare the multidimensional (person and item parameter) estimates from the MIRT calibration model, which are vector-valued, to the parameter estimates from the relative UIRT model in order to answer my first research question. As stated briefly earlier in Chapter One, I will incorporate a new technique developed by Luecht and Ackerman (2017) that employs the expected response function (ERF) approach—which is based on the expected raw scores (ERS)—to compare different components of residuals and errors in the test across different calibration models. Detailed procedures of the approach and the simulation study will be delineated in Chapter Three.

CHAPTER III

METHODS

Simulation studies allow researchers to answer specific questions about data analysis, statistical power, and best practices for obtaining accurate results in empirical research. Such studies also enable any number of experimental conditions that may not be readily observable in real testing situations to be tested and carefully controlled. Moreover, simulation enables researchers to replicate study conditions easily and consistently that would have been very expensive when conducted with live subjects. Although it is acknowledged that a simulation of educational testing situations will never accurately feature the true complexity and inherent context of real data (Luecht & Ackerman, 2017) and therefore does not permit for conclusive conclusions, simulations are useful for framing general patterns and trends of a limited selection of phenomena of interest. I will use a computer simulation to address the research questions presented in Chapter One based on the substantive literature review in Chapter Two to conduct a "simulation study-in context". I will delineate all the methods involved in this section.

Constant of Study

I will use a constant number of primary dimensions for data generation. I will use four primary dimensions to represent the intended constructs of the assessments which is in line with the K-12 CCR assessments for ELA and math, in which, in general, each content domain consists of four subdomains (see Multidimensionality of the Intended Assessment Constructs for the Next Generation Assessments in Chapter One). Most simulation studies in educational testing, especially on subscores (e.g., Haberman & Sinharay, 2010; Yao & Boughton, 2007), tend to employ three to four dimensions (as used in operational setting).

Conditions of Study

I will use two sets of conditions to generate the data. The first set is based on six different test configurations. This set is divided into two separate sets of crossed-condition data (i.e., test formats) based on: (1) Test Format 1: the presence of nuisance dimension (zero nuisance dimension versus one nuisance dimension); and (2) Test Format 2: the presence of two nuisance dimensions with different correlations between them. For the purpose of computatitonal simplicity, I will replicate all conditions 10 times. Table 10 display the combination of simulation designs for the first and second test formats. I will then apply a second set of the study condition (i.e., different IRT calibration strategies) to the generated data sets. I describe the conditions which I will use to generate response data for the simulation in the next two subsections.

Table 10. Complete Simulation Design

Number of nuisance dimension	disc levels	vcorr	ncorr	Sample size = 1,000		Sample size = 5,000	
				Number of items per dimensions		Number of items per dimensions	
				10	20	10	20
0	low	0.40	NA	10 iterations	10 iterations	10 iterations	10 iterations
		0.80		10 iterations	10 iterations	10 iterations	10 iterations
	high	0.40		10 iterations	10 iterations	10 iterations	10 iterations
		0.80		10 iterations	10 iterations	10 iterations	10 iterations
1	low	0.40	NA	10 iterations	10 iterations	10 iterations	10 iterations
		0.80		10 iterations	10 iterations	10 iterations	10 iterations
	high	0.40		10 iterations	10 iterations	10 iterations	10 iterations
		0.80		10 iterations	10 iterations	10 iterations	10 iterations
2	low	0.40	0.00	10 iterations	10 iterations	10 iterations	10 iterations
			0.40	10 iterations	10 iterations	10 iterations	10 iterations
			0.70	10 iterations	10 iterations	10 iterations	10 iterations
		0.80	0.00	10 iterations	10 iterations	10 iterations	10 iterations
			0.40	10 iterations	10 iterations	10 iterations	10 iterations
			0.70	10 iterations	10 iterations	10 iterations	10 iterations
	high	0.40	0.00	10 iterations	10 iterations	10 iterations	10 iterations
			0.40	10 iterations	10 iterations	10 iterations	10 iterations
			0.70	10 iterations	10 iterations	10 iterations	10 iterations
		0.80	0.00	10 iterations	10 iterations	10 iterations	10 iterations
			0.40	10 iterations	10 iterations	10 iterations	10 iterations
			0.70	10 iterations	10 iterations	10 iterations	10 iterations

Note. disc levels: item discrimination levels for all primary dimensions; vcorr: correlations between primary dimensions; ncorr: correlation between nuisance dimensions; NA: not applicable

Sample Sizes

I chose two sample sizes (1,000 and 5,000 examinees) to represent the upper and lower ends of the typical number of examinees that might be administered a single form of a large-scale mixed-format CCR assessment. Employing 1,000 examinees as the lower bound of my sample size condition is consistent with the meaningful minimum number of simulees used in most simulation studies on subscores (e.g., de la Torre & Patz, 2005; de la Torre & Song, 2009; Sinharay, 2010; Tate, 2004; Yao & Boughton, 2007, Yao, 2010). The upper end choice—5,000 examinees—reflects the average cohort size for 2014/2015 in the Guilford County Schools (Public Schools of North Carolina (NC DPI), 2016).

Number of Nuisance Dimensions

With the emergent use of innovative item formats that assess complex problem solving skills and higher critical thinking ability using a computer interface and devices in the next generation, K-12 CCR assessments, at least two potential irrelevant constructs, as I mention in Chapter Two, may exist: item/response format and interfering linguistic complexity. At the same time, it is important to examine the case in which no nuisance dimension is present (baseline). I will therefore use zero, one, and two nuisance dimension(s).

Correlations between Primary Dimensions

I will vary the level of association between the four traits. I will set the levels of associations as follows:

$$\rho_{\theta_1,\theta_2} = \rho_{\theta_1,\theta_3} = \rho_{\theta_1,\theta_4} = \rho_{\theta_2,\theta_3} = \rho_{\theta_2,\theta_4} = \rho_{\theta_3,\theta_4} = .40;$$
$$\rho_{\theta_1,\theta_2} = \rho_{\theta_1,\theta_3} = \rho_{\theta_1,\theta_4} = \rho_{\theta_2,\theta_3} = \rho_{\theta_2,\theta_4} = \rho_{\theta_3,\theta_4} = .80.$$

I assume that the subdomains of a content domain CCR assessment are likely to be somewhat moderately correlated. Uncorrelated subdomains are not likely to be observed. Operational tests reported by other researchers have shown average correlations between subscores that range between .42 and .77 with average disattenuated correlations between .75 and 1.00 (Sinharay, 2010; Sinharay, Puhan, & Haberman, 2010).

Correlations between Nuisance Dimensions

As shown in Table 5, when there is no nuisance dimension present or when the number of nuisance dimension is at most one, I will not alter the correlation structure for the nuisance dimension. When two nuisance dimensions are simulated to be present in a given dataset, I will set the levels of associations between the two nuisance dimensions as follows:

$$\rho_{\theta\,\text{nuisance}\,1,\theta\,\text{nuisance}\,2} = .00, .40, .70$$

I hypothesize that item/response format construct will be either not related, somewhat related, or highly related with the interfering linguistic complexity construct. However, these choices are tentative and exploratory since I have found no studies so far that have investigated the association between linguistic complexity and item/response format.

Discrimination Levels on Primary Dimension

I will also test the effect of item discrimination level. I will use two different combinations of item discrimination levels for the four primary dimensions: (1) all four dimensions have high item discriminations (high:high:high:high) and (2) all four dimensions have high item discriminations (low:low:low:low). I will set high item discriminations between .90 and 1.30 and low item discriminations between .40 and .70. I will set all nuisance dimensions to having low item discrimination.

Number of Items per Dimension

I will include either 10 or 20 items in each subdomain so that a test of four subdomains, each with 10 items, will have a total of 40 items and the same test using 20 items per subdomain will result in a total of 80 items. I selected these numbers of items to represent what would be considered an average length subdomain and long subdomain (respectively) in a large-testing program. Yao & Boughton (2007) simulated item response data based on item parameters from a large scale Grade 8 math assessment with four objectives/dimensions and 12 to 18

items per objective/dimension. Several simulation studies on subscores estimation (de la Torre & Patz, 2005; de la Torre & Song, 2010; de la Torre, Song, & Hong, 2011) consistently manipulated each subscore length as 10, 20, or 30 items. Given proper considerations on the practicality and feasibility (i.e., financial cost and administration time (Wainer & Feinberg, 2015)) of the new generation assessments, the subscores' lengths I have chosen to use reflect the reality in most operational settings (e.g., SBAC & PARCC).

Calibration/Modeling Strategies

I will use two different strategies to calibrate the crossed-condition item response data: (1) a UIRT model for each primary latent trait where dichotomous and polytomous items are simultaneously calibrated (also referred as concurrent calibration) using the 2PL model and the GRM, respectively; and, (2) a full-information, confirmatory, compensatory MIRT model for the four primary traits in which the calibration of the mixed-item formats is conducted simultaneously but dichotomous items are calibrated using the M2PL model and polytomous items are calibrated using the MGRM.

Data Generation

I will generate the data for all of the replications of the six test configurations using R (R Core Team, 2016), specifically two of its packages: 'MASS' (Venables & Ripley, 2002) and 'mirt; (Chalmers, 2012). Data generation process in R can be described in five different steps: (1) simulation of the vector of 'true' item

discrimination parameters, $\mathbf{a}_i$, (2) simulation of the 'true' intercept parameters, d_i and the 'true' 'pseudo guessing' parameters, (3) simulation of the vector of 'true' latent ability parameters, $\boldsymbol{\theta}_j$, (4) simulation of one mixed format item response data given the generated parameters by programming a create_data( ) function, and (5) simulation of item response data, fully crossed for all levels of study conditions. I describe these five steps below.

Step 1

I will use a random uniform distribution to generate the vector of $\mathbf{a}_i$ parameters. To generate items with high discriminations, I will set the minimum value to .90 and the maximum value to 1.30. I will set the minimum and maximum values for the items with low discrimination to .40 and .70, respectively. Vector $\mathbf{a}_i$ will have a $1\times p+b$ structure with p indicating the number of primary dimensions or latent factors in the coordinate space and b representing the number of nuisance dimensions. It is true that $\mathbf{a}_i$ parameters are often simulated using lognormal distribution to ensure that the generated values are in positive values. However, using lognormal distribution can make the control of $\mathbf{a}_i$ difficult. Using the condition for 10 items per dimension with four primary constructs, the structure of the $\mathbf{a}_i$ parameters with zero, one, and two nuisance dimension(s) are shown (respectively) in Figures 6(a), 6(b), and 6(c). When one nuisance dimension is present, the $\mathbf{a}_i$ parameters will follow the five-dimensional MIRT model structure. With the

presence of two nuisance dimensions, the structure of the $\mathbf{a}_i$ parameters will follow the structure of the six-dimensional MIRT model. I will model the primary factor(s) using nuisance dimension(s) with low level item discriminations that load on all items to ensure simplicity and more readily explainable results.

Figure 6. Three Different Item Pattern Matrices for a Test with 40 Items

$$\boldsymbol{\alpha}_{0\,\text{nuisance}} = \begin{bmatrix} \alpha_{11} & 0 & 0 & 0 \\ \vdots & 0 & 0 & 0 \\ \alpha_{10,1} & 0 & 0 & 0 \\ 0 & \alpha_{11,2} & 0 & 0 \\ 0 & \vdots & 0 & 0 \\ 0 & \alpha_{20,2} & 0 & 0 \\ 0 & 0 & \alpha_{21,3} & 0 \\ 0 & 0 & \vdots & 0 \\ 0 & 0 & \alpha_{30,3} & 0 \\ 0 & 0 & 0 & \alpha_{31,4} \\ 0 & 0 & 0 & \vdots \\ 0 & 0 & 0 & \alpha_{40,4} \end{bmatrix}$$

(a)

$$\boldsymbol{\alpha}_{1\,\text{nuisance}} = \left[\begin{array}{c|cccc} \alpha_{11} & \alpha_{12} & 0 & 0 & 0 \\ \alpha_{\ldots,1} & \vdots & 0 & 0 & 0 \\ \alpha_{10,1} & \alpha_{10,2} & 0 & 0 & 0 \\ \alpha_{11,1} & 0 & \alpha_{11,3} & 0 & 0 \\ \alpha_{\ldots,1} & 0 & \vdots & 0 & 0 \\ \alpha_{20,1} & 0 & \alpha_{20,3} & 0 & 0 \\ \alpha_{21,1} & 0 & 0 & \alpha_{21,4} & 0 \\ \alpha_{\ldots,1} & 0 & 0 & \vdots & 0 \\ \alpha_{30,1} & 0 & 0 & \alpha_{30,4} & 0 \\ \alpha_{31,1} & 0 & 0 & 0 & \alpha_{31,5} \\ \alpha_{\ldots,1} & 0 & 0 & 0 & \vdots \\ \alpha_{40,1} & 0 & 0 & 0 & \alpha_{40,5} \end{array}\right]$$

(b)

$$\boldsymbol{\alpha}_{2\,\text{nuisance}} = \left[\begin{array}{cc|cccc} \alpha_{11} & \alpha_{12} & \alpha_{13} & 0 & 0 & 0 \\ \alpha_{\ldots,1} & \alpha_{\ldots,2} & \vdots & 0 & 0 & 0 \\ \alpha_{10,1} & \alpha_{10,2} & \alpha_{10,3} & 0 & 0 & 0 \\ \alpha_{11,1} & \alpha_{11,2} & 0 & \alpha_{11,4} & 0 & 0 \\ \alpha_{\ldots,1} & \alpha_{\ldots,2} & 0 & \vdots & 0 & 0 \\ \alpha_{20,1} & \alpha_{20,2} & 0 & \alpha_{20,4} & 0 & 0 \\ \alpha_{21,1} & \alpha_{21,2} & 0 & 0 & \alpha_{21,5} & 0 \\ \alpha_{\ldots,1} & \alpha_{\ldots,2} & 0 & 0 & \vdots & 0 \\ \alpha_{30,1} & \alpha_{30,2} & 0 & 0 & \alpha_{30,5} & 0 \\ \alpha_{31,1} & \alpha_{31,2} & 0 & 0 & 0 & \alpha_{31,6} \\ \alpha_{\ldots,1} & \alpha_{\ldots,2} & 0 & 0 & 0 & \vdots \\ \alpha_{40,1} & \alpha_{40,2} & 0 & 0 & 0 & \alpha_{40,6} \end{array}\right]$$

(c)

Step 2

I will generate the intercept/location parameters, d_i, for both test formats using a random uniform distribution with minimum value of -3.00 and maximum value of 3.00. To generate the pseudo-guessing parameters, I use the beta distribution.

Step 3

I will create the primary latent ability and nuisance constructs using multivariate standard normal distributions with various correlations conditions. I will create three pattern matrix structures for four primary constructs each with 10 or 20 items using the 'MASS' package (Venables & Ripley, 2002) in R (R Core Team, 2016). I will set the vector of means (i.e., $\boldsymbol{\mu}$) for all dimensions to .00; the structures of the positive-definite symmetric matrices specifying the correlation matrix of the dimensions (i.e., $\boldsymbol{\sigma}$) for all possible crossed-conditions are shown in Tables 11(a) and 11(b). Table 11(a) provides the structure of $\boldsymbol{\sigma}$ (sigma) for tests when there is no nuisance dimension or when there is only one nuisance dimension. Table 11(b) provides the structure of $\boldsymbol{\sigma}$ for tests with two nuisance dimensions.

Table 11(a). Structure of Sigma for Test Format when There is No Nuisance Dimension or One Nuisance Dimension

No. of Nuisance Dimension = 0										
	Correlation between Primary Dimensions = .40					Correlation between Primary Dimensions = .80				
	P1	P2	P3	P4			P1	P2	P3	P4	
P1	1					P1					
P2	0.4	1				P2	0.8	1			
P3	0.4	0.4	1			P3	0.8	0.8	1		
P4	0.4	0.4	0.4	1		P4	0.8	0.8	0.8	1	
No. of Nuisance Dimension = 1											
	P1	P2	P3	P4	N1		P1	P2	P3	P4	N1
P1	1					P1	1				
P2	0.4	1				P2	0.8	1			
P3	0.4	0.4	1			P3	0.8	0.8	1		
P4	0.4	0.4	0.4	1		P4	0.8	0.8	0.8	1	
N1	.00	.00	.00	.00	1	N1	.00	.00	.00	.00	1

Note. P: primary dimension; N: nuisance dimension

Table 11(b). Structure of Sigma for Test Format when There are Two Nuisance Dimensions

No. of Nuisance Dimension = 2																			
Correlation between Primary Dimensions = .40																			
	Correlation between Nuisance Dimensions = .00							Correlation between Nuisance Dimensions = .40							Correlation between Nuisance Dimensions = .70						
	P1	P2	P3	P4	N1	N2		P1	P2	P3	P4	N1	N2		P1	P2	P3	P4	N1	N2
P1	1						P1	1						P1	1					
P2	0.4	1					P2	0.4	1					P2	0.4	1				
P3	0.4	0.4	1				P3	0.4	0.4	1				P3	0.4	0.4	1			
P4	0.4	0.4	0.4	1			P4	0.4	0.4	0.4	1			P4	0.4	0.4	0.4	1		
N1	.00	.00	.00	.00	1		N1	.00	.00	.00	.00	1		N1	.00	.00	.00	.00	1	
N2	.00	.00	.00	.00	.00	1	N2	.00	.00	.00	.00	.40	1	N2	.00	.00	.00	.00	.70	1
Correlation between Primary Dimensions = .80																			
	Correlation between Nuisance Dimensions = .00							Correlation between Nuisance Dimensions = .40							Correlation between Nuisance Dimensions = .70						
	P1	P2	P3	P4	N1	N2		P1	P2	P3	P4	N1	N2		P1	P2	P3	P4	N1	N2
P1	1						P1	1						P1	1					
P2	0.8	1					P2	0.8	1					P2	0.8	1				
P3	0.8	0.8	1				P3	0.8	0.8	1				P3	0.8	0.8	1			
P4	0.8	0.8	0.8	1			P4	0.8	0.8	0.8	1			P4	0.8	0.8	0.8	1		
N1	.00	.00	.00	.00	1		N1	.00	.00	.00	.00	1		N1	.00	.00	.00	.00	1	
N2	.00	.00	.00	.00	.00	1	N2	.00	.00	.00	.00	.40	1	N2	.00	.00	.00	.00	.70	1

Note. P: primary dimension; N: nuisance dimension

Step 4

Using the generated item parameters from Steps 1 and 2 and the latent ability/constructs parameters from Step 3, I will use a function, which I named create_data( ), written in R (R Core Team, 2016), to simulate the mixed format item response data. Specifically, I will use the 'mirt' package (Chalmers, 2012) to simulate the mixed response data (see Luecht & Ackerman (2017) for better explanations on how the random number generator generates the observed item response data). I will simulate the dichotomous intercept/location parameters using the M3PL model. I will use MGRM to generate the polytomous intercept/location parameters. I will create each polytomous item with five score categories, where u=0,1,2,3,4.

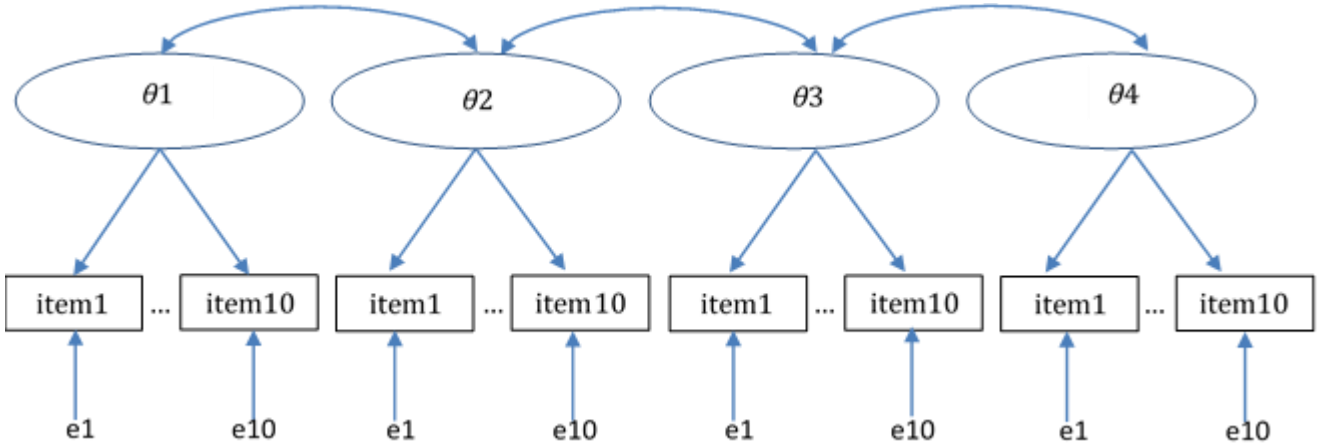
Step 5

Finally, I will use the test configuration conditions for Test Formats 1 and 2 to generate the crossed-condition datasets by using the 'for loop' function in R (R Core team, 2016) and my function create_data( ) in Step 4. The first test format will have a total of 32 crossed-conditions while the second item format will have 48 crossed-conditions resulting in a total of 80 crossed-conditions. I will replicate the conditions using 10 iterations, resulting in a total of 8,000 item response data points for data analysis.

Structure of the Generated Response Data

Figures 7, 8, and 9 show how I will manipulate the dimensional structures in the data generation process using a path diagram format. Figure 7 displays the data structure for the baseline condition (no nuisance dimension present) in which the data structure has 10 items per dimension. At the top of each Figure, four ovals labeled θ_1 to θ_4, represent four primary latent traits or subdomains for a given content domain from either the CCSS or NGSS. The curved arrow between these traits signifies that the traits are associated with one of the correlation levels I will be testing. Below the primary latent traits are rectangular boxes that represent the items in a given test that measure the latent traits. The single headed arrow connecting each item to its corresponding latent trait is a factor loading that describes the strength of the relationship between the item and trait. Each of these factor loadings, in the context of IRT, represents the set of true item discrimination parameters ($\mathbf{a}_i$) that I will be simulating in the data generation process (see Figure 6(a)). Finally, the "e" terms in Figures 7, 8, and 9 are the measurement error variables. For instance, e1 is the measurement error for item 1 and e10 is the measurement error for item 10.

Figure 7. Schematic Diagram of the Structure of Generated Data for 10 Items per Subtest with No Nuisance Dimension



Figures 8 and 9 illustrate how I will manipulate the dimensional structures to form locally dependent item sets due to unintended underlying dimensionality (when one nuisance or two nuisance dimension(s) exist(s)). The previous descriptions of Figures 7 also apply to Figures 8 and 9. Each latent trait in the two figures has 10 items. In addition to the four primary latent traits, the two latter figures also show secondary latent factor(s) associated with the test items. The secondary factor is labelled as $\theta_{\text{nui}1}$ in Figure 8 to indicate the presence of only one nuisance trait or irrelevant construct. In Figure 9, the two secondary factors represented by $\theta_{\text{nui}1}$ and $\theta_{\text{nui}2}$ indicate that two nuisance factors exist in the test and are associated with the test items. When two nuisance factors are present in a given test, as shown in Figures 9, a curved arrow is shown between the traits in each of the figure to indicate that the nuisance factors may be associated with one another based on the correlation-between-nuisance-dimension levels I will be testing.

As discussed earlier, the data structure with one nuisance dimension (Figure 8) follows the five dimensional MIRT model and the data structure with two nuisance dimensions follows the four dimensional MIRT model with both dimensions loaded on all items (Figure 9).

Table 12 provides the summary of the constant and conditions of my simulation study along with the corresponding rationales of selecting such constant and conditions.

Figure 8. Schematic Diagram of the Structure of Generated Data for 10 Items per Subtest with the Presence of One Nuisance Dimension

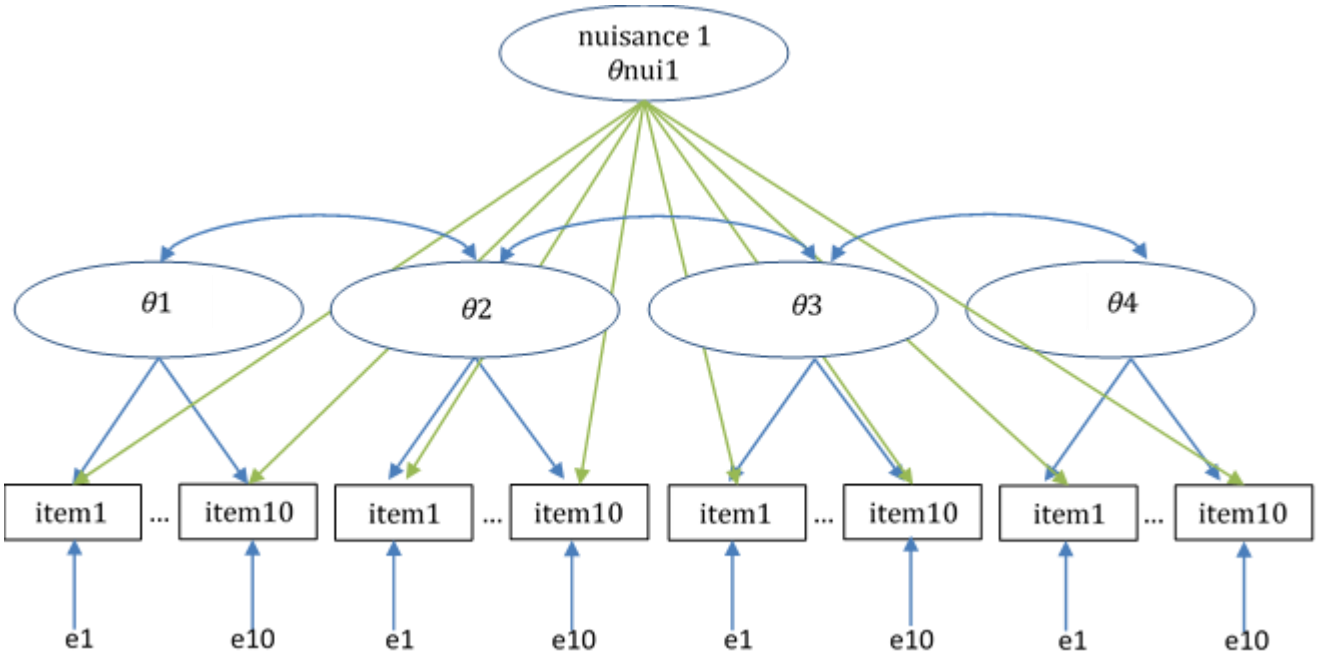


Figure 9. Schematic Diagram of the Structure of Generated Data for 10 Items per Subtest with the Presence of Two Nuisance Dimension

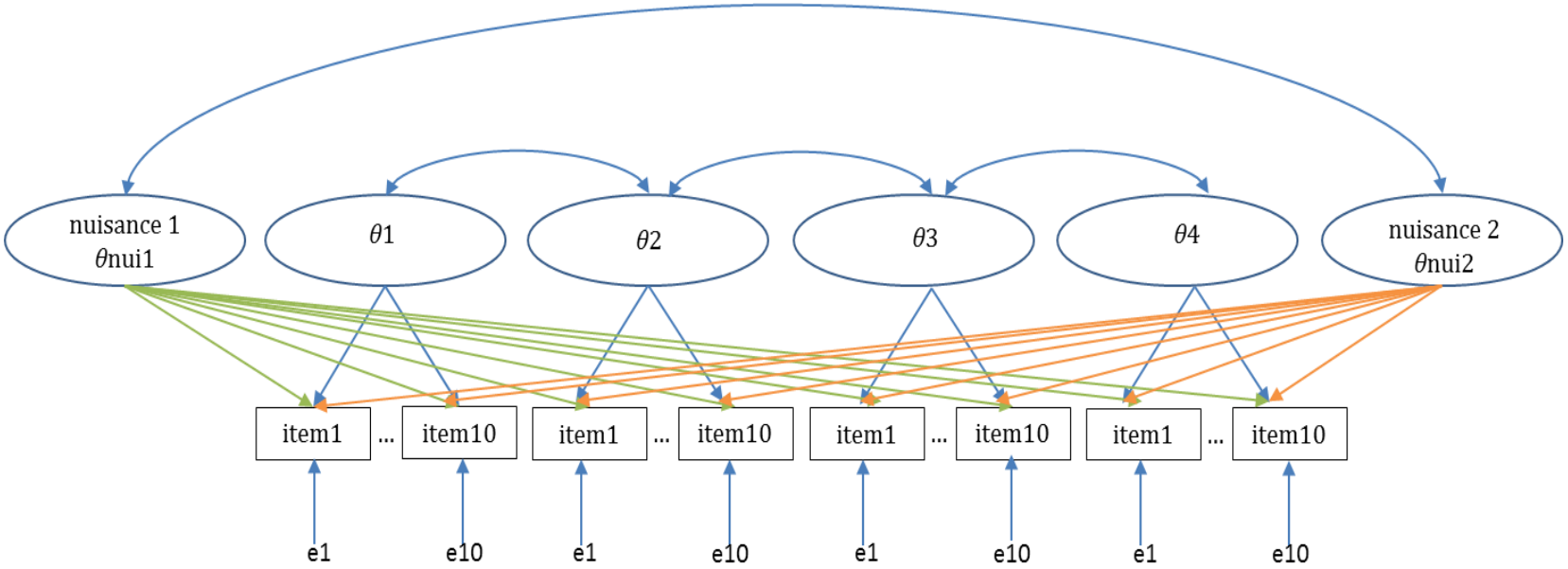


Table 12. Summary & Corresponding Rationale of the Constant & Conditions of Simulation Study

			Rationale
Constant	Number of primary dimensions	• 4	• Consistent with the content domains of next generation assessments
Variables	Number of nuisance dimensions	• 0 • 1 • 2	• The case in which no nuisance dimension is present (baseline) • At least two potential irrelevant constructs may exist: item/response format & interfering linguistic complexity
	Correlations between primary dimensions	• .40 • .80	• Operational tests have shown average correlations between subscores that range between .42 and .77 (Sinharay, 2010; Sinharay, Puhan, & Haberman, 2010).
	Correlations between nuisance dimensions	• .00 • .40 • .70	• When two nuisance dimensions are simulated to be present in a given dataset, the levels of associations between the two nuisance dimensions will be: not related, somewhat related, and highly related • These choices are tentative and exploratory • No studies so far have investigated the association between linguistic complexity and item/response format
	Discrimination levels on primary dimensions	• All high • All low	• Two different combinations of item discrimination for the four primary domains • High item discrimination: .90 to 1.30 • Low item discrimination: .40 to .70
	Number of items per dimension	• 10 items • 20 items	• 10 and 20 items per dimension is practical and feasible (i.e., financial cost and administration time (Wainer & Feinberg, 2015)) • Reflect the reality in most operational settings for next generation assessments(e.g., SBAC & PARCC).
	Sample sizes	• 1,000 • 5,000	• 1,000 examinees are used in most simulation studies on subscores (e.g., de la Torre & Patz, 2005; de la Torre & Song, 2009; Sinharay, 2010; Tate, 2004; Yao & Boughton, 2007; Yao, 2010). • 5,000 examinees reflects the average cohort size for 2014/2015 in the Guilford County Schools (Public Schools of North Carolina (NC DPI), 2016).

Item Parameter Estimation

I will conduct the item calibrations using the flexMIRT® 3.0 commercial software program (Vector Psychometric Group, 2017). This software employs the estimation using two estimation methods: (1) the expectation-maximization (EM) algorithm (Bock & Aitkin 1981) and (2) the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm (Cai, 2010b, 2010c) that uses a stochastic approximation algorithm.

Although the EM algorithm is widely used in estimating item parameters in IRT, the estimation method becomes unwieldy, time consuming, and computationally expensive (Cai, 2010b, 2010c; Chalmers, 2012; Houts, & Cai, 2015) when high-dimensional models are used. Houts and Cai (2015, p. 74) noted that, "[a]s the number of dimensions increases linearly, the number of Gauss-Hermite quadrature points increases exponentially". Given the high-dimensional models ($p \geq 4$) in this study, I will use estimation procedures implemented with the MH-RM algorithm to estimate the item parameters.

Estimation/Scoring of Latent Ability

In a large-scale Monte Carlo study comparing the performance of 10 latent trait estimators, which included two latent ability estimation Bayes methods, the expected-*a-posteriori* (EAP) (Bock & Mislevy, 1982), and maximum-*a-posteriori* (MAP) (Samejima, 1980; Bock, 1983), Wainer and Thissen (1987) concluded that EAP performed better than the other estimators when used with the 3PL model and

was most likely the best choice among the latent ability estimates that have been developed and made widely available in commercial software. Thissen et al. (2001) employed the EAP estimator with the GRM and GPCM to estimate scale scores for an end-of-course exam. They found that, for the most part, scale scores for the same response pattern from either the GRM or the GPCM differed by less than .10 standard units. Thissen and Orlando (2001) concurred with Wainer and Thissen (1987) by suggesting that the EAP estimation method provided the best scale score computed using item response models. I will therefore employ EAP in order to estimate/score the latent ability estimations with standard normal prior in flexMIRT® 3.0 program (Vector Psychometric Group, 2017).

Criteria for Evaluating the Results: Luecht and Ackerman's Expected Response Function Approach

Luecht and Ackerman (2017) introduce an innovative approach to the model-based simulation study. Using the expected response function (ERF) approach, they encourage the use of complex IRT-based data generation models that could, to an extent, resemble complex testing features, including the following: the representation of complex causal factors; non-random sources of missing data; test-wiseness and cheating tendency; assessment-related method factors (e.g., response/item formats, test accommodations); construct-irrelevant dimensions (e.g., interfering linguistic complexity); intended assignation of bias; and, many other potential simulation conditions "that could contaminate [the] estimated

metrics, misrepresent the assumed properties of those metrics, or simply lead to estimation issues such as instability or bias in the parameters estimates" (p.4).

Their new approach involves a simple transformation of the ERFs for expected raw scores (ERS) that would allow for more realistic model-based simulation studies to be conducted. Specifically, their approach employs the ERS-based residuals in place of the commonly used criterion of IRT parameter recovery. The ERF approach cleanly separates residuals due to data-model fit from the estimation error even under different calibration models (e.g., UIRT versus MIRT models) and will be highly useful in my model-based simulation study to distinguish between mean, variance, and covariance functions of those residuals. The approach also eliminates arbitrary scaling constraints (e.g., location and unit size), shrinkage issues associated with Bayesian estimation, and other complications encountered when dealing with competing models and estimation strategies.

Luecht and Ackerman remind researchers conducting simulation studies of the importance of replicating the various factors and the interaction among factors in order to capture the complexity of the real data and testing situations. In my simulation study I attempt to replicate real data complexity and realistic testing situations by integrating potential nuisance factor(s) along with different testing configurations based on relevant studies in the literature in order to situate my simulation study in the context of the next generation CCR assessment that now includes different examinees from different subpopulations in its assessment

system (i.e., ELLs and SWDs). I will detail Luecht and Ackerman's (2017) ERF approach below.

The ERF for item *i* is given mathematically (Luecht & Ackerman, 2017, p. 7) as:

$$f_i(u \mid \boldsymbol{\theta}) = E(u_i \mid \boldsymbol{\theta}; \boldsymbol{\xi}_i) = \sum_{k=1}^{m} X_{ik} P_{ik}(\boldsymbol{\theta}) \tag{52}$$

where i is the individual test item, u is the observed response to item i, $\boldsymbol{\theta}$ is the person parameters (in this case, it is a vector which indicates multidimensionality of the latent abilities/traits), $\boldsymbol{\xi}_i$ is the item parameters for item i for a given (or for a combination of) IRT model, m is possible score category where $\mathbf{X}_i = \{X_{i1}, X_{i2}, ..., X_{im}\}$. The ERF simplifies further to $P_i(\boldsymbol{\theta})$ for all dichotomously scored items in all IRT models. For the polytomous items in my study, $\mathbf{X}_i = \{X_{i1}, X_{i2}, X_{i3}, X_{i4}\}$ since m = 4, where u=0,1,2,3,4.

Building on the ERF in equation (52), Luecht and Ackerman (2017) ask researchers to consider different types of ERFs in their own simulation studies. In any simulation study, there are at least two types of ERFs: (1) the true ERF and (2) the estimated ERF. If the ERF in equation (52) is the true ERF, the researchers thus know the 'true' (i.e., the generated) values of the person and the item parameters. When estimated person and item parameters are used to compute the ERF, the ERF has now become an estimated quantity. Luecht and Ackerman (2017) further suggest that the estimated ERF is not an observable variable but a model-based

representation of the observed variable(s). The mathematical function of the estimated ERF is given in equation (53).

$$\hat{f}_i(u \mid \hat{\boldsymbol{\theta}}) = E(u_i \mid \hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\xi}}_i) = \sum_{k=1}^{m} X_{ik} P_{ik}(\hat{\boldsymbol{\theta}}) \tag{53}$$

In my simulation study in the context of the next generation CCR assessments, I explicitly model some 'potential' nuisance factor(s) (what I termed as innovative item formats or/and interfering linguistic complexity) when simulating the simulated examinees' response data. Using Luecht and Ackerman's (2017) ERF approach, I can further model four different types of true and estimated ERFs. The four ERFs are: (1) $f0$: ERF based on my 'true'/generated IRT model (contaminated), (2) $f1$: ERF based on my 'true' IRT model of interest (purified), (3) $\hat{f}1_{\text{MIRT}}$: estimated ERF for the MIRT calibration model, and (4) $\hat{f}1_{\text{UIRT}}$: estimated ERF for the UIRT calibration model. The mathematical expressions of the four ERFs are given briefly in equations (54) to (57) below, respectively.

$$f0_i(u \mid \boldsymbol{\theta}_{1:4/5/6}, \mathbf{a}_i, c_i, d_i) = \sum_{k=1}^{m} X_{ik} [P_{ik}(\boldsymbol{\theta}_{1:4/5/6})_{\text{M3PL\&MGRM}}] \tag{54}$$

$$f1_i(u \mid \boldsymbol{\theta}_{1:4}, \mathbf{a}_i, d_i) = \sum_{k=1}^{m} X_{ik} [P_{ik}(\boldsymbol{\theta}_{1:4})_{\text{M2PL\&MGRM}}] \tag{54}$$

$$\hat{f}1_{i\,\text{MIRT}}(u \mid \hat{\boldsymbol{\theta}}_{1:4}, \hat{\mathbf{a}}_i, \hat{d}_i) = \sum_{k=1}^{m} X_{ik} [P_{ik}(\hat{\boldsymbol{\theta}}_{1:4})_{\text{M2PL\&MGRM}}] \tag{56}$$

$$\hat{f}1_{i\,\mathrm{UIRT}}(u \mid \hat{\theta}_{1:4}, \hat{a}_i, \hat{b}_i, \hat{d}_i) = \sum_{k=1}^{m} X_{ik}[P_{ik}(\hat{\boldsymbol{\theta}}_{1:4})_{\mathrm{2PL\&GRM}}] \quad (57)$$

In equation (54), when there is no nuisance dimension present, there will be only four 'true'/generated person parameters ($\boldsymbol{\theta}_{1:4}$) and four item discrimination parameters ($\mathbf{a}_i$). If there is one nuisance dimension, there will be five person parameters ($\boldsymbol{\theta}_{1:5}$) and five item discrimination parameters ($\mathbf{a}_i$). Similarly, when two nuisance dimensions are modeled, there will be six generated person parameters ($\boldsymbol{\theta}_{1:6}$) with six item discrimination parameters ($\mathbf{a}_i$). Equation (55) is the mathematical expression for the 'true' ERF based on the four dimensions of interest (i.e., the content areas of interest such as in ELA, math, and science), known earlier as $f1$. Equations (56) and (57) are the mathematical expressions for the estimated ERFs when the observed response data are calibrated using the MIRT ($\hat{f}1_{\mathrm{MIRT}}$) or UIRT ($\hat{f}1_{\mathrm{UIRT}}$) calibration model, respectively.

Given the item parameters that I generated in Steps 1 and 2 (see Data Generation section in this chapter), the person parameters generated in Step 3, the observed response data in Step 4, and my four true and estimated ERFs in equations (54) to (57), I now am able to compute different types of residuals (Luecht & Ackerman, 2017) to answer my two research questions.

Luecht and Ackerman (2017) describe four different types of residuals based on their ERF approach: (1) e0: the total residuals/errors, (2) e1: bias-induced

residuals, (3) e2: parameter-estimation residuals, and (4) e3: estimated model-data fit residuals. Following the previous notations, these residuals can be expressed mathematically in equations (58) to (61) below.

$$\varepsilon 0 = u - f0 = \varepsilon 1 + \varepsilon 2 + \varepsilon 3 \tag{58}$$

$$\varepsilon 1 = f1 - f0 \tag{59}$$

$$\varepsilon 2 = \hat{f} - f1 \tag{60}$$

$$\varepsilon 3 = u - \hat{f} \tag{61}$$

Applying Luecht and Ackerman's different types of ERF-based residuals to the variables in my study, I will have six different residuals in which I will have a pair of parameter estimation residuals for MIRT and UIRT calibrations and a pair of model-data fit residuals for both estimated MIRT and UIRT. All ERF-based residuals in my study are shown below:

$$\varepsilon 0 = u - f0 = \varepsilon 1 + \varepsilon 2_{\text{MIRT}} + \varepsilon 3_{\text{MIRT}}$$

or

$$\varepsilon 0 = u - f0 = \varepsilon 1 + \varepsilon 2_{\text{UIRT}} + \varepsilon 3_{\text{UIRT}} \tag{62}$$

$$\varepsilon 1 = f1 - f0 \tag{63}$$

$$\varepsilon 2_{\text{MIRT}} = \hat{f}_{\text{MIRT}} - f1 \tag{64}$$

$$\varepsilon 2_{\text{UIRT}} = \hat{f}_{\text{UIRT}} - f1 \tag{65}$$

$$\varepsilon 3_{\text{MIRT}} = u - \hat{f}_{\text{MIRT}} \tag{66}$$

$$\varepsilon 3_{\text{UIRT}} = u - \hat{f}_{\text{UIRT}} \tag{67}$$

Essentially, the ERF-based approach developed by Luecht and Ackerman (2017) is very useful and provides significant contribution in the educational measurement field, especially for researchers who intend to conduct simulation studies that could resemble the complexity of the real data.

> Re-conceptualizing the bias errors, estimation errors and model-data fit errors in terms of these four types of residuals rather cleanly avoids across-model or across-estimator comparability issues by simply transforming all of the simulated and estimated parameters into the number-correct score space. ... This re-conceptualization of the residuals offers some important substantive advantages insofar as interpreting the absolute magnitude of various types of "errors." (Luecht & Ackerman, 2017, p. 11)

They added:

> These transformations therefore let [researchers] assess the direction and magnitude of bias, estimation error or model-data fit on a common metric that works for items or examinees—regardless of the complexity of the IRT model(s) used. We might therefore call these types of ERF transformations and the ensuing residuals, "**metric neutral**". An added bonus is that the residual metric is based on the number-correct item or test score scale and therefore allows the magnitude of error to be directly interpreted relative to "*difference that matte*rs" (DTM) criteria (c.f., Dorans & Feigenbaum, 1994). (Luecht & Ackerman, 2017, p. 14)

Before discussing in detailed how I will use the ERF-based residuals approach by Luecht and Ackerman (2017) as criteria to evaluate both of my research questions, it is also noteworthy to briefly describe the concept of the DTM (Dorans & Feigenbaum, 1994). The logic of DTM in test equating, scaling, and linking

(Kolen & Brennan, 2004) is that less than half a reported score unit of the combined group linking at a given raw score point is ignorable. To state this differently, the DTM denotes the difference in scores that would cause a significant change in reported scores. The DTM may be defined differently in different contexts. On the SAT—a standardized test widely used for college admissions in the US—the DTM is defined as a difference of five points since SAT scores are reported in ten point intervals. Adding five points to any score would result in a student's score being rounded to the next highest score (Holland & Dorans, 2006). The key in defining the DTM in any context is to define the DTM so that it reflects a change in scores that would make a practical difference in the scores reported to examinees.

Criteria for Evaluating the Results for Research Question 1: Comparison of the ERF-Based, Metric-Neutral, Residuals for MIRT and UIRT Models

As a reminder, my first research question concerns the amount of ERF-based residual covariance produced by different, more parsimonious, IRT calibration models when the generated ('true') model represents a more complex reality with nuisance dimensions such as in the next generation, mixed-method, CCR online assessments. Which calibration method performs best:

a. When the nuisance dimension(s) is/are present?
b. When correlations between nuisance dimensions vary?
c. When correlations between primary dimensions vary?
d. When item discrimination levels on primary dimensions vary?

e. When the number of item in each primary dimension varies?

f. Over various sample sizes?

More specifically, I will examine the performance of two IRT calibration models using Luecht and Ackerman's (2017) ERF-based residuals approach. The calibration models include a UIRT model and a MIRT model.

To answer this question, I will first summarize the descriptive statistics of all six different types of the ERF-based residuals for each sub-research question. I will conduct all computations to obtain the residuals in the R program (R Core Team, 2016) and the Microsoft Excel version 2016. Specifically, I provide the mean, standard deviation, minimum, and maximum residuals amount. I will also illustrate and describe relevant figures to present my findings.

Two types of residuals (Luecht & Ackerman, 2017) are my primary interest: the estimated model data fit residuals for both MIRT and UIRT calibrations (hereafter e3MIRT and e3UIRT) and the parameter estimation residuals from the two calibration models (hereafter e2MIRT and e2UIRT). The e3MIRT is the difference between the observed response data (u) and the estimated parameters from the M2PL model and the MGRM ($\hat{f}1_{\text{MIRT}}$). The e3UIRT is the difference between the observed response data (u) and the ERF computed using the estimated parameters from the 2PL model and GRM ($\hat{f}1_{\text{UIRT}}$). The e2MIRT is the difference between $\hat{f}1_{\text{MIRT}}$ and the ERF computed using the generated parameters of interest from the M3PL model and MGRM ($f1$: true ERF of interest). The e2UIRT is the

difference between $\hat{f}1_{\text{UIRT}}$ and the $f1$. The summary of these residuals of interest is shown in Table 10 in Chapter Four.

To determine whether there are mean differences in the levels of conditions for the study, I will conduct a series of two- and three- way factorial analysis of variance (ANOVA) models using SPSS 19 (IBM Corp. Released 2010). Due to the large data set size, I will set the significance level for the ANOVA tests to .001. Additionally, an effect size, partial eta-squared (η^2), accompanies the results of the ANOVAs. Effect size heuristics are based on those of Gray & Kinnear (2012) in which *small effect size* ranges from .01 to less than .06, *medium effect size* ranges from .06 to less than .14, and *large effect size* occurs when η^2 is equal to or greater than .14.

Criteria for Evaluating the Results for Research Question 2:

Examination of Bias-Induced Residual Covariance ($r_{e1i,e1h}$)

In my second research question, I will examine the impact of the following six testing configurations on the amount of of modeled, bias-induced, residual covariance:

a. The presence of nuisance dimension(s)?
b. The strength of correlations between nuisance dimensions?
c. The strength of correlations between primary dimensions?
d. Changes in item discrimination levels on the primary dimensions?
e. The number of items in each primary dimension?

f. Changes in sample size?

To answer this question, I will summarize the modeled, bias-induced, residual covariances ($r_{\mathrm{e1}_i,\mathrm{e1}_h}$) for each sub-research question. As a reminder, the bias-induced residual (hereafter e1) is the difference between the ERF computed using the generated parameters from the M3PL model and MGRM when the presence of nuisance factor(s) vary and the ERF computed using the generated parameters of interest from the M3PL model and MGRM (see also Table 10). I compute the covariances/correlations for the e1 residuals using the R program (R Core Team, 2016).

Specifically, I will report the mean, standard deviation, minimum, and maximum residual covariance. To determine whether there are mean differences in the levels of conditions for the study, I will again conduct a series of factorial ANOVA models. Effect size heuristics for η^2 are based on Gray & Kinnear (2012).

CHAPTER IV

RESULTS

I present my results in two primary sections, one for each of my research questions. I divide each section into subsections based on the subquestions for each research question. Table 13 provides descriptions of the different types of residuals in this study which appear in equations (63) to (67). Reference to this table should prove useful in interpreting my results. I report descriptive statistics for all residuals based on conditional residuals. I conditioned the residuals on percentage scores. Also, I standardized the residuals from longer tests to ensure comparability with the residuals from shorter tests. Since I do not specifically focus on e0, which is the total error or total residual (see equations (58) and (62)), I will provide the descriptive statistics for e0 in Appendix E. The interested reader can use those statistics to confirm the amount of the rest of the residuals under investigation. I will start reporting my results with a descriptive summary in the beginning of each section of the research question answered in that section to provide the overall findings for all crossed conditions, after which I will provide the findings for each of the specific subquestions for that section.

Table 13. Descriptions & Operational Definitions of Residuals used as Criteria to Answer the Research Questions

Residual Label	Definitional Formula	Description of the Residual
e1	$\varepsilon 1 = f1 - f0$	Bias-induced residual Residual between • the true ERF with nuisance induced from the M3PL model & MGRM and • the true ERF of interest (purified) from the M3PL model & MGRM
e2MIRT	$\varepsilon 2_{\text{MIRT}} = \hat{f}_{\text{MIRT}} - f1$	Parameter estimation residual from MIRT calibration Residual between • the estimated ERF from the M2PL model & MGRM and • the true ERF of interest from the generated M3PL model & MGRM
e2UIRT	$\varepsilon 2_{\text{UIRT}} = \hat{f}_{\text{UIRT}} - f1$	Parameter estimation residual from UIRT calibration Residual between • the estimated ERF from the 2PL model & GRM and • the true ERF of interest from the generated M3PL model & MGRM
e3MIRT	$\varepsilon 3_{\text{MIRT}} = u - \hat{f}_{\text{MIRT}}$	Model-data fit residual from MIRT calibration Residual between • observed response data and • estimated ERF from M2PL model & MGRM
e3UIRT	$\varepsilon 3_{\text{UIRT}} = u - \hat{f}_{\text{UIRT}}$	Model-data fit residual from UIRT calibration Residual between • observed response data and • estimated ERF from 2PL model & GRM

Note.
MIRT: multidimensional item response theory; UIRT: unidimensional item response theory; ERF: expected response function;
M3PL: multidimensional three-parameter logistic; M2PL: multidimensional two-parameter logistic;
MGRM: multidimensional graded response model; 2PL: two-parameter logistic; GRM: graded response model

Results for Research Question 1: Comparison of the ERF-Based Residuals for MIRT and UIRT Models

In this section, I discuss the results related to my first research question. I asked about the amount of ERF-based residuals produced by different, more parsimonious, IRT calibration models when the generated ('true') model represents a more complex reality with nuisance dimensions such as those which exist in the next generation, mixed-method, CCR online assessments. Which calibration method performs best:

a. When the nuisance dimension(s) is(are) present?
b. When correlations between nuisance dimensions vary?
c. When correlations between primary dimensions vary?
d. When item discrimination levels on primary dimensions vary?
e. When the number of item in each primary dimension varies?
f. Over various sample sizes?

Descriptive Statistics for All Study Conditions

Tables 14 and 15 provide the summary for all parameter estimation residuals for MIRT and UIRT calibration, respectively. Table 14 summarizes the e2MIRT while Table 15 summarizes the e2UIRT. Tables 16 and 17 provide the summary for all model-data fit residuals for MIRT and UIRT calibration, respectively. Table 16 summarizes the e3MIRT while Table 17 summarizes the e3UIRT. All four tables (Tables 14 to 17) present the means, standard deviations,

and the ranges of the respective residuals based on all 80 crossed conditions of the study.

The purpose of the separate tables for each type of residual is to allow for overall comparisons of the residual amounts and structures across all crossed conditions. However, it is better to compare each residual type for different unidimensional and multidimensional calibrations to relate the result to the primary question asked.

On average, the parameter estimation residuals from the MIRT calibration (e2MIRT) in Table 14 are smaller and closer to zero than their UIRT counterparts (see Table 15). This is as I expected since the estimated ERF from the MIRT calibration is very similar to the generated model of interest. This result (the slight discrepancy between the e2MIRT and zero) might be due to a difference in the 'pseudo guessing' parameter from the generated model.

A closer examination reveals that the conditional means of all e2UIRT in Table 15 values are negative due to the obvious discrepancy in the estimated and generated ERFs. Again, the pseudo guessing induced in the data by the M3PL generating model substantiates the discrepancies. Moreover, the negative residuals from the UIRT calibration may also be explained by the compensatory nature of the generating multidimensional model. The model of interest is generated using the four dimensional compensatory model in which the compensatory model is based on a linear combination of coordinates of θ (Reckase, 2009). The linear combination is used to specify the probability of a response. The linear combination

of θ-coordinates can produce the same sum with various combinations of θ-values. If one θ-coordinate is low, the sum will be the same if another θ-coordinate is sufficiently high. Such features are not represented by and could not be captured by the UIRT model, hence, the UIRT model tends to underestimate the examinees' raw scores.

Similar to the findings from e2MIRT and e2UIRT, an overall comparison between the model-data fit residuals between the e3MIRT (Table 16) and e3UIRT (Table 17) suggests that the MIRT model on average fits the observed response data better than the UIRT model. Examination of the e3MIRT shows that the magnitude of the residuals from the MIRT calibration is lower and closer to zero than their UIRT counterparts. Again, this was as expected because the observed response data was generated using the multidimensional person and item parameters.

Table 14. Descriptive Statistics for Conditional e2MIRT (based on Percentage Scores) for All Crossed Conditions

1000 examinees																		
			10 items per dimension								20 items per dimension							
nuisance	vcor	ncor	low item discrimination				high item discrimination				low item discrimination				high item discrimination			
			mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max
0	0.40	NA	-0.50	2.72	-8.15	5.95	0.15	5.83	-13.92	21.37	-0.19	2.84	-8.44	8.55	0.01	0.90	-2.88	2.57
	0.80		-0.17	3.39	-9.87	10.20	-0.61	7.33	-19.32	20.10	-0.01	3.91	-13.21	12.10	0.04	1.02	-4.16	3.05
1	0.40	NA	-0.16	4.77	-14.59	12.55	0.36	6.53	-19.47	17.09	-0.03	4.52	-11.26	13.62	0.11	2.34	-8.08	7.02
	0.80		-0.22	4.92	-13.89	14.27	-0.04	8.10	-22.67	25.28	0.07	5.49	-15.55	15.06	0.01	2.54	-7.77	8.40
2		0.00	-0.20	5.87	-15.75	14.38	-0.75	7.25	-19.81	18.09	0.07	5.82	-14.68	16.35	0.11	3.91	-14.10	10.87
	0.40	0.40	0.07	6.46	-14.64	16.30	0.54	7.68	-19.14	19.37	0.00	6.19	-15.83	16.34	0.38	4.65	-14.18	15.52
		0.70	-0.09	7.37	-17.99	19.18	0.28	8.26	-20.77	19.63	0.22	6.61	-24.24	18.79	-0.13	5.10	-14.84	13.86
		0.00	-0.14	6.82	-22.16	15.54	0.25	8.66	-26.41	24.15	-0.15	5.87	-14.39	14.48	-0.26	3.48	-9.51	15.28
	0.80	0.40	0.39	7.14	-16.96	17.73	0.25	9.11	-25.84	21.03	0.12	6.53	-16.76	16.81	0.37	4.40	-12.72	16.05
		0.70	0.07	7.92	-18.26	24.30	-0.43	9.25	-26.00	20.43	-0.32	7.03	-20.94	19.05	-0.02	4.71	-15.55	15.90
5000 examinees																		
			10 items per dimension								20 items per dimension							
nuisance	vcor	ncor	low item discrimination				high item discrimination				low item discrimination				high item discrimination			
			mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max
0	0.40	NA	-0.05	3.32	-9.46	7.33	-0.07	6.74	-18.17	17.58	0.18	3.49	-9.64	10.89	0.09	0.72	-3.24	3.69
	0.80		0.15	4.20	-12.30	13.33	-0.04	8.77	-28.04	21.78	0.08	4.25	-11.53	13.29	0.00	0.69	-2.75	3.09
1	0.40	NA	-0.06	5.87	-15.22	13.09	0.02	8.00	-22.35	22.28	0.17	5.67	-14.50	13.72	0.09	2.72	-9.50	7.88
	0.80		0.00	6.33	-16.58	14.37	-0.02	9.77	-32.91	27.68	0.14	6.12	-16.09	17.59	0.04	2.43	-9.89	9.55
2		0.00	0.12	7.53	-22.17	18.69	0.27	9.23	-26.50	24.98	-0.02	7.10	-21.78	18.88	-0.29	4.00	-11.58	11.39
	0.40	0.40	-0.38	8.33	-21.35	14.91	0.38	9.01	-24.51	20.96	0.01	7.48	-18.03	18.58	-0.04	5.07	-18.45	16.56
		0.70	0.00	8.29	-17.62	18.62	0.05	9.66	-23.41	22.60	0.07	8.25	-20.28	18.43	-0.12	5.69	-19.78	18.21
		0.00	0.24	7.43	-15.59	16.64	-0.09	9.75	-36.49	26.37	0.44	7.28	-17.89	19.48	0.07	3.69	-13.42	15.51
	0.80	0.40	-0.09	8.24	-19.50	18.80	-0.71	10.65	-33.18	29.95	-0.02	7.83	-20.29	19.89	-0.10	4.37	-13.50	16.95
		0.70	0.65	8.94	-20.54	21.21	0.14	10.45	-31.34	26.31	0.44	8.75	-21.65	23.74	0.03	4.91	-16.90	15.23

Note. nuisance:number of nuisance dimension(s) present; vcor: correlations between primary dimensions; ncor: correlation between two nuisance dimensions

Table 15. Descriptive Statistics for Conditional e2UIRT (based on Percentage Scores) for All Crossed Conditions

1000 examinees																	
			10 items per dimension								20 items per dimension							
			low item discrimination				high item discrimination				low item discrimination				high item discrimination			
nuisance	vcor	ncor	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max
0	0.40	NA	-4.36	2.53	-12.10	3.13	-3.82	4.44	-14.60	10.28	-4.13	2.18	-12.59	1.38	-4.29	8.81	-26.67	14.53
	0.80		-4.42	2.86	-12.44	4.13	-5.46	5.67	-23.36	9.66	-4.41	3.09	-14.53	5.19	-5.01	10.55	-29.97	24.39
1	0.40	NA	-4.29	3.20	-12.37	3.26	-3.75	4.49	-17.95	9.26	-5.33	3.39	-15.97	4.03	-4.72	8.39	-26.68	23.76
	0.80		-3.61	3.75	-13.89	8.26	-4.77	5.69	-19.91	15.12	-4.80	3.79	-15.26	6.33	-5.76	10.28	-33.14	19.84
2	0.40	0.00	-3.59	4.04	-14.70	6.47	-4.28	5.72	-21.24	12.04	-5.38	4.16	-18.14	5.69	-5.23	8.09	-27.59	14.31
		0.40	-4.08	4.67	-17.09	9.36	-4.10	6.06	-25.24	11.97	-4.97	4.83	-17.62	7.75	-6.38	8.03	-28.28	11.55
		0.70	-3.90	5.37	-18.24	8.64	-5.88	5.78	-22.93	15.10	-5.50	4.89	-21.40	7.37	-5.76	8.45	-33.06	19.21
	0.80	0.00	-4.39	4.72	-16.60	9.95	-4.81	6.35	-20.68	15.30	-5.29	4.57	-15.47	9.66	-7.37	10.40	-39.37	21.59
		0.40	-4.82	5.14	-17.84	14.57	-4.91	6.54	-30.97	18.37	-6.17	4.92	-19.85	10.05	-5.03	10.39	-31.60	24.88
		0.70	-5.47	5.47	-18.93	9.32	-6.05	7.12	-29.04	14.05	-6.42	4.97	-19.79	8.37	-5.79	9.96	-32.83	21.59
5000 examinees																		
			10 items per dimension								20 items per dimension							
			low item discrimination				high item discrimination				low item discrimination				high item discrimination			
nuisance	vcor	ncor	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max
0	0.40	NA	-3.42	2.05	-10.22	4.64	-4.18	4.17	-15.15	5.92	-3.62	2.32	-10.00	3.34	-3.92	10.07	-27.27	19.66
	0.80		-3.91	2.95	-12.96	6.64	-3.44	6.04	-20.06	17.52	-3.36	2.93	-13.87	5.34	-4.13	13.34	-34.18	25.47
1	0.40	NA	-3.20	4.13	-14.47	7.35	-3.26	5.18	-19.20	10.02	-4.00	3.93	-15.23	6.35	-5.08	10.03	-26.76	20.16
	0.80		-2.97	4.32	-16.08	7.41	-4.14	6.96	-25.47	17.25	-4.48	4.31	-18.49	8.91	-5.78	12.91	-42.84	25.55
2	0.40	0.00	-3.31	5.10	-17.19	11.63	-4.75	6.19	-19.33	16.08	-4.36	5.05	-18.00	9.59	-5.74	9.77	-28.69	19.92
		0.40	-3.97	5.66	-20.06	9.83	-4.62	6.29	-19.93	16.65	-4.61	5.42	-17.34	11.91	-5.83	10.02	-34.15	17.12
		0.70	-5.46	6.00	-22.58	7.81	-3.45	6.91	-25.30	12.46	-4.20	5.95	-20.01	10.79	-5.00	10.05	-29.26	20.02
	0.80	0.00	-3.57	5.57	-18.50	9.62	-4.64	7.16	-25.20	18.30	-5.36	5.21	-18.58	11.04	-5.32	12.56	-37.75	27.94
		0.40	-4.37	6.08	-22.17	9.81	-4.41	7.63	-27.37	16.75	-5.34	5.70	-19.05	10.67	-6.59	12.36	-36.44	23.16
		0.70	-3.42	6.44	-18.12	13.62	-4.67	7.67	-26.70	15.02	-5.49	6.28	-25.24	13.90	-7.21	12.63	-35.88	22.18

Note. nuisance:number of nuisance dimension(s) present; vcor: correlations between primary dimensions; ncor: correlation between two nuisance dimensions

Table 16. Descriptive Statistics for Conditional e3MIRT (based on Percentage Scores) for All Crossed Conditions

1000 examinees																	
			10 items per dimension								20 items per dimension							
			low item discrimination				high item discrimination				low item discrimination				high item discrimination			
nuisance	vcor	ncor	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max
0	0.40	NA	-0.18	2.27	-5.47	4.38	-0.09	1.26	-3.93	3.27	-0.07	1.24	-2.66	3.18	-0.03	0.64	-1.55	2.03
	0.80		0.02	1.98	-4.18	4.93	-0.03	0.99	-2.80	2.85	0.01	0.98	-2.92	3.23	-0.02	0.57	-1.71	1.66
1	0.40	NA	-0.05	1.58	-3.66	3.86	-0.13	0.88	-2.46	2.24	-0.06	0.72	-2.43	1.59	-0.08	0.53	-2.11	1.56
	0.80		0.00	1.50	-4.54	3.60	-0.01	0.90	-2.28	2.06	-0.02	0.79	-2.20	2.32	-0.02	0.53	-1.84	1.69
2	0.40	0.00	-0.12	1.22	-2.88	3.21	-0.13	0.83	-2.31	2.58	-0.02	0.64	-1.91	2.18	-0.01	0.44	-1.64	1.19
		0.40	-0.06	1.02	-3.47	2.15	-0.09	0.77	-2.48	2.35	-0.04	0.57	-1.82	1.56	0.02	0.49	-1.74	1.61
		0.70	-0.07	0.97	-3.18	3.37	-0.06	0.69	-2.12	2.16	0.00	0.55	-1.69	2.03	0.01	0.43	-1.71	1.46
	0.80	0.00	-0.08	1.16	-3.06	3.09	-0.12	0.81	-2.82	2.48	-0.01	0.69	-2.07	2.24	0.00	0.49	-1.82	1.47
		0.40	0.00	1.01	-2.67	2.94	0.01	0.75	-2.92	2.67	-0.02	0.58	-2.19	1.80	0.02	0.46	-1.27	1.37
		0.70	-0.03	0.92	-3.72	2.07	-0.03	0.74	-2.35	1.83	0.02	0.57	-1.80	1.92	0.01	0.48	-2.03	1.54
5000 examinees																		
			10 items per dimension								20 items per dimension							
			low item discrimination				high item discrimination				low item discrimination				high item discrimination			
nuisance	vcor	ncor	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max
0	0.40	NA	-0.11	2.72	-6.19	4.83	-0.12	1.35	-3.73	3.02	-0.02	1.35	-3.12	3.18	-0.03	0.70	-2.39	2.19
	0.80		-0.01	2.26	-5.02	4.94	-0.09	1.18	-3.31	3.34	-0.02	1.20	-3.29	2.79	-0.02	0.60	-2.44	1.80
1	0.40	NA	-0.09	1.77	-4.31	4.41	-0.12	1.19	-3.45	2.87	-0.01	0.86	-2.70	2.36	-0.03	0.56	-2.28	1.94
	0.80		-0.06	1.60	-3.83	4.41	-0.03	0.99	-3.37	2.67	-0.01	0.83	-1.99	2.15	-0.03	0.53	-1.81	2.01
2	0.40	0.00	0.03	1.40	-3.96	3.51	-0.04	0.93	-2.42	2.08	-0.01	0.68	-2.43	1.90	-0.06	0.47	-2.23	1.50
		0.40	-0.08	1.27	-3.46	3.19	-0.10	0.85	-3.40	1.99	-0.04	0.59	-2.10	1.92	-0.02	0.43	-1.95	1.43
		0.70	-0.13	1.11	-3.25	3.41	-0.05	0.76	-2.32	2.31	-0.01	0.54	-1.99	1.77	-0.02	0.41	-1.79	1.64
	0.80	0.00	0.00	1.37	-3.44	4.02	-0.02	0.89	-2.25	2.63	0.03	0.71	-2.15	1.92	0.00	0.45	-1.53	1.40
		0.40	-0.05	1.22	-3.13	3.76	-0.07	0.78	-2.61	2.13	-0.03	0.63	-2.13	1.73	0.02	0.44	-2.04	1.63
		0.70	0.04	1.07	-2.24	2.81	-0.02	0.78	-1.86	2.77	0.03	0.60	-1.69	1.94	0.02	0.42	-1.50	1.53

Note. nuisance:number of nuisance dimension(s) present; vcor: correlations between primary dimensions; ncor: correlation between two nuisance dimensions

Table 17. Descriptive Statistics for Conditional e3UIRT (based on Percentage Scores) for All Crossed Conditions

1000 examinees																	
			10 items per dimension								20 items per dimension							
			low item discrimination				high item discrimination				low item discrimination				high item discrimination			
nuisance	vcor	ncor	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max
0	0.40	NA	3.68	6.65	-14.13	18.04	3.89	10.03	-20.24	28.49	3.87	5.27	-8.16	15.12	4.26	9.38	-15.59	28.12
	0.80		4.27	7.30	-11.92	18.89	4.82	12.02	-22.12	32.44	4.40	6.71	-11.83	21.89	5.03	10.97	-24.30	30.74
1	0.40	NA	4.09	8.96	-18.76	23.59	3.97	10.71	-18.71	26.94	5.24	7.86	-14.01	25.43	4.75	10.27	-27.03	29.84
	0.80		3.40	9.49	-21.50	23.54	4.71	12.65	-23.30	30.37	4.86	9.03	-16.97	27.86	5.75	11.88	-21.14	38.49
2	0.40	0.00	3.26	10.64	-20.12	25.12	3.40	12.50	-27.23	37.44	5.44	9.88	-16.47	34.21	5.33	11.21	-21.37	34.26
		0.40	4.09	11.62	-25.29	30.50	4.55	13.13	-21.48	35.96	4.93	10.86	-19.24	34.41	6.77	11.73	-22.04	34.60
		0.70	3.74	13.12	-22.90	35.30	6.10	13.34	-29.26	35.51	5.73	11.36	-30.35	38.23	5.64	12.71	-27.04	47.00
	0.80	0.00	4.18	11.95	-26.85	30.24	4.95	14.11	-38.24	36.02	5.13	10.17	-18.12	29.91	7.11	12.87	-25.50	42.20
		0.40	5.21	12.65	-30.93	36.68	5.18	14.65	-36.59	40.01	6.26	10.96	-21.49	33.07	5.42	13.65	-30.82	38.20
		0.70	5.51	13.53	-23.83	40.45	5.58	15.25	-30.41	36.11	6.12	11.58	-27.46	33.84	5.79	13.48	-33.93	38.51
5000 examinees																		
			10 items per dimension								20 items per dimension							
			low item discrimination				high item discrimination				low item discrimination				high item discrimination			
nuisance	vcor	ncor	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max
0	0.40	NA	3.26	7.72	-12.18	17.18	3.98	11.75	-21.16	28.51	3.78	6.73	-10.65	19.27	3.98	10.84	-20.29	30.85
	0.80		4.05	8.92	-14.96	23.22	3.31	15.01	-29.24	39.47	3.41	7.88	-13.45	22.28	4.11	13.93	-26.75	39.06
1	0.40	NA	3.05	11.49	-19.15	23.89	3.16	13.72	-30.45	33.37	4.15	10.17	-16.99	26.73	5.14	12.86	-24.42	32.59
	0.80		2.91	11.98	-22.60	25.91	4.08	16.64	-38.22	37.03	4.61	10.84	-19.65	29.93	5.79	15.12	-26.48	43.86
2	0.40	0.00	3.46	13.78	-36.86	32.26	4.98	15.77	-25.59	40.53	4.33	12.50	-28.48	33.59	5.40	13.77	-27.33	36.98
		0.40	3.51	15.01	-28.04	31.54	4.90	15.63	-39.11	38.36	4.58	13.22	-23.12	32.56	5.76	14.99	-27.30	44.41
		0.70	5.33	15.18	-27.38	36.93	3.44	16.87	-30.63	41.77	4.26	14.48	-31.55	37.29	4.86	15.60	-29.37	43.95
	0.80	0.00	3.81	14.05	-24.55	30.95	4.53	16.97	-33.28	44.58	5.83	12.78	-20.32	33.55	5.39	16.00	-27.27	44.59
		0.40	4.23	15.18	-28.26	38.72	3.63	18.17	-38.71	38.70	5.29	13.75	-24.54	35.80	6.51	16.50	-30.44	40.92
		0.70	4.12	16.15	-30.58	37.77	4.78	18.14	-35.64	44.59	5.97	15.16	-26.80	38.53	7.26	17.26	-33.78	45.78

Note. nuisance:number of nuisance dimension(s) present; vcor: correlations between primary dimensions; ncor: correlation between two nuisance dimensions

Results for Research Question 1a

In this question, I ask more specifically about the amount of ERF-based residual produced by IRT models when nuisance dimensions are present. The results are provided in Table 18 and Figure 10.

An examination of the e2MIRT shows that the MIRT model tends to produce small errors. As shown in the three graphs in the bottom of Figure 10, MIRT model tend to underestimate low scores and overestimate high scores. The trend is more manifest with the presence of at least a nuisance dimension. This is consistent with what Reese (1995) observed, although Reese examined LID using Yen's Q3 (1984, 1993) when calibrating with the 3PL model. Reese found that the LID caused low scores to be underestimated and high scores to be overestimated, especially in sets of items that exhibit high LID. This effect caused the score distribution to spread out at the tails and flatten in the middle.

An examination of the e2UIRT shows that the UIRT model (combination of 2PL and GRM) tends to produce more errors compared to the MIRT model. Specifically, the unidimensional model underestimate examinees' raw scores even when there is no nuisance dimension. It underestimates examinees' raw scores even more (its worst performance) with the presence of two nuisance dimensions. This is clearly evident in the bottom right graph shown in Figure 10.

In terms of model-data fit residuals, a multidimensional model tends to fit the data better on average than the unidimensional model. UIRT tends to fit the data badly on average. e3MIRT deviates below zero on average regardless of the

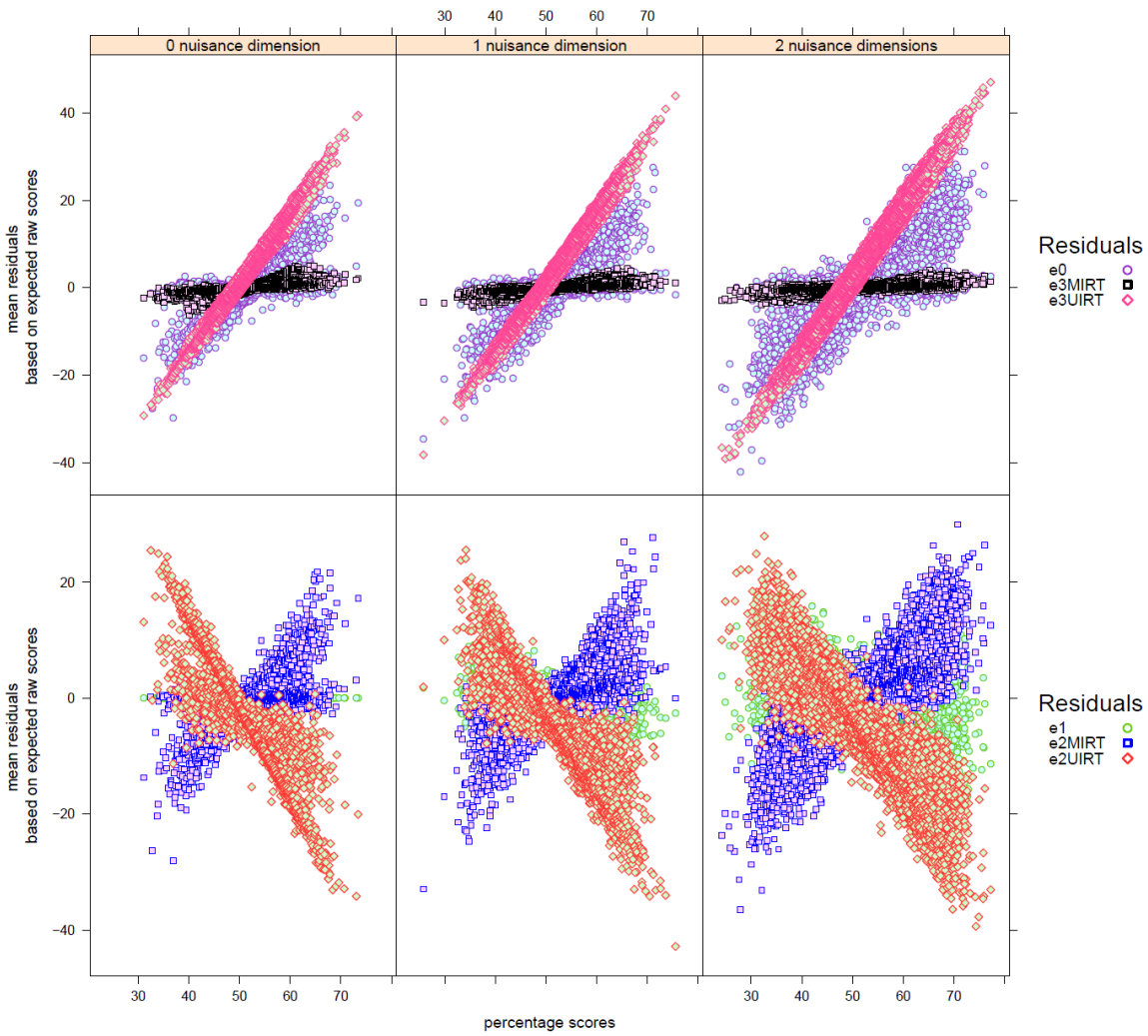
presence of nuisance dimensions or the number of nuisance dimensions. Overall, the MIRT model tends to slightly deviate from the examinees' observed response. e3UIRT deviates above zero on average with the presence of at least one nuisance dimension. Specifically, it deviates more on average with the presence of two nuisance dimensions as can be seen in the top-right graph in Figure 10. The UIRT model tends to deviate greatly from the examinees' observed response data.

In conclusion, the multidimensional model produces less error on average than the unidimensional model (estimated minus the intended dimensionality). The MIRT model produces a larger error on average with the presence of one nuisance dimension. With two nuisance dimensions, the error from MIRT is still large but is smaller on average than when only one nuisance dimension exists.

Table 18. Descriptive Statistics for e2 and e3 Residuals from MIRT and UIRT Calibrations Given the Amount of Nuisance Dimension

presence of nuisance dimension(s)	e2MIRT		e2UIRT		e3MIRT		e3UIRT	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
0	-0.029	4.190	-4.108	7.424	-0.044	1.293	4.035	10.342
1	0.045	5.497	-4.552	7.437	-0.043	0.965	4.554	12.030
2	0.035	7.042	-5.176	7.783	-0.021	0.735	5.190	14.174

Figure 10. Distributions of Conditional Mean Residuals (based on Percentage Scores) Given the Amount of Nuisance Dimension



Results for Research Question 1b

This question asks more specifically about the amount of ERF-based residuals produced by IRT models when correlations between nuisance dimensions vary. The results are presented in Table 19 and Figure 11.

The multidimensional model produces less error on average than the unidimensional model. Examination of e2MIRT shows MIRT model produces a larger error on average when the nuisance dimensions are more associated with each other (nuisance correlations of .40 and .70). In these cases, the four dimensional MIRT model might detect the correlated nuisance dimensions as another new dimension. However, the errors seem much more inconsequential compared to the errors produced by the UIRT model. Based on e2UIRT, I found that the UIRT model tends to underestimate examinees' raw scores on average regardless of the correlations between nuisance dimensions. It underestimates examinees' raw scores the worst when the nuisance dimensions are more associated with each other (nuisance correlation of .70).

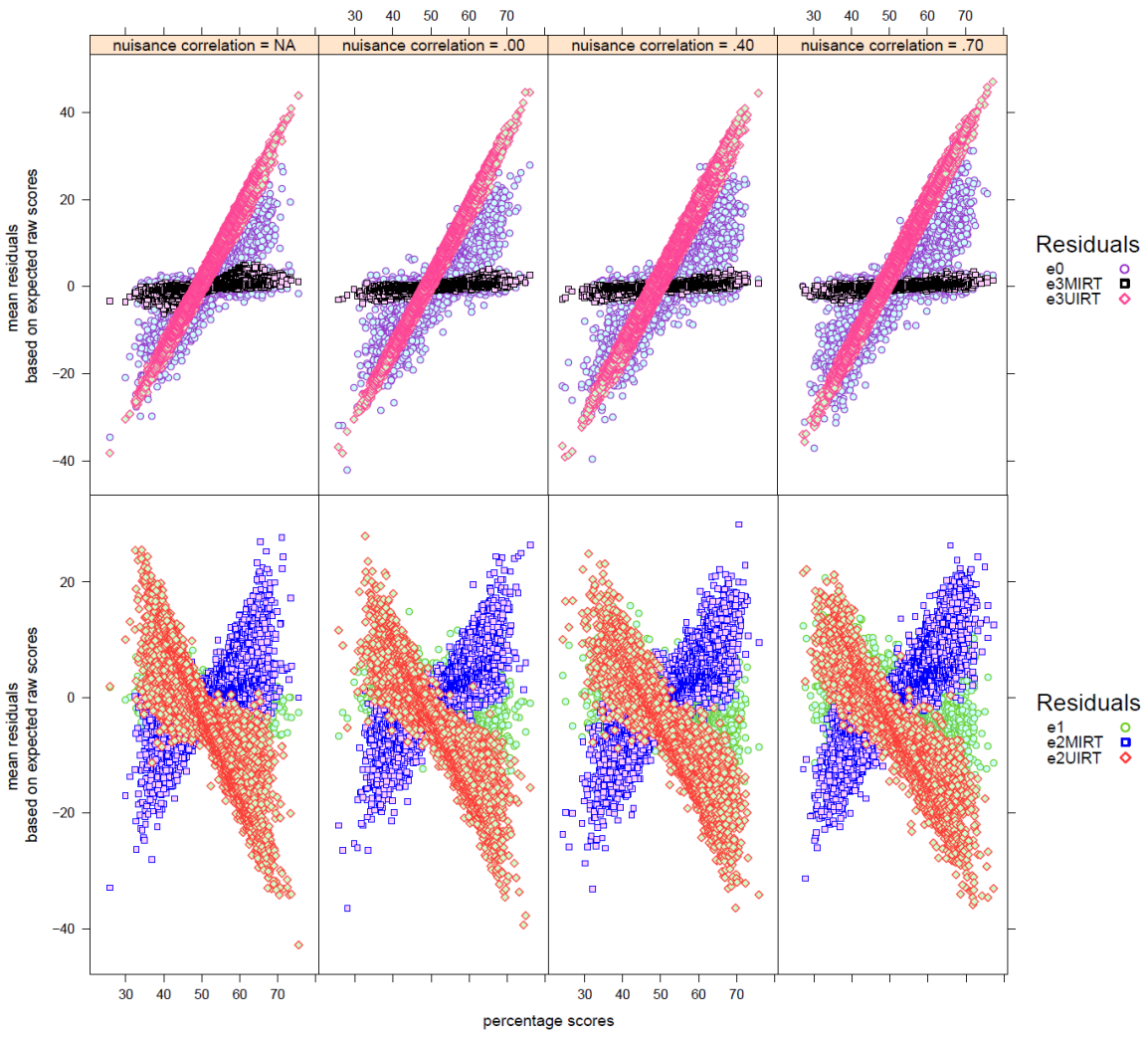
In terms of model-data fit residuals, the multidimensional model tends to fit the data better on average than the unidimensional model. The plots at the top of Figure 11 also suggest this same conclusion. e3MIRT deviates slightly below zero on average regardless of the associations between nuisance dimensions. It also deviates further below zero on average when the nuisance dimensions are slightly associated with each other (nuisance correlation of .40). e3UIRT deviates above zero on average regardless of the associations between nuisance dimensions. It

deviates further, above zero on average when the nuisance dimensions are more associated with each other (nuisance correlations of .40 and .70).

Table 19. Descriptive Statistics for e2 and e3 Residuals from MIRT and UIRT Calibrations Given the Strength of Correlations between Nuisance Dimensions

correlations between nuisance dimensions	e2MIRT		e2UIRT		e3MIRT		e3UIRT	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
NA	0.011	4.938	-4.347	7.434	-0.043	1.128	4.315	11.287
0.00	-0.002	6.478	-5.001	7.589	-0.026	0.808	4.973	13.379
0.40	0.048	7.031	-5.184	7.760	-0.028	0.725	5.204	14.156
0.70	0.056	7.527	-5.326	7.974	-0.010	0.672	5.372	14.875

Figure 11. Distributions of Conditional Mean Residuals (based on Percentage Scores) Given the Strength of Correlations between Nuisance Dimensions



Results for Research Question 1c

This question asks about the amount of ERF-based residuals produced by IRT models when correlations between primary dimensions vary. The results are summarized in Table 20 and Figure 12.

An examination of the parameter-estimation residuals suggests that the multidimensional model produces less error on average than the unidimensional model. An examination of e2MIRT shows that MIRT model produces a very small error on average regardless of the associations between primary dimensions. An examination of e2UIRT also shows similar pattern in which UIRT model tends to underestimate examinees' raw scores on average regardless of the correlations between primary dimensions. As shown in the bottom-right graph in Figure 12, the e2UIRT are spreading out across the percentage scores. The UIRT model tends to underestimate examinees' raw scores more poorly as the primary dimensions are more associated with each other (correlation of .80). Such finding may be due to the compensatory nature of the generating model. UIRT model is not able to perceive the effect of all four closely-associated dimensions and detects the items in the four dimensions as similar to each other.

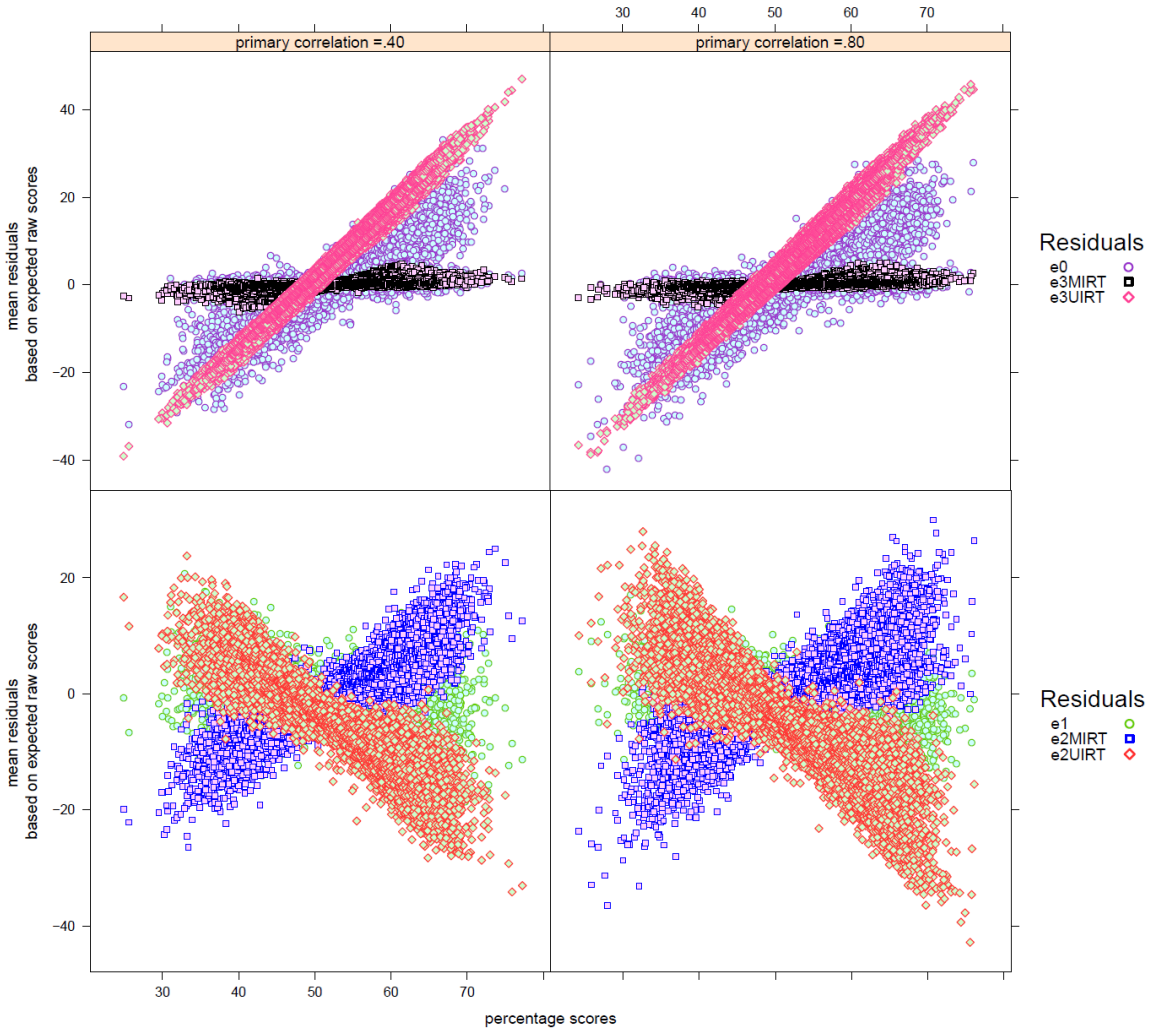
In terms of model-data fit residuals, the multidimensional model tends to fit the data better on average than the unidimensional model. e3MIRT deviates below zero on average regardless of the associations between primary dimensions. It deviates further below zero on average when the primary dimensions are more associated with each other (correlation of .80). On the other hand, e3UIRT deviates

above zero on average regardless of the associations between primary dimensions. It deviates further above zero on average when the primary dimensions are more associated with each other (correlation of .80).

Table 20. Descriptive Statistics for e2 and e3 Residuals from MIRT and UIRT Calibrations Given the Strength of Correlations between Primary Dimensions

correlations between primary dimensions	e2MIRT		e2UIRT		e3MIRT		e3UIRT	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
0.40	0.024	6.159	-4.629	6.823	-0.048	0.925	4.605	12.524
0.80	0.029	6.595	-5.126	8.376	-0.011	0.859	5.144	13.878

Figure 12. Distributions of Conditional Mean Residuals (based on Percentage Scores) Given the Strength of Correlations between Primary Dimensions



Results for Research Question 1d

This question asks about the amount of ERF-based residuals produced by IRT models when item discrimination levels on primary dimensions vary. Table 21 and Figure 13 summarize and illustrate the findings.

The multidimensional model produces less error on average than the unidimensional model. Examination of e2MIRT shows that the MIRT model produces a small error on average regardless of the item discrimination levels. However, with high item discrimination, the error produces by MIRT model is very small. An examination of e2UIRT shows that the UIRT model tends to underestimate examinees' raw scores on average regardless of the item discrimination levels. Again, this model underestimates examinees' raw scores the worst when the item discrimination level is high (between 1.30 and .90). This is evident in the bottom-right graph in Figure 13. One plausible reason for this is that the UIRT model is not able to capture the high item discriminations (slopes) in each of the compensatory dimensions in the generating model. Because the generating slopes are in separate dimensions, the UIRT model fails to estimate the slope parameters correctly.

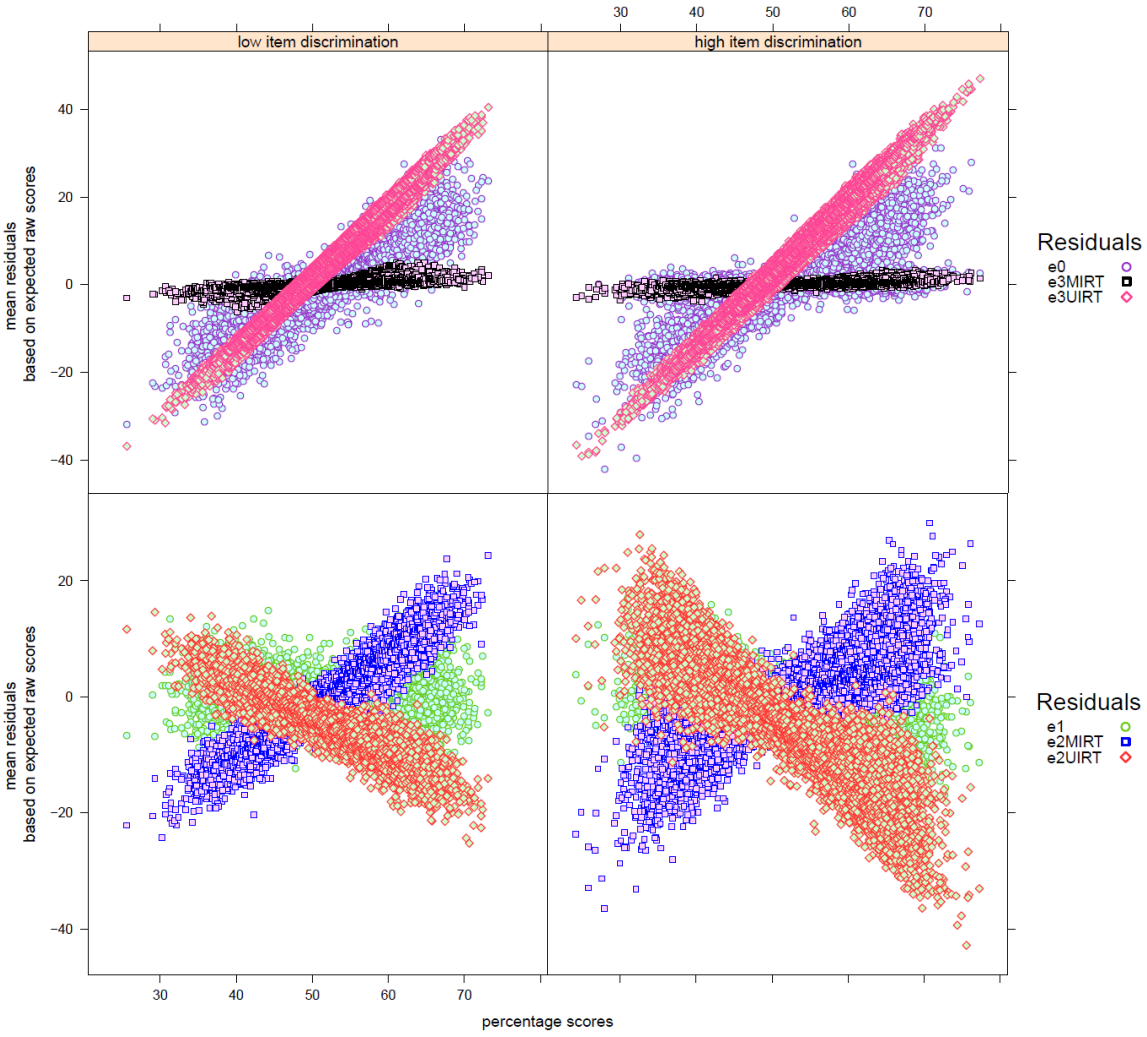
An examination of the model-data fit residuals shows that the multidimensional model tends to fit the data better than the unidimensional model on average. e3MIRT deviates below zero on average regardless of the item discrimination levels. Specifically, the residuals tend to deviate further below zero on average when the item discrimination level is high. e3UIRT deviates above zero

on average regardless of the item discrimination levels and deviates further above zero on average when item discrimination level is high.

Table 21. Descriptive Statistics for e2 and e3 Residuals from MIRT and UIRT Calibrations Given Different Item Discrimination Levels on the Primary Dimensions

item discrimination	e2MIRT		e2UIRT		e3MIRT		e3UIRT	
levels	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
all low	0.054	6.683	-4.572	4.884	-0.026	1.085	4.600	11.928
all high	0.004	6.132	-5.152	9.386	-0.031	0.689	5.125	14.257

Figure 13. Distributions of Conditional Mean Residuals (based on Percentage Scores) Given Different Item Discrimination Levels on the Primary Dimensions



Results for Research Question 1e

This question asks about the amount of ERF-based residuals produced by IRT models when the test length varies. The results are shown in Table 22 and Figure 14.

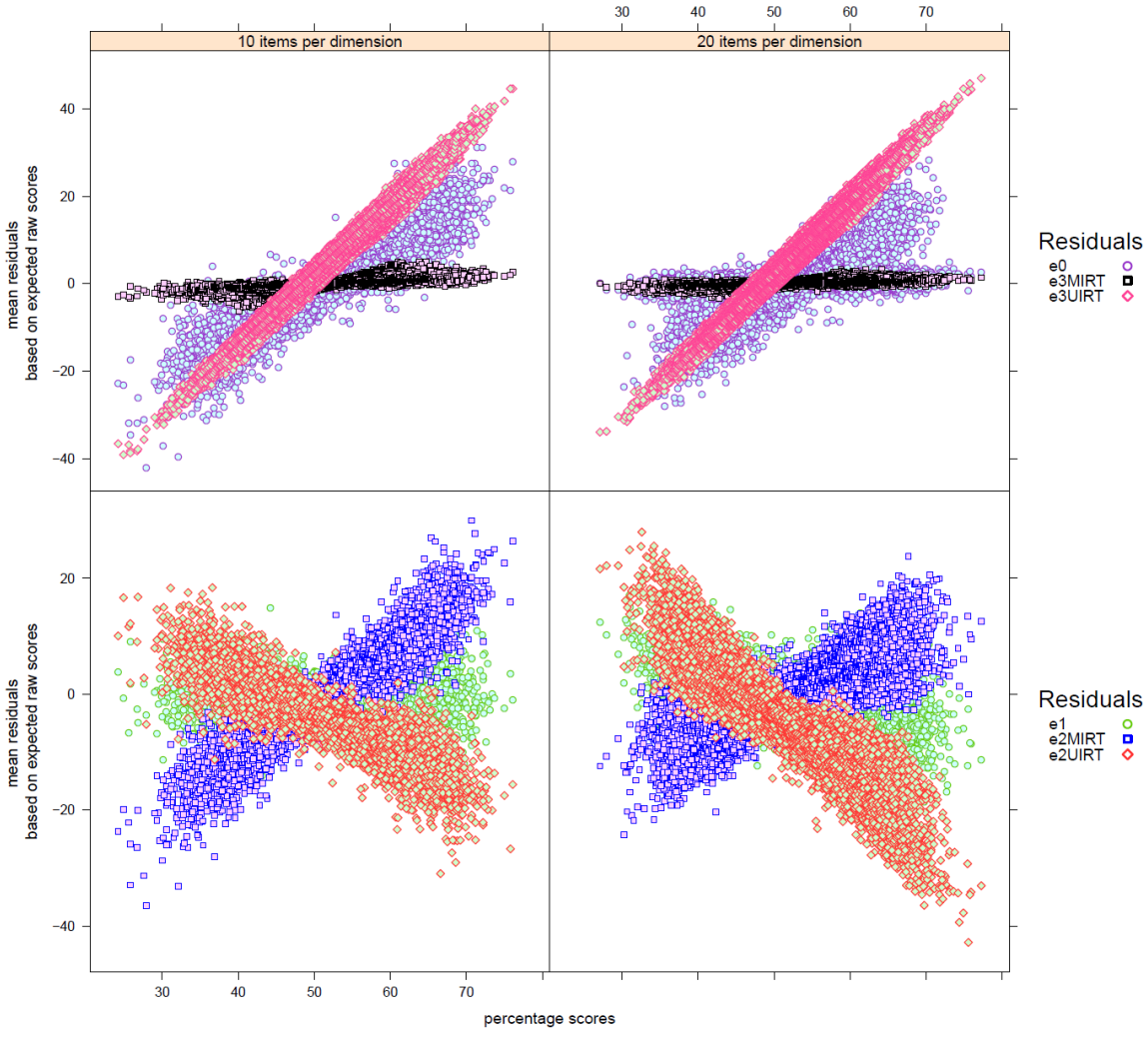
On average, the multidimensional model produces less error than the unidimensional model. An examination of e2MIRT suggests that the MIRT model produces a small error on average regardless of the test length. However, the MIRT model produces less error in a test with more items per each dimension than in a test with fewer items (see also bottom-right graph in Figure 14). An examination of the e2UIRT shows that the UIRT model tends to underestimate examinees' raw scores on average regardless of the test length.

In terms of model-data fit residuals, the multidimensional model tends to fit the data better on average than the unidimensional model. The e3MIRT tends to deviate below zero on average regardless of the test length. The residuals deviate further below zero on average when there are fewer items in a test. The e3UIRT deviates above zero on average regardless of the test length. It deviates further above zero on average when the number of items in a test increases since the nuisance dimension(s) was(were) set to load onto all items regardless of the test length. In other words, the assumptions of unidimensionality and local independence were violated further when UIRT models are used in calibration.

Table 22. Descriptive Statistics for e2 and e3 Residuals from MIRT and UIRT Calibrations Given Number of Items in each Primary Dimensions

number of items per primary dimension	e2MIRT		e2UIRT		e3MIRT		e3UIRT	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
10 items	-0.001	7.953	-4.259	5.740	-0.057	1.210	4.201	13.891
20 items	0.043	5.256	-5.257	8.589	-0.012	0.633	5.288	12.838

Figure 14. Distributions of Conditional Mean Residuals (based on Percentage Scores) Given Number of Items in each Primary Dimensions



Results for Research Question 1f

This question asks about the amount of ERF-based residuals produced by IRT models over varying sample sizes. The results are summarized in Table 23 and Figure 15.

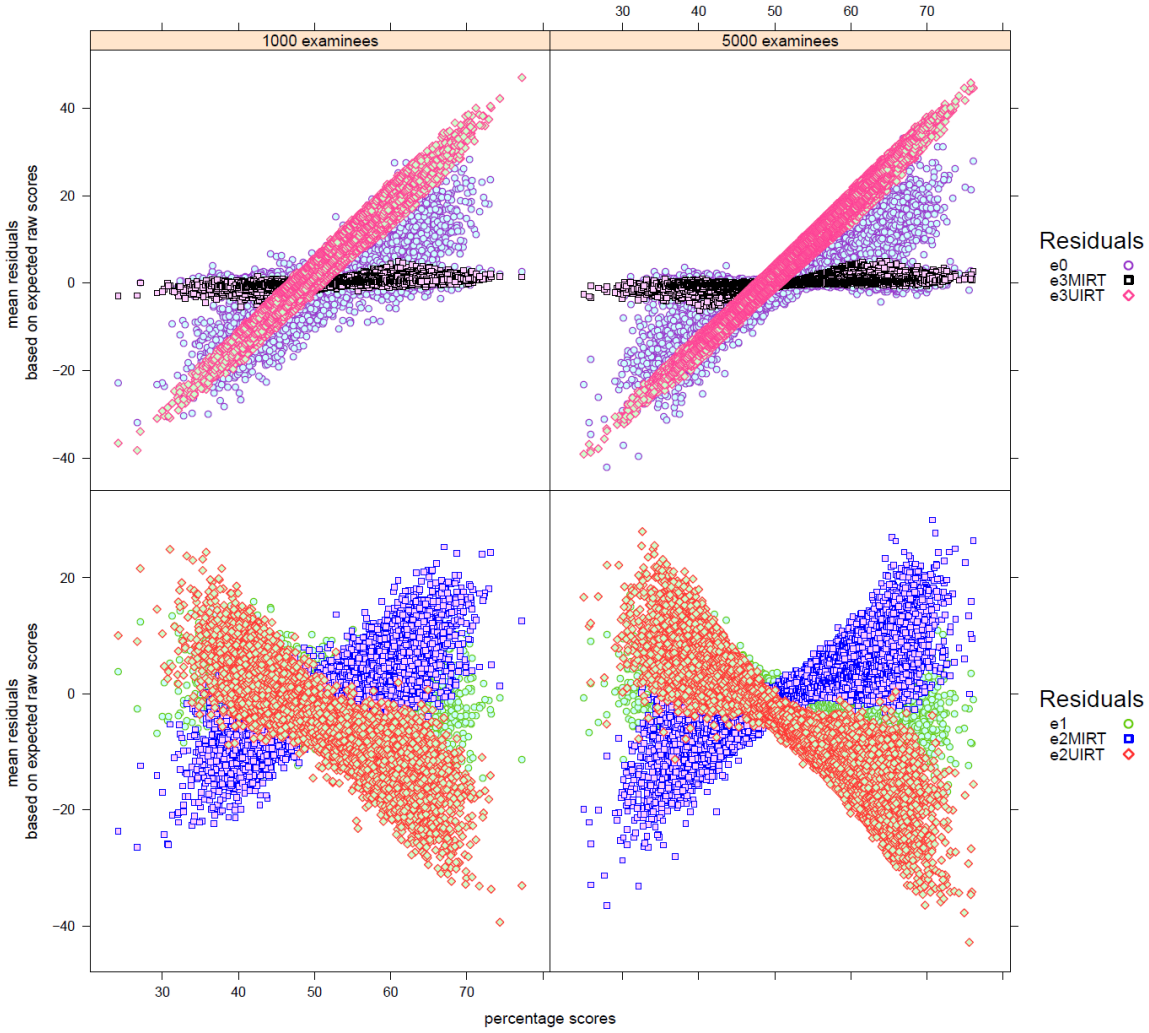
An examination of the parameter estimation residuals shows that the multidimensional model tends to produce less error on average regardless of the number of examinees. Similarly, an examination of e2UIRT suggests that the UIRT model tends to underestimate examinees' raw scores on average regardless of the number of examinees. However, both e2MIRT and e2UIRT tend to shrink closer to zero with 5,000 examinees.

In terms of model-data fit residuals, the e3MIRT tends to deviate slightly below zero, on average, regardless of the number of examinees. Specifically, it deviates further below zero on average with fewer examinees. e3UIRT deviates above zero on average regardless of number of examinees and deviates further above zero on average with fewer examinees. However, the multidimensional model produces less error on average than the unidimensional model regardless of the sample size.

Table 23. Descriptive Statistics for e2 and e3 Residuals from MIRT and UIRT Calibrations Given Different Sample Sizes

sample sizes	e2MIRT		e2UIRT		e3MIRT		e3UIRT	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1,000 examinees	0.004	5.818	-5.128	6.798	-0.031	0.845	5.101	11.544
5,000 examinees	0.043	6.774	-4.712	8.248	-0.027	0.924	4.728	14.359

Figure 15. Distributions of Conditional Mean Residuals (based on Percentage Scores) Given Different Sample Sizes



Results from the Factorial ANOVA of the Estimated ERF-Based Residuals

In this section, I will only report my findings from the factorial ANOVA analyses when the two-way interaction effect for a given combination of factors is significant. All three-way interaction effects for all possible combinations of factors on e2MIRT, e3MIRT, and e3UIRT were not statistically significant and thus are not reported.

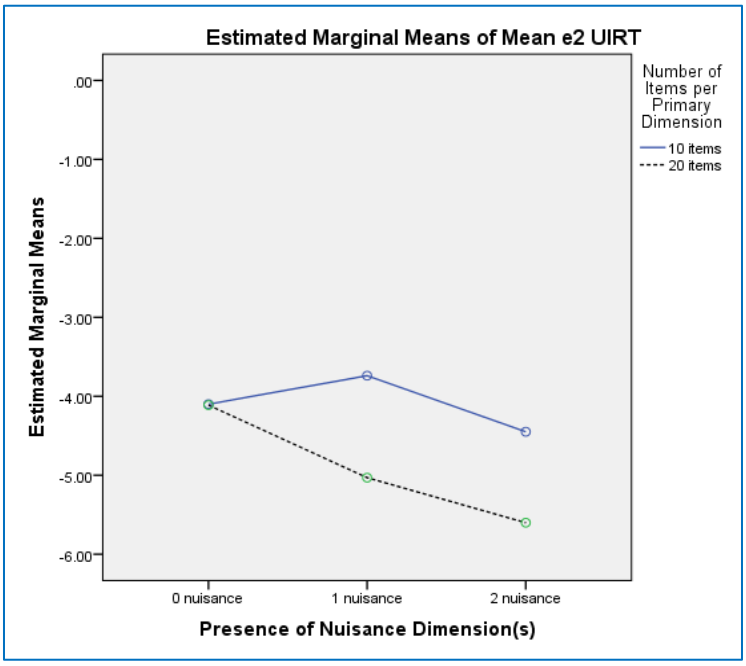
For the two-way ANOVA, I conducted a 3 x 2 factorial ANOVA between the presence of nuisance dimensions and the number of items for each primary dimension on the e2UIRT. The overall analysis indicates that the main effects for the two factors along with their interaction effects are statistically significant. In particular, the interaction effects are statistically significant but yield a very small effect size ($F(2, 28916)= 10.849$, $p<.001$, $\eta^2=.001$). Given the significant interaction effect, I conducted an analysis of the simple effects. As can be seen in the simple effects table in Table 24 and in the profile plot in Figure 16, it is evident that the e2UIRT significantly decreases when there is no nuisance dimension present especially with 20 items per subtest. There is a small effect size for test length when one nuisance dimension is present ($F(1, 28916) = 36.102$, $p < .001$, $\eta^2 = .001$). Similarly, there is also a small effect size for test length with the presence of two nuisance dimensions ($F(1, 28916) = 99.706$, $p < .001$, $\eta^2 = .003$).

Table 24. Summary Table for Two-Factorial ANOVA on the Conditional Mean e2UIRT

Overall Analysis						
Source	*SS*	*df*	*MS*	*F*	*p*	η^2
Nuisance Dimension	3803.822	2	1901.911	32.541	.000**	0.002
Number of Item	3106.563	1	3106.563	53.152	.000**	0.002
2-Way Interaction	1268.161	2	634.081	10.849	.000**	0.001
Error	1690055	28916	58.447			
Total	2393877	28922				
Analysis of Simple Effects						
Source	*SS*	*df*	*MS*	*F*	*p*	η^2
Number of Item at 0 nuisance dimension	0.147	1	0.147	0.003	0.96	0
Number of Item at 1 nuisance dimension	2110.034	1	2110.034	36.102	.000**	0.001
Number of Item at 2 nuisance dimension	5827.549	1	5827.54	99.706	.000**	0.003
Error	1690055	28916	58.447			

**p<.001

Figure 16. Profile Plots for Two-Factorial ANOVA on the Conditional Mean e2UIRT



Summary of Overall Findings for Research Question 1

I conclude from my comparison of parameter estimation residuals (e2MIRT & e2UIRT) that the unidimensional model (2PL&GRM) produces large estimation errors when the model of interest is multidimensional and compensatory in nature (M3PL&MGR), regardless of the testing configuration. The multidimensional model (M2PL&MGR), as expected, produces small estimation errors when the model of interest (M3PL&MGR) is of a similar nature. The small amount of error may be due to the guessing induced in the data by the M3PL model.

The comparison of model-data fit residuals (e3MIRT & e3UIRT) shows that the unidimensional model tends to consistently deviate further away from the examinees' observed response data regardless of the testing configuration. The multidimensional model tends to fit the observed response data better with only a slight misfit.

The two-way interaction effect on e2UIRT is statistically significant as the number of items per subtest differs with the presence of nuisance dimension(s), but the effect size is very small. My test of the simple effects and examination of the profile plot for the two factors indicate that the e2UIRT tends to be smaller with the existence of one nuisance dimension in a shorter test. However, its effect size is very small.

Results for Research Question 2: Examination of Bias-Induced Residual Covariance ($r_{e1i,e1h}$)

In my second research question, I examined how the amount of modeled, bias-induced, residual covariance was impacted by six testing configurations:

a. The presence of nuisance dimension(s)?
b. The strength of correlations between nuisance dimensions?
c. The strength of correlations between primary dimensions?
d. Changes in item discrimination levels on the primary dimensions?
e. The number of items in each primary dimension?
f. Changes in sample size?

To answer this question, I first summarize the modeled, biased-induced, residual covariance for each sub-research question by providing the mean, standard deviation, minimum, and maximum residual covariance. I provide a series of factorial ANOVA models to determine whether there are mean differences in the levels of conditions. Again, I based the effect size heuristics for η^2 on Gray & Kinnear (2012) in which *small effect size* ranges from .01 to less than .06, *medium effect size* ranges from .06 to less than .14, and *large effect size* occurs when η^2 is equal to or greater than .14.

Descriptive Statistics for All Study Conditions

First, I present the descriptive statistics of the conditional mean for bias-induced residuals (e1) in Table 25. Next, I provide the descriptive statistics of the correlations for the bias-induced residuals ($r_{e1_i,e1_h}$) in Table 26. I will discuss the interpretations of the results for the magnitude and correlations of e1 in an integrated fashion. In Figures 17(a) and 17(b), I illustrate the distribution of the bias-induced residual correlations using box and whisker plots. These distributions are closely related to the descriptive statistics in Table 26. Figure 17(a) displays the crossed conditions for five testing configurations and presents the distribution of residual correlations when two nuisance dimensions exist with 1,000 examinees, with different test lengths, different levels of item discrimination, when correlations among primary dimensions vary, and when the correlations among the two nuisance dimensions vary. Figure 17(b) provides a similar depiction but with 5,000 examinees. In each figure, there is a vertical line crossing at correlation of .90 to mark a high residual correlation value to enable better distinctions between each of the plots.

In Table 25, with the absence of nuisance dimension, the conditional bias-induced residuals (e1) influence the results for baseline analyses, explaining the zero statistics values. Similarly, when there is no nuisance dimension affecting the items, the residual covariances ($r_{e1_i,e1_h}$) in Table 26 are all zero. In the baseline analysis, zero covariance is expected, regardless of the study conditions. These

findings are consistent with the performance of other LID indices described in the literature such as Yen's Q3 (e.g., Yen, 1993; see also Goodman, 2008; Lee, 2004).

As expected, with the presence of one nuisance dimension, the magnitude of e1 and $r_{e1_i,e1_h}$ are now non-zero. The magnitude of e1 changes in both negative and positive directions depending on the conditioning variables. The residual correlations ($r_{e1_i,e1_h}$) are now larger in magnitude than those of the baseline analysis. For both sample sizes of 1,000 and 5,000, the $r_{e1_i,e1_h}$, on average, are smaller when the discrimination levels for all primary dimensions are high. This finding provides some initial indication that high discrimination levels of items in a test will tend to reduce both e1 and the e1 correlations. Moreover, when discrimination levels are high, the e1 and e1 correlations tend to decrease as the number of items in each dimension increase. Again, this is consistent across sample sizes (see also Figures 17(a) and 17(b)). A similar pattern occurs when the discrimination levels are low for 5000 examinees. As the primary dimensions in the tests have greater degrees of association with one another, the e1 correlations of the items tend to increase despite the high item discrimination.

Similar findings result when two unassociated nuisance dimensions are present, particularly when the correlation among the primary dimensions of interest is .40. With these conditions, the e1 correlations tend to decrease with high item discrimination levels in the primary dimensions. As the primary dimensions become more highly associated with one another, such a trend is no longer

observed. In general, with the presence of two nuisance dimensions in the test items, a consistent trend can be observed for e1 correlations when the nuisance correlations vary across number of sample sizes: as the number of examinees increases, the e1 correlations decrease for all conditions of nuisance correlations.

Figures 17(a) and 17(b) display a similar pattern in which the mean e1 residual correlations increase with low item discrimination in the primary dimensions, regardless of the amount of nuisance dimension. A brief examination of Figure 17(b) indicates that a different amount of correlation between two nuisance dimensions does not have an apparent effect on the residual correlations.

Table 25. Descriptive Statistics for Conditional e1 (based on Percentage Scores) for All Crossed Conditions

1000 examinees																	
			10 items per dimension								20 items per dimension							
nuisance	vcor	ncor	low item discrimination				high item discrimination				low item discrimination				high item discrimination			
			mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max
0	0.40	NA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.80		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.40	NA	-0.17	1.72	-6.24	5.21	-0.10	1.52	-4.74	4.63	-0.09	1.90	-6.53	8.12	-0.05	2.26	-7.25	7.20
	0.80		0.01	1.54	-4.43	7.15	-0.11	1.69	-5.66	4.97	-0.15	2.03	-10.11	5.83	-0.01	2.38	-7.27	7.88
2	0.40	0.00	-0.05	2.56	-9.36	14.90	0.05	2.41	-9.91	9.62	0.10	3.23	-12.39	11.07	-0.04	3.82	-11.33	13.67
		0.40	-0.18	3.11	-14.13	7.96	0.21	2.91	-7.63	10.15	0.13	3.60	-10.47	13.96	-0.32	4.60	-16.23	15.82
		0.70	-0.04	3.40	-12.68	9.39	0.01	3.08	-9.23	10.19	0.27	3.61	-13.42	13.35	0.11	4.90	-14.06	14.56
	0.80	0.00	0.18	2.31	-8.94	7.38	0.18	2.43	-7.31	9.88	0.17	2.81	-8.79	11.56	0.25	3.34	-12.81	9.45
		0.40	-0.69	2.96	-12.20	10.03	-0.29	2.94	-9.89	10.04	-0.09	3.32	-12.24	10.60	-0.35	4.35	-14.59	14.48
		0.70	0.05	3.17	-10.43	12.47	0.07	3.25	-14.38	11.73	0.08	3.87	-11.64	13.43	-0.03	4.57	-14.80	15.19

5000 examinees																		
			10 items per dimension								20 items per dimension							
nuisance	vcor	ncor	low item discrimination				high item discrimination				low item discrimination				high item discrimination			
			mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max
0	0.40	NA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.80		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.40	NA	-0.12	1.30	-5.01	5.78	-0.07	1.16	-7.14	4.00	0.01	1.31	-5.63	5.43	-0.05	2.53	-7.24	9.17
	0.80		0.07	1.18	-5.61	4.98	-0.12	1.31	-6.53	5.58	0.08	1.46	-6.83	8.08	0.00	2.30	-8.26	7.98
2	0.40	0.00	-0.04	1.91	-11.21	8.39	-0.04	1.79	-10.08	6.32	0.17	2.10	-6.21	11.41	0.34	3.80	-10.55	9.94
		0.40	-0.12	2.07	-10.01	10.59	0.09	1.68	-8.41	6.35	0.26	2.26	-10.43	12.70	0.07	4.91	-17.37	17.58
		0.70	0.15	2.54	-13.84	10.11	0.01	2.25	-11.22	13.10	-0.08	2.81	-13.69	14.69	0.14	5.55	-18.64	20.59
	0.80	0.00	-0.18	2.08	-7.21	9.48	-0.07	1.81	-13.09	7.98	-0.06	2.26	-14.03	9.90	-0.09	3.62	-14.05	11.65
		0.40	0.15	2.15	-9.65	11.76	0.03	2.08	-8.87	9.44	0.25	2.43	-11.20	11.45	0.06	4.23	-16.36	13.62
		0.70	-0.35	2.27	-14.43	10.80	-0.23	2.34	-12.16	9.68	-0.13	2.91	-13.44	11.54	-0.04	4.79	-14.74	16.98

Note. nuisance:number of nuisance dimension(s) present; vcor: correlations between primary dimensions; ncor: correlation between two nuisance dimensions

Table 26. Descriptive Statistics for the Bias-Induced Residual Correlations for All Crossed Conditions

1000 examinees																		
			10 items per dimension								20 items per dimension							
			low item discrimination				high item discrimination				low item discrimination				high item discrimination			
nuisance	vcor	ncor	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max
0	0.40	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	0.80		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
1	0.40	NA	0.851	0.067	0.632	0.977	0.827	0.063	0.573	0.963	0.865	0.061	0.564	0.987	0.822	0.064	0.561	0.974
	0.80		0.866	0.066	0.620	0.978	0.833	0.066	0.631	0.973	0.870	0.062	0.596	0.985	0.827	0.069	0.543	0.968
2	0.40	0.00	0.862	0.053	0.667	0.974	0.839	0.058	0.600	0.966	0.867	0.057	0.631	0.975	0.832	0.057	0.567	0.971
		0.40	0.868	0.063	0.596	0.979	0.833	0.059	0.623	0.969	0.873	0.056	0.651	0.981	0.843	0.053	0.616	0.975
		0.70	0.880	0.051	0.689	0.972	0.846	0.054	0.670	0.977	0.873	0.053	0.613	0.980	0.832	0.060	0.528	0.973
	0.80	0.00	0.865	0.061	0.632	0.978	0.839	0.065	0.588	0.968	0.867	0.056	0.624	0.980	0.828	0.067	0.539	0.968
		0.40	0.883	0.048	0.709	0.978	0.848	0.060	0.639	0.969	0.870	0.061	0.565	0.984	0.842	0.068	0.538	0.980
		0.70	0.858	0.061	0.623	0.970	0.845	0.060	0.628	0.962	0.873	0.059	0.646	0.982	0.830	0.064	0.553	0.968
5000 examinees																		
			10 items per dimension								20 items per dimension							
			low item discrimination				high item discrimination				low item discrimination				high item discrimination			
nuisance	vcor	ncor	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max
0	0.40	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	0.80		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
1	0.40	NA	0.870	0.059	0.634	0.978	0.832	0.052	0.681	0.958	0.869	0.060	0.521	0.980	0.827	0.059	0.598	0.977
	0.80		0.874	0.064	0.614	0.983	0.842	0.069	0.600	0.970	0.871	0.060	0.598	0.982	0.830	0.068	0.514	0.972
2	0.40	0.00	0.853	0.058	0.656	0.975	0.820	0.065	0.621	0.966	0.862	0.063	0.633	0.977	0.819	0.063	0.600	0.974
		0.40	0.878	0.052	0.657	0.979	0.837	0.052	0.653	0.963	0.865	0.056	0.653	0.978	0.830	0.057	0.591	0.969
		0.70	0.853	0.065	0.602	0.979	0.827	0.054	0.657	0.962	0.861	0.061	0.647	0.982	0.829	0.060	0.603	0.966
	0.80	0.00	0.887	0.053	0.691	0.985	0.840	0.061	0.589	0.962	0.876	0.058	0.649	0.984	0.835	0.066	0.507	0.967
		0.40	0.873	0.062	0.631	0.984	0.834	0.064	0.619	0.963	0.878	0.053	0.665	0.986	0.836	0.066	0.597	0.971
		0.70	0.870	0.060	0.652	0.988	0.828	0.069	0.601	0.963	0.875	0.054	0.645	0.982	0.846	0.063	0.601	0.980

Note. nuisance:number of nuisance dimension(s) present; vcor: correlations between primary dimensions; ncor: correlation between two nuisance dimensions

Figure 17(a). Distribution of Bias-Induced Residual Correlations with the Existence of Two Nuisance Dimensions for 1,000 Examinees and for the Crossed Conditions of the Remaining Four Testing Conditions

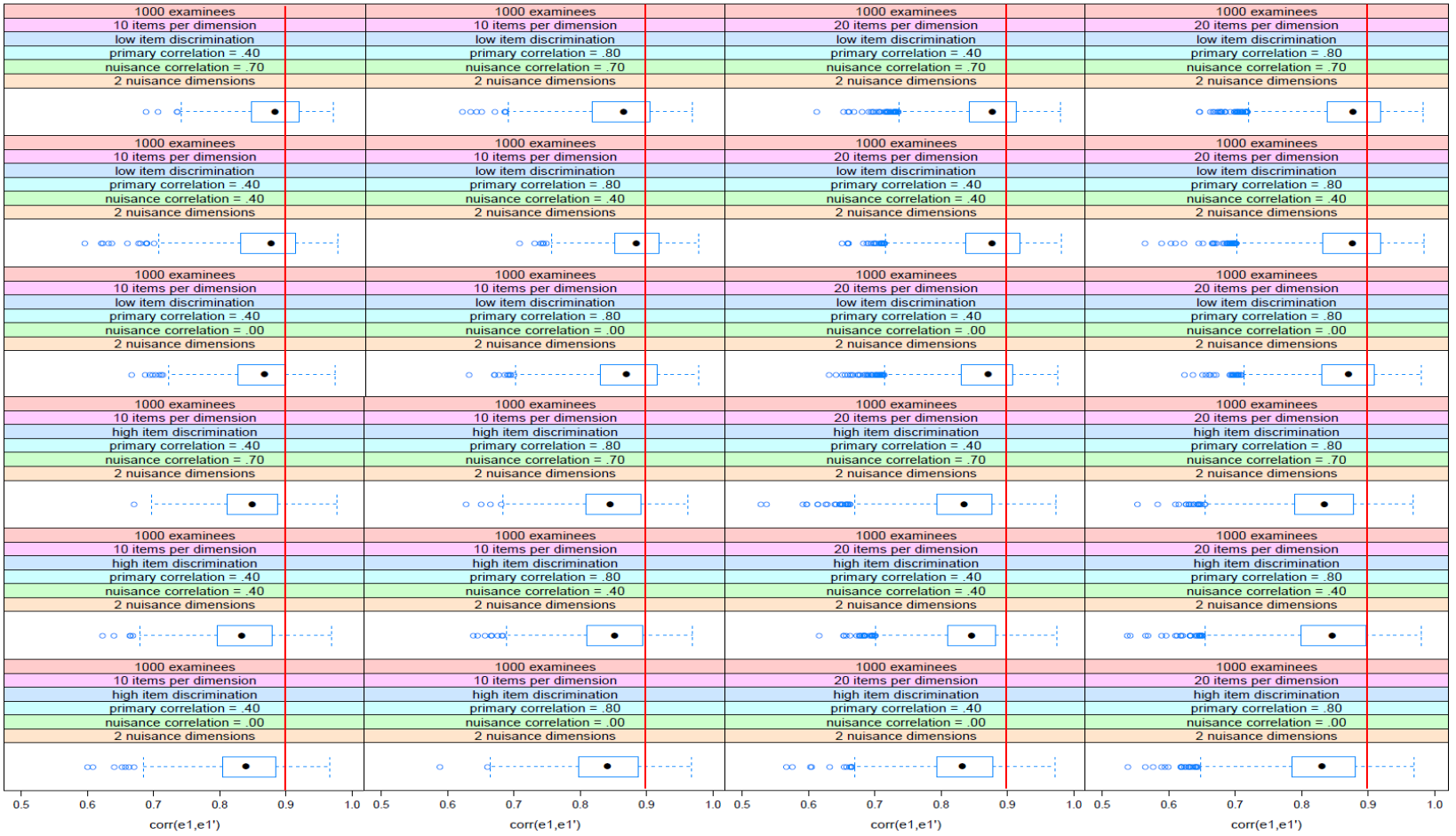
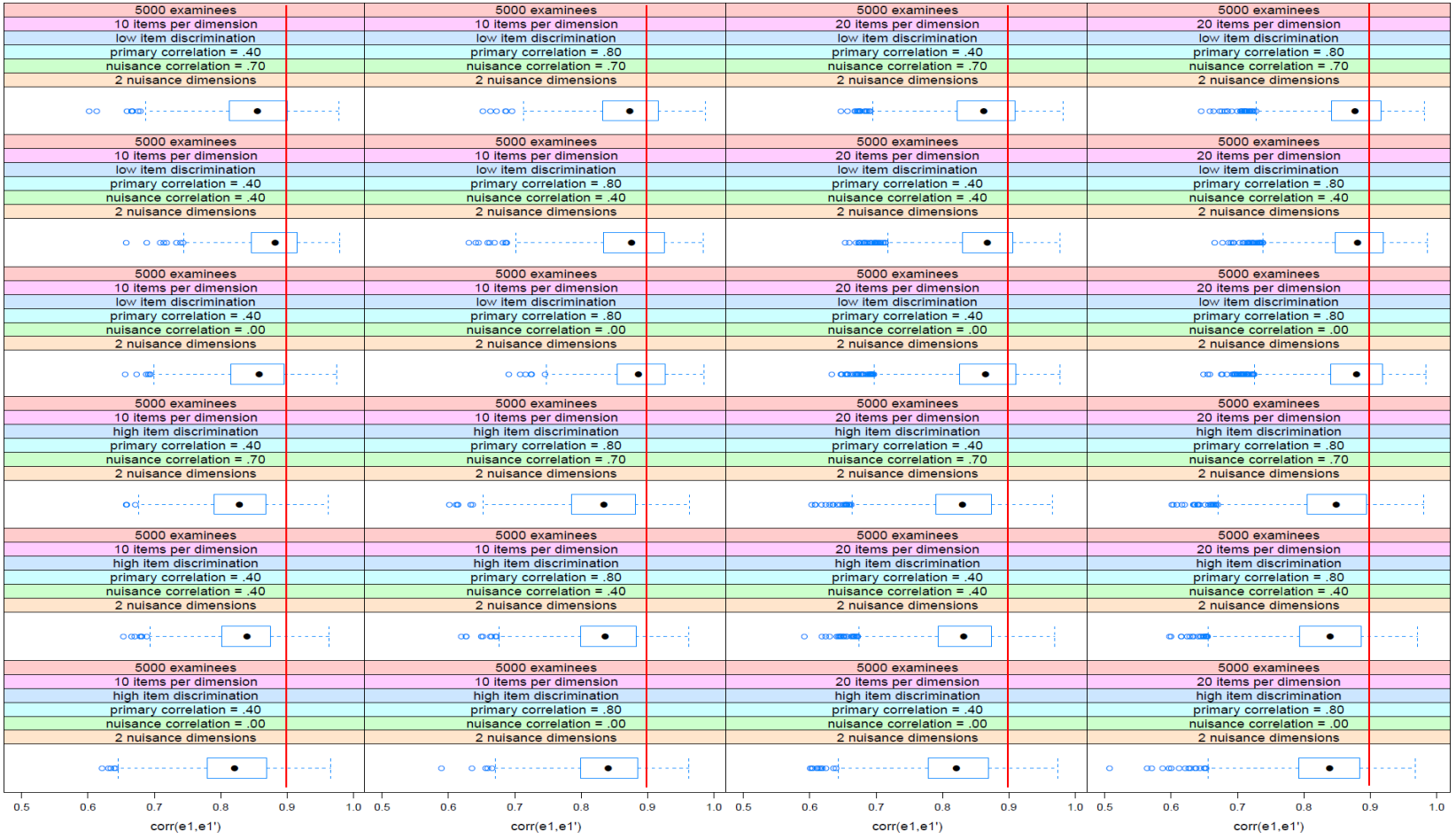


Figure 17(b). Distribution of Bias-Induced Residual Correlations with the Existence of Two Nuisance Dimensions for 5,000 Examinees and for the Crossed Conditions of the Remaining Four Testing Conditions



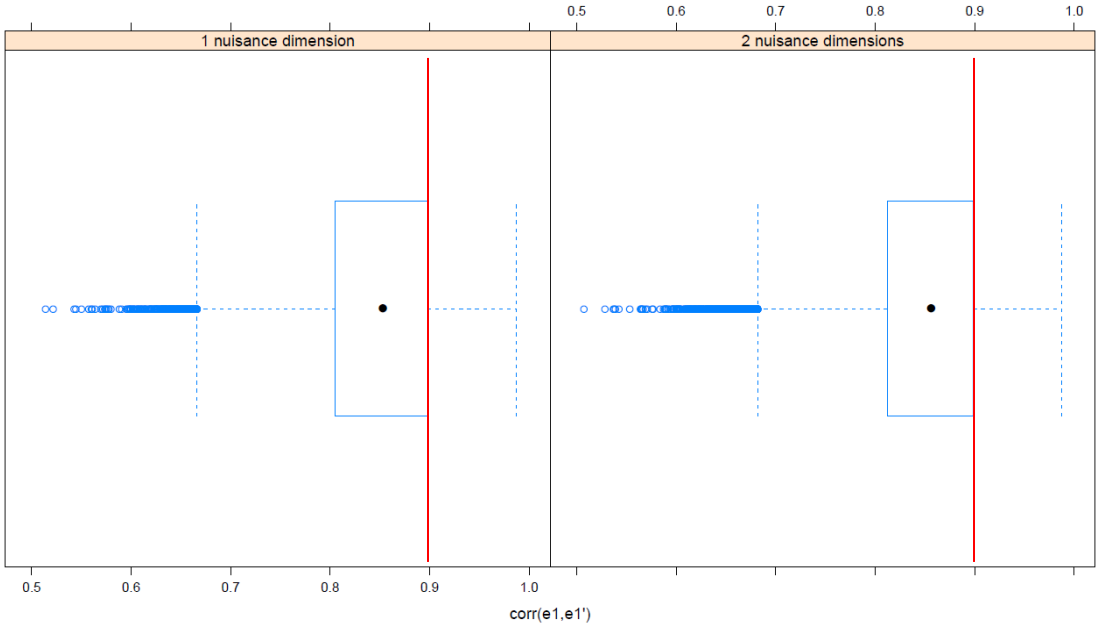
Results for Research Question 2a

This question asks how is the amount of the bias-induced residual correlations ($r_{e1_i, e1_h}$) impacted by the presence of nuisance dimensions. Table 27 and the box and whisker plots in Figure 18 provide the findings, namely, zero covariance in the baseline condition where there is no nuisance dimension. The residual correlations increase with the presence of at least one nuisance dimension. The amount of residual is similar although the amount of nuisance dimension increases to two dimensions.

Table 27. Descriptive Statistics for Bias-Induced Residual Correlations Given the Amount of Nuisance Dimension

presence of nuisance dimension(s)	*M*	*SD*	min	max
0	NA	NA	NA	NA
1	0.848	0.066	0.514	0.987
2	0.852	0.063	0.507	0.988

Figure 18. Distribution of Bias-Induced Residual Correlations Given the Amount of Nuisance Dimension



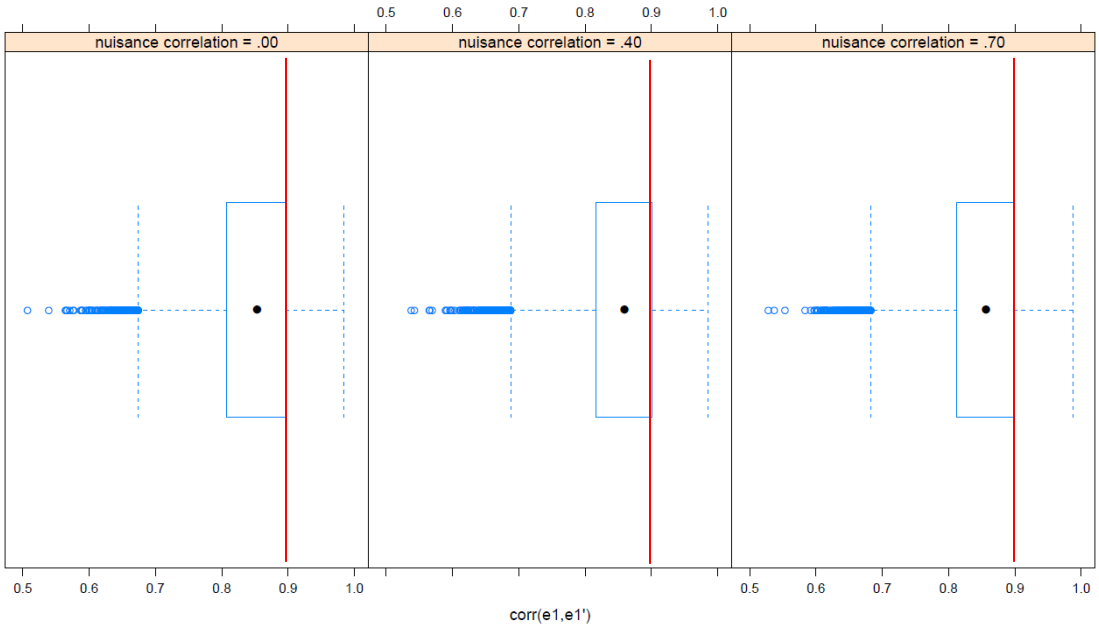
Results for Research Question 2b

This question asks how is the amount of the bias-induced residual correlations ($r_{e1_i, e1_h}$) impacted by the strength of correlations between nuisance dimensions. Table 28 and the box and whisker plots in Figure 19 provide the findings, namely, that the amount of residual correlations is similar regardless of the degree of correlation between two nuisance dimensions. On average, the mean of residual correlations are .849, .855, and .852 for correlations of .00, .40, and .70, respectively.

Table 28. Descriptive Statistics for Bias-Induced Residual Correlations Given the Strength of Correlations between Nuisance Dimensions

correlation between nuisance dimensions	*M*	*SD*	min	max
0.00	0.849	0.064	0.507	0.985
0.40	0.855	0.061	0.538	0.986
0.70	0.852	0.062	0.528	0.988

Figure 19. Distribution of Bias-Induced Residual Correlations Given the Strength of Correlations between Nuisance Dimensions



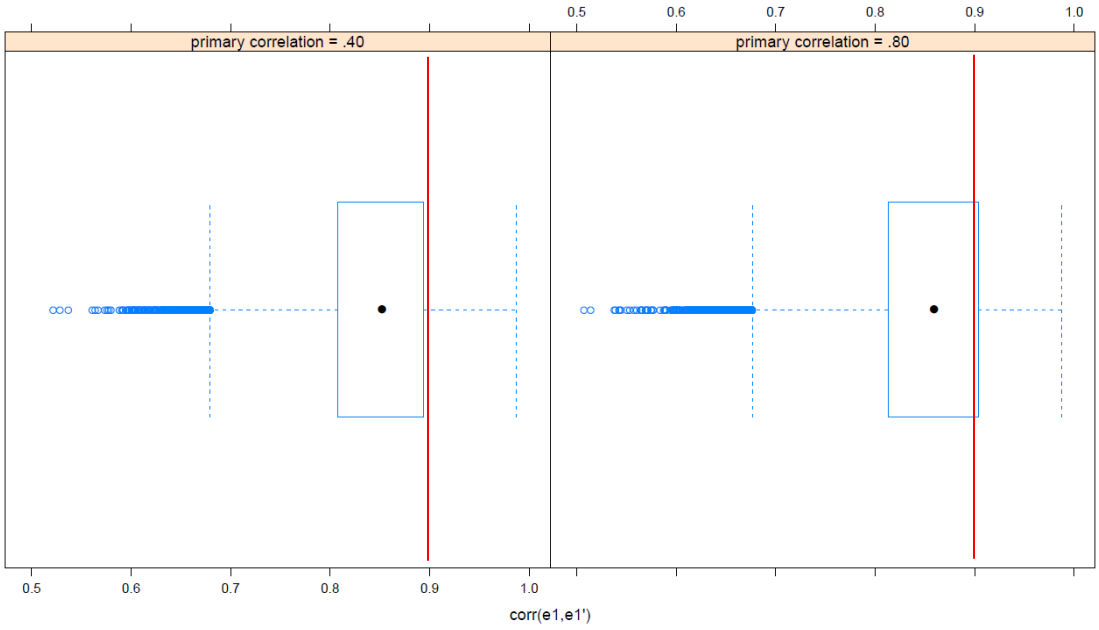
Results for Research Question 2c

This question asks how is the amount of the bias-induced residual correlations ($r_{e1_i, e1_h}$) impacted by the strength of the correlations between the primary dimensions. Table 29 and Figure 20 provide the findings, namely, that the amount of residual correlations is similar regardless of the correlations between the four primary dimensions. On average, the mean of residual correlations is .848 and .854 for correlations of .40 and .80, respectively.

Table 29. Descriptive Statistics for Bias-Induced Residual Correlations Given the Strength of Correlations between Primary Dimensions

correlations between primary dimensions	*M*	*SD*	min	max
0.40	0.848	0.062	0.521	0.987
0.80	0.854	0.065	0.507	0.988

Figure 20. Distribution of Bias-Induced Residual Correlations Given the Strength of Correlations between Primary Dimensions



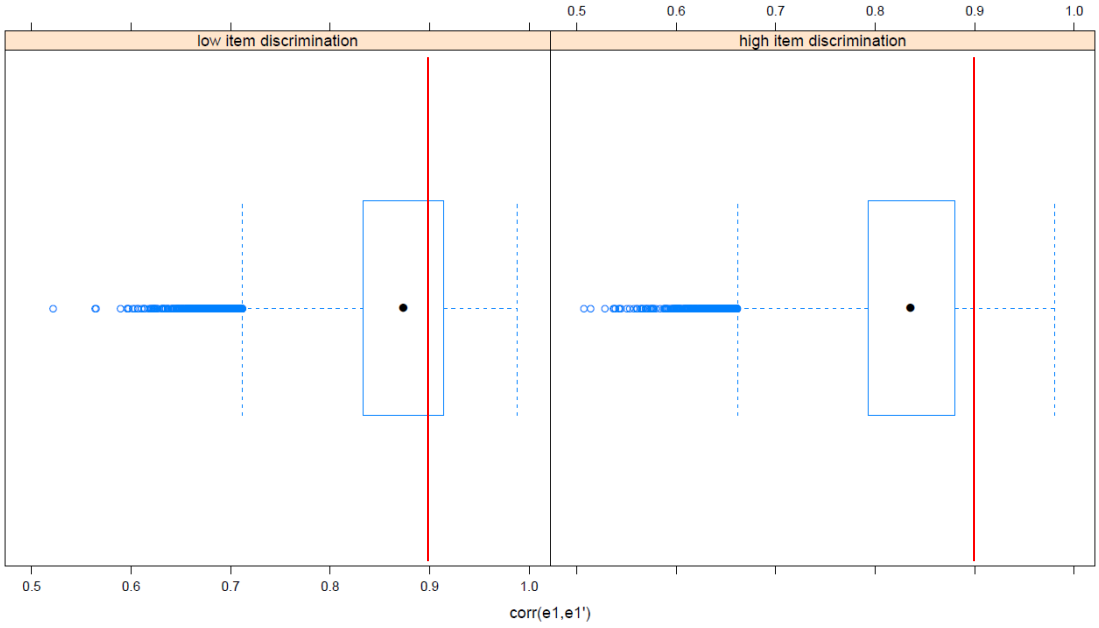
Results for Research Question 2d

This question asks how is the amount of the bias-induced residual correlations ($r_{e1_i,e1_h}$) impacted by changes in the item discrimination levels of the primary dimensions. Table 30 and Figure 21 provide the findings, namely that, on average, the amount of residual correlations decreases as the item discrimination levels increase (*M*=.833).

Table 30. Descriptive Statistics for Bias-Induced Residual Correlations Given Different Item Discrimination Levels on the Primary Dimensions

item discrimination levels	*M*	*SD*	min	max
all low	0.869	0.059	0.521	0.988
all high	0.833	0.063	0.507	0.980

Figure 21. Distribution of Bias-Induced Residual Correlations Given Different Item Discrimination Levels on the Primary Dimensions



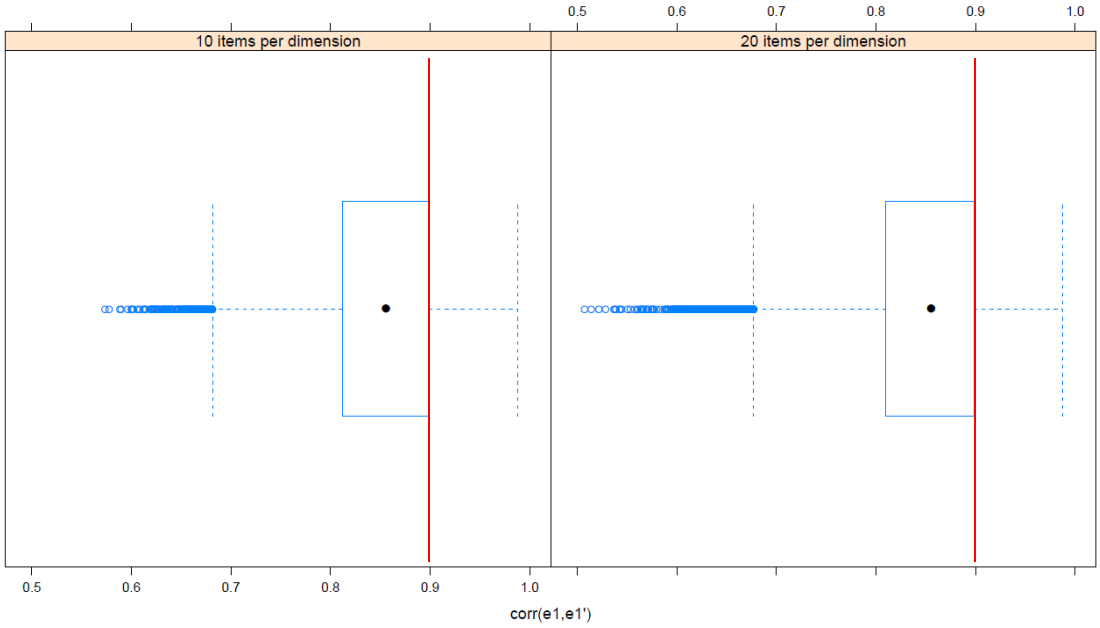
Results for Research Question 2e

This question asks how is the amount of the bias-induced residual correlations ($r_{e1_l, e1_h}$) impacted by the number of items in each primary dimensions. Table 31 and Figure 22 provide the findings, namely, that the amount of residual correlations is similar regardless of the test length. On average, the mean of the residual correlations is .852 and .851 for 40 and 80 test items, respectively.

Table 31. Descriptive Statistics for Bias-Induced Residual Correlations Given Number of Items in each Primary Dimensions

number of items per primary dimension	*M*	*SD*	min	max
10 items	0.852	0.063	0.573	0.988
20 items	0.851	0.064	0.507	0.987

Figure 22. Distribution of Bias-Induced Residual Correlations Given Number of Items in each Primary Dimensions



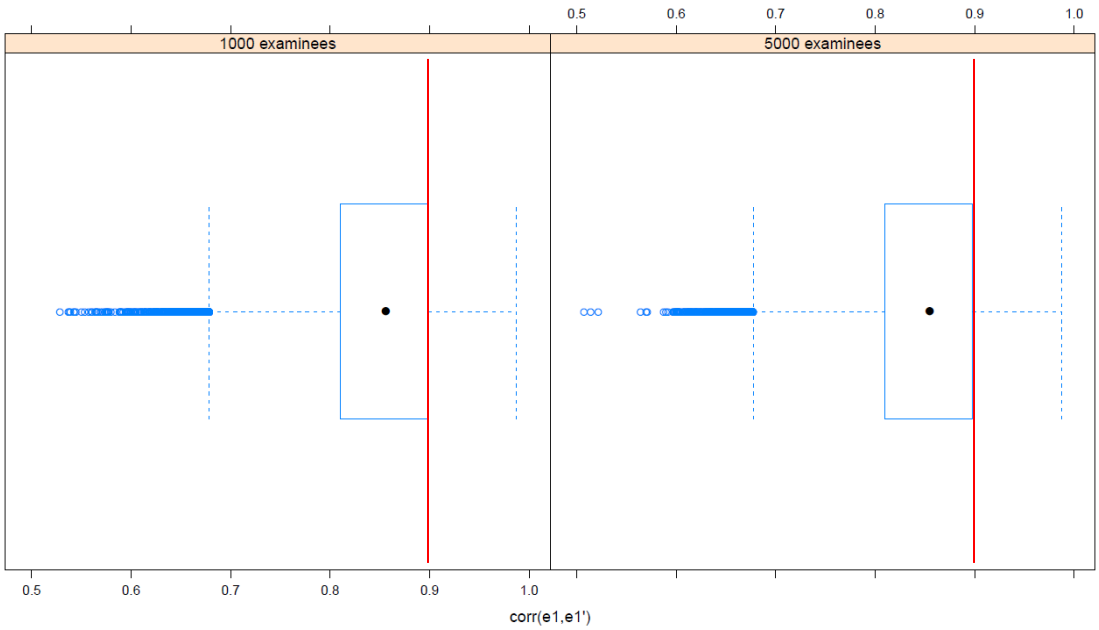
Results for Research Question 2f

This question asks how is the amount of the bias-induced residual correlations ($r_{e1_i,e1_h}$) impacted by changes in sample size. Table 32 and Figure 23 provide the findings, namely, that the amount of residual correlations is similar regardless of the test length. The mean of residual correlations is .851 for both 1,000 and 5,000 examinees.

Table 32. Descriptive Statistics for Bias-Induced Residual Correlations Given Different Sample Sizes

sample sizes	*M*	*SD*	min	max
1,000 examinees	0.851	0.063	0.528	0.987
5,000 examinees	0.851	0.064	0.507	0.988

Figure 23. Distribution of Bias-Induced Residual Correlations Given Different Sample Sizes



Results from Factorial ANOVA on the Bias-Induced Residual Correlations

To test the statistical differences across conditions, I incorporated the factorial ANOVA analyses. I conducted an investigation of the statistical effects of the presence and of nuisance dimension(s) (zero, one, and two nuisance dimension(s)) with the levels of item discriminations in the primary dimensions (all high and all low) on the e1 correlations using a 3 x 2 factorial ANOVA. I used a nominal Type 1 error rate of .001 for all ANOVAs in this research question to account for the large sample size of a

simulation study. As shown in Table 33, both the main effects of the nuisance dimensions and the level of item discrimination are statistically significant: $F(2, 157594)= 3078783.063$, $p<.001$ and $F(1, 157594)= 6554.937$, $p<.001$, respectively. The effect size of the presence of nuisance dimension factor was very large ($\eta^2=.975$). This is potentially because the e1 covariances were zero in the absence of any nuisance dimension, but with the presence of just one nuisance dimension the e1 correlations increased significantly. The effect size of the levels of item discrimination for all primary dimensions was small ($\eta^2=.04$). More importantly, the interaction effect of the two factors was statistically significant with a small effect size ($F(2,157594)=1463.646$, $p<.001$, $\eta^2=.018$).

Given the significant interaction, I conducted a test of simple structure as a follow-up. From Table 33 in the simple effects summary table and from Figure 24, it is evident that differences due to item discrimination levels occur with the presence of at least one nuisance dimension, although the sum of squares for one nuisance dimension is only about half of that when two nuisance dimensions are present in the test items. Specifically, a small effect size is shown for item discrimination in the presence of one nuisance dimension ($F(1, 157594) = 4258.002$, $p < .001$, $\eta^2 = .026$) and moderate effect size is shown for item discrimination with two nuisance dimensions ($F(1, 157594) = 10238.336$, $p < .001$, $\eta^2 = .061$).

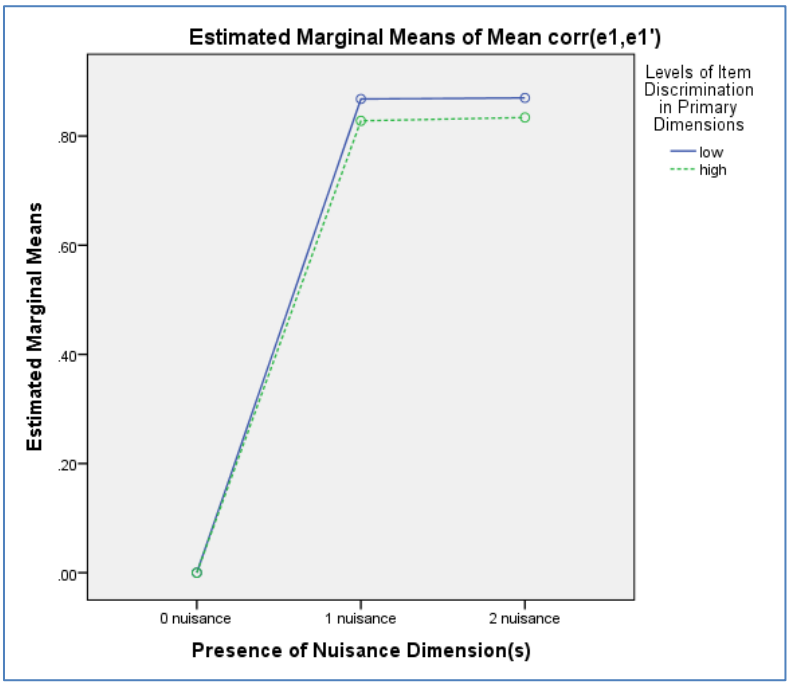
The profile plot in Figure 24 illustrates that, with the existence of two nuisance dimensions in the test items, the e1 correlations tend to be larger for items with lower discrimination levels.

Table 33. Summary Table for Two-Factorial ANOVA on the e1 Correlations

Overall Analysis						
Source	*SS*	*df*	*MS*	*F*	*p*	η^2
nuisance dimension	18258.576	2	9129.288	3078783.063	0.000**	0.975
item discrimination	19.437	1	19.437	6554.937	0.000**	0.040
2-Way Interaction	8.680	2	4.340	1463.646	0.000**	0.018
Error	467.302	157594	0.003			
Total	91801.634	157600				
Analysis of Simple Effects						
Source	*SS*	*df*	*MS*	*F*	*p*	η^2
item discrimination at 0 nuisance dimension	.000	1	.000	.000	1.000	.000
item discrimination at 1 nuisance dimension	12.626	1	12.626	4258.002	0.000**	.026
item discrimination at 2 nuisance dimensions	30.359	1	30.359	10238.336	0.000**	.061
Error	467.302	157594	.003			

**p<.001

Figure 24. Profile Plots for Two-Factorial ANOVA on the e1 Correlations



Next, I performed a 2 x 2 factorial ANOVA to explore the effect of sample size and correlations of the primary dimensions. As shown in Table 34, the interaction effect between the two factors is not statistically significant, (F(1,157596)=5.002, p=.025, $\eta^2 = .000$). Both the main effects of sample size and correlations between primary dimensions are also not statistically significant (F(1,157596)=.107, p=.743, $\eta^2 = .000$) and (F(1,157596)=6.769, p=.009, $\eta^2 = .000$, respectively).

Table 34. Summary Table for Two-Factorial ANOVA on the e1 Correlations

Overall Analysis						
Source	*SS*	*df*	*MS*	*F*	*p*	η^2
N	.013	1	.012	.107	.743	.000
Vcorr	.806	1	.806	6.769	.009	.000
N*vcorr	.596	1	.596	5.002	.025	.000
Error	18767.449	157596				
Total	91801.634	157600				

**p<.001

Table 35. Summary Table for Two-Factorial ANOVA on the e1 Correlations

Overall Analysis						
Source	*SS*	*df*	*MS*	*F*	*p*	η^2
N	.060	1	.060	.795	.373	.000
Ncorr	6928.752	3	2309.584	30741.324	0.000**	.369
N*ncorr	.271	3	.090	1.202	.307	.000
Error	11839.827	157592	.075			
Total	91801.634	157600				

**p<.001

To further explore the effect of number of sample size and three correlations of nuisance dimensions, I performed a 2 x 4 factorial ANOVA (see Table 35). The interaction effect between the two study conditions is not statistically significant (F(3,157592)=1.202, p=.307, $\eta^2 = .000$). The main effect of sample size is also not statistically significant (F(1,157592)=.795, p=.373, $\eta^2 = .000$). However, the main effect of the correlations of nuisance dimensions is statistically significant (F(3,157592)=30741.324, p<.001) with a large effect size (η^2=.369). Despite the large effect size, the multiple comparison analysis with Tukey's Honest Significant Difference (HSD: Tukey, 1949) with significance level of .001, showed that the e1 correlations were not significantly different for each pair-wise comparison of the nuisance correlations (.00, .40. and .70). The significant mean difference was only observed when nuisance correlation is NA (i.e., not applicable), in which when there was only one or zero nuisance dimension present.

Similar findings are shown in Table 36 for the analysis of a three-way factorial ANOVA with sample size (1000 and 5000 examinees), correlations on primary dimensions (.40 and .80), and correlations between nuisance dimensions (NA, .00, .40, and .70) as factors. There are no significant two-way or three-way interactions. While the main effect of the correlations from primary dimensions is statistically significant, its effect size was very small (F(1,157592)=12.282, p<.001, $\eta^2 = .000$). Lastly, similar to the previous finding of the two-way ANOVA between sample size and correlations between nuisance dimensions, the main effect

of correlations between nuisance dimensions was statistically significant with a large effect size ($F(1,157592)=30745.739$, $p<.001$, $\eta^2=.369$).

Table 36. Summary Table for Two-Factorial ANOVA on the e1 Correlations

Overall Analysis						
Source	*SS*	*df*	*MS*	*F*	*p*	η^2
N	0.06	1	0.060	0.795	0.373	0.000
Vcorr	0.923	1	0.923	12.282	0.000**	0.000
Ncorr	6928.752	3	2309.584	30745.739	0.000**	0.369
N*vcorr	0.908	1	0.908	12.086	0.001	0.000
N*ncorr	0.271	3	0.090	1.202	0.307	0.000
vcorr*ncorr	0.183	3	0.061	0.811	0.488	0.000
N*vcorr*ncorr	0.717	3	0.239	3.180	0.023	0.000
Error	11839.827	157592	.075			
Total	91801.634	157600				

**p<.001

Summary of Overall Findings for Research Question 2

I conclude that, with the absence of a nuisance dimension in the test items (baseline), the residual covariance is zero, regardless of the study conditions. With the presence of one nuisance dimension, the e1 residual correlations are large on average and are closer to the e1 residual correlations of tests with two nuisance dimensions.

On average, the e1 residual correlations are larger when item discrimination level for all primary dimensions is low. However, the e1 residual correlations are large and similar in magnitude on average—and thus are not affected—when: the correlations between two nuisance dimensions vary; the correlations between the primary dimensions vary; test lengths vary; and when sample sizes vary.

My test of simple effects and examination of the profile plot for two-factorial ANOVA between the presence of nuisance dimensions and item discrimination levels indicate that the e1 residual correlations tend to become larger with the existence of at least one nuisance dimension in the test items for items with a lower discrimination level.

CHAPTER V

CONCLUSIONS

In this chapter, I will provide a brief summary of the study, provide implications for practitioners, highlight the significance of this study and how it may be able to fill existing gaps in the research on the use of a model-based simulation study for educational measurement, discuss the lessons learned, and explore possible directions for future research.

Summary of Findings and Implications for Practice

For each research question, I will provide the summary and an explication of the findings, and address how my findings may inform the practice of educational assessment, especially in the context of the K-12 College and Career Readiness (CCR) assessments.

I have termed both of my research questions based on six testing conditions: (1) the presence of (a) nuisance dimensions, (2) the strength of correlations between two nuisance dimensions, (3) the strength of correlations between primary dimensions, (4) the levels of item discrimination on the primary dimensions, (5) test length, and (6) number of examinees. These conditions were chosen to mimic to a certain extent the practical and complex reality of assessments in general and in the next generation K-12 CCR assessments for education in particular.

Summary and Implications: Research Question 1

I began with the question of how well do different, more parsimonious IRT calibration models perform when calibrating simulated examinees' response data intentionally generated with the existence of unintended and irrelevant constructs under the six aforementioned simulated testing conditions. To evaluate the different unidimensional and multidimensional calibrations, I employed two out of the four types of ERF-based residuals introduced by Luecht & Ackerman (2017; see also Table 10 in Chapter Four): (1) e2, which is the parameter-estimation residual; and, (2) e3, which is the estimated model-data fit. I found that the residuals from the UIRT parameter estimation (e2UIRT) and the UIRT estimated model-data fit (e3UIRT) consistently provide large ERS residuals and consistently deviate further from the simulated examinees' observed response data for all conditions and crossed conditions of the study. I therefore conclude that, on average, the MIRT calibration model tends to produce less error and tends to fit the data better than the UIRT model.

My findings have some major implications, particularly in terms of test uses and interpretations, which is a key validity concern in educational testing (AERA et al., 2014; ITC, 2013a, Kane, 2013). Most assessment programs still employ UIRT despite acknowledging the possibility of multidimensionality in the examinees response data (e.g., SBAC, 2016; PARCC, 2014). With the implementation of assessments aligned to the CCR content standards (e.g., North Carolina Testing Program, 2016; PARCC, 2014; SBAC, 2016), students' academic achievement

standards have been reported as (1) Level 4 and above: on track for being prepared for college and career at the conclusion of high school and (2) Level 3 and above: demonstrating preparedness to be successful at the next grade level. Table 37 illustrates the proficiency descriptors and cut scores from the North Carolina 2013/2014 End-of-Grade (EOG) tests for mathematics (NCDPI, 2014a).

Table 37. Academic Achievement Descriptors and Cut Scores for North Carolina End-of-Grade Math Test for Year 2013/2014

Achievement Level	Brief Description	Meets Grade-Level Proficiency Standard	Meet Common Core State Standards	Level Ranges & Percent Correct
Level 5	Superior command	**Yes**	**Yes**	≥ 460 86-100%
Level 4	Solid command	**Yes**	**Yes**	451-459 66-85%
Level 3	Sufficient command	**Yes**	No	448-450 57-65%
Level 2	Partial command	No	No	440-447 39-56%
Level 1	Limited command	No	No	≤ 439 0-38%

Adopted and modified from the North Carolina Department of Public Instruction (NCDPI), March 2014

Given the large errors in UIRT calibration, with on average about three to six percent from total score percentage being underestimated (see Table 12), such big difference could seriously impact both grade-level proficiency and the CCSS Standards cut scores (despite the small effect size). Such implications could cause students to be held back a grade level or could abstain them (especially the ELLs and SWDs) from entering high schools and colleges. Using the difference that matter (DTM) empirical criteria (cf. Dorans & Feigenbaum, 1994), losing six percent of the EOG scores in math test (see Table 34) may refrain students to achieve Level 4 in order to meet the CCSS. Things become more complicated as the UIRT model which is used to calibrate shorter tests violates the assumption of local independence (LInd) (Reese, 1995) with the presence of different item formats (Hohensinn & Kubinger, 2012 Rabinowitz & Brandt, 2001; Randall, Sireci, Li, & Kaira, 2012; Sireci & Zenisky, 2006; Taylor et al., 1999; Zenisky & Sireci, 2002; Zenisky & Sireci, 2006) or linguistic complexity (Abedi & Linquanti, 2012; Bailey & Wolf, 2012; Bunch, Kibler, & Pimental, 2012; Fillmore & Fillmore, 2012; Lee, Quinn, & Valdes, 2013; Moschkovich, 2012; Turner & Danridge, 2014; Wolf, Wang, Blood, & Huang, 2014), or both.

Summary and Implications: Research Question 2

My second research question concerned how the aforementioned six testing configurations interact to affect the amount of the bias-induced, residual correlations of test items ($r_{e1_i, e1_h}$). Analyses of the ($r_{e1_i, e1_h}$) suggest that the residual correlations increase with the presence of at least one nuisance dimension but tend to decrease with high item discriminations. My findings provide some evidence that the number of nuisance dimensions does not affect the structure of the bias-induced residual correlations. To state this in the context of the next generation CCR assessments, the test items may exhibit statistical dependencies (i.e., LID) when there is at least one unintended dimensionality affecting the test items—regardless of whether the irrelevant constructs are due to the innovative item/response formats or the presence of interfering linguistic complexity or both.

LID (e.g., Chen & Thissen, 1997; Edwards & Cai, 2008, 2011; Yen, 1984, 1993) is an assumption for many psychometric models, especially in the UIRT models (de Ayala, 2009; Embretson & Reise, 2000; Hambleton & Swaminathan, 2010). When the assumption of LID is met, there should be no significant residual covariance between items after conditioning on the examinees' ability. However, the newer educational assessments, which are more innovative and challenging, are often intentionally or unintentionally multidimensional in nature. The complex interactions between examinees and task/items and among the items themselves are often unexplained and may result in some conditional associations between the items/tasks. Hence, a central question is how can test developers address and

attempt to reduce the presence of unintended dimensionality in attempting to design a test that is comprehensive, more innovative, and critical?

Test developers can conduct feasibility reviews (Popp et al., 2016; Zenisky & Sireci 2006, 2001) to judge the appropriateness of the various items/response formats based on psychometric, operational, and contextual criteria. Bias and sensitivity reviews by content experts during item and task development (PARCC, 2014; SBAC, 2016) could help determine whether the constructs are measuring what was intended (Lissitz & Samuelson, 2007a, 2007b, AERA et al., 1999, 2014) and whether the constructs are essential to the purpose, use, and intended interpretations of the test (AERA et al., 1999, 2014; ITC, 2013a, Kane, 2013). Various tutorials and documentations (PARCC, 2014; SBAC, 2016a, 2016b), such as an informational guide, high level blueprints, guidelines on test practice, use of technology, and source of cognitive complexity from the assessment consortia are also made available in an effort to acknowledge, address, and reduce the unintended multidimensionality of the tests (see also other guidelines from international testing standards (AERA et al., 2014; ITC, 2005a)).

The use of test accommodations for the ELL and SWD subgroups is another effort to help reduce any possible unnecessary contamination from unintended constructs. However, more research is required to examine whether such accommodations help facilitate examinees or contribute to another nuisance dimension and thereby contribute to LID in test items (Abedi, 2006; Chapelle & Douglas, 2006; Popp, et. al, 2016; SBAC, 2016; Zenisky & Sireci, 2001). More pilot

tests need to be conducted and the findings from such tests could help inform test developers with respect to test development. The findings could also help researchers to determine and prioritize the relevant criteria (Kane, 2004, 2014) for their conditions when conducting simulation studies.

Essentially, by creating more discriminating items relevant to the construct of interest, test developers may be able to compensate for the effect of unintended constructs on the residual covariance.

Discussion

The simplest and most frequently used IRT models are the models that specify a single or unidimensional latent ability. UIRT models are stable, easy to understand, and employ parameter estimation methods that are, to some extent, computationally friendly (de Ayala, 2009; Embretson & Reise, 2000; Hambleton & Swaminathan, 2010; Reckase, 2009). In practice, examinees response data seldom meet the rigorous assumptions of the UIRT models. The nature of educational tests, especially the ones instigated by the CCSS and the NGSS, are inherently complex and often not unidimensional. Thus, it is usually not appropriate to fully define the latent ability space with only one latent factor. The various assumptions of the UIRT model have also made its application to the multicomponent and complex test designs and formats somewhat limited. UIRT models assume that each item within a test measures the same unidimensional construct and that item responses, given the latent construct, are locally independent. Violation of this conditional independence

will affect the psychometric properties of the test (Ackerman, 1987; Edwards & Cai, 2010; Chen & Thissen, 1997; Oshima, 1994; Reese, 1995; Sireci et al., 1991; Tuerlinckx & De Boeck, 2001; Thissen et al., 1989; Wainer & Thissen, 1996; Wainer & Wang, 2000; Yen, 1993), hence jeopardizing the validity of test scores (Lissitz & Samuelson, 2007a, 2007b) and their interpretations and uses (AERA et al., 1999, 2014; Kane, 2013).

I found that employing UIRT model in an estimation process together with the presence of construct-irrelevant factors will result in the underestimation of examinees' raw scores (see also Reese, 1995, although Reese did not employ the Luecht and Ackerman's (2007) ERF-based residuals approach). Nonetheless, the UIRT model is still preferred by many state assessment programs and consortia (e.g., NCDPI, 2016; PARCC, 2014; SBAC, 2016) as a calibration model in next generation assessments due to its simplicity (i.e., parsimony). Incorporating more complex model comes with a cost. More often, more complex models in general tend to have less stability (Reckase, 2009) and could result in potential convergence issues (San Martin, Gonzalez, & Tuerlinckx, 2015). Again, the important consideration is to determine when the more complex model is necessary and when does dimensionality contribute to the problem (i.e., construct relevance and irrelevance, different characteristics of examinees from various subpopulations, etc.). As I have shown, unidimensional estimation, whether unintended or intended, could underestimate examinees' raw scores, which could hold them back for a grade level or cost them admission to college.

Standards 12.3 and12.6 from Chapter 12: Educational Testing and Assessment of the *Standards* (AERA, et al., 2014) mandate careful test design and development, as well as comprehensive documentation of supporting evidence on the feasibility of CBT (see Popp, et al., 2016; Zenisky & Sireci, 2001) to gather information about the construct, to avoid CIV, and to uphold accessibility for all examinees. Although the implementation of computer-based tests in the next generation assessments is promising, there is limited research into the possibility that such tests might introduce CIV (Haladyna & Downing, 2004, Huff & Sireci, 2001; Lakin, 2014) and affect the residual covariance structure due to such 'nuisance', interfering dimensionality. Introducing new or unfamiliar computerized item formats to examinees from different subpopulations creates particular challenges for test developers because examinees need to quickly and accurately understand what the test items require (Haladyna & Downing, 2004) as well as to understand the differences that may exist across formats (e.g., Pearson Educational Measurement, 2005; Scalise, 2012, 2009; Scalise and Gifford, 2006; Sireci & Zenisky, 2006). The critical challenge is how best to introduce a task so that all examinees are able to respond to the format as intended by the test developers.

Through my comprehensive examination of the impact of different realistic test configuration on the components of residual that are independent on the IRT scale and through my explicit consideration of the issue of residual covariance and potential construct irrelevant factors in the context of the next generation assessments, my findings will benefit psychometricians and test developers in

refining the test design and development process. Proper considerations of the various factors that may impact test scores and the covariance structures should be taken as part of the test development process such as in an assessment engineering (AE) framework (e.g., Luecht, 2006a, 2006b, 2007, 2013; Luecht, Gierl, Tan, & Huff, 2006) and the universal design (UD) principle (Ketterlin-Geller, 2008; Thompson et al., 2002; Thompson et al., 2004; Stone et al., 2016). An AE encourages the treatment of dimensionality to be addressed proactively in test development through the development of principled multidimensional information by specifying the number of traits of interest and by identifying potential irrelevant traits. Universally designed assessments are designed and developed to allow participation of the broadest possible range of students to provide valid inferences about performance on grade-level standards for all students who participate in the assessment. More importantly, my results should provide insight to psychometricians about the most effective calibration method to be used with next generation assessments to ensure valid interpretations and uses of the test scores and to uphold fairness in testing practices.

Although it is acknowledged that a simulation of educational testing situations will never accurately portray the true complexity and inherent context of real data (Luecht & Ackerman, 2017) and therefore does not permit firm conclusions, simulations are still useful for framing the general patterns and trends of a limited selection of phenomena of interest. For that reason, when conducting simulation studies, researchers should generate the observed response data that

capture the complex reality of and potential 'contaminations' in testing practices and assessment programs.

Researchers with a methodological and technical background or from different school of thoughts might find it difficult to reconcile their approaches with the idea of situating a simulation study in the context of a given assessment when no attempt of generalization to a specific context is actually made. Others from the socio-cognitive paradigm (see Bachman & Palmer, 1996, 2010; Chalhoub-Deville, 2003, 2009; Deville & Chalhoub-Deville, 2006; Mislevy & Duran, 2014; Snow, 2008, 2010; Snow, 1994) may posit views on test score variances as a result of different types of persons and tasks interactions such that they may shy away from framing the variability in test scores as errors. Such consideration is important and interesting but is not within the scope of this dissertation. My concern is with special populations such as the ELLs and SWD and the potential challenges that they might face with the new innovative next generation CCR assessments.

As I have stated previously, I am hoping to establish some context for my research interest and by prioritizing some evidence (Kane, 2004, 2013) in the literature in order to appreciate and employ new innovations in model-based simulation study (Luecht & Ackerman, 2017). The ERF-based residuals approach (Luecht & Ackerman, 2017) resolves some fundamental limitations in conducting a simulation study-in context. Essentially, their approach has shed important light on the model-based simulation study in educational measurement, for which their approach provides a useful and clean separation of different types of errors from

different sources and enables comparison across psychometric models, estimators from different commercial software, and scaling choices (i.e., metric-neutral).

Limitations and Directions for Future Research

My work here is only a first step at systemically incorporating and examining the sources of and potential approaches for limiting the influence of the potential CIV in the next generation K-12 assessments. Thus, the findings may not be generalizable across assessment programs despite my effort to situate the study in context.

I may not have generated sufficient amounts of simulated data or used multiple different algorithms to accurately mirror the complexity of real data. To conserve time and for the sake of simplicity, I only used ten iterations for each crossed condition and employ the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm (Cai, 2010b, 2010c) in flexMIRT®3.0RC (Vector Psychometric Group, 2017). Future researchers should conduct their simulation studies using a larger number of replications and should consider different estimation methods from multiple commercial software applications (Luecht & Ackerman, 2017).

During data generation process, I also dropped several simulation conditions (i.e., other combinations of item discrimination levels for the four primary dimensions and different ratios of the number of dichotomous items to the number of polytomous items for each content area (i.e., subscore)) to reduce the complexity of my study. Therefore, other researchers may expand the test format variables to

include conditions that represent the other situations possibly observed in practice. Moreover, future researchers might address the structure of the response data in which there are associations between the primary and the nuisance dimensions. In this study, I did not consider such associations. Another researcher might attempt to create TEs and CPAs of different lengths, yielding different numbers of dichotomous measurement opportunities or polytomous score units. Expanding this idea would allow the polytomous score units that are created to contain differing number of score categories, unlike my study in which I constrained each polytomous unit formed to contain exactly five score categories.

Also, researchers can conduct logical subsequent simulation study (or studies) by applying the methods to outcomes from testing programs that contain test configurations similar to the ones incorporated in my simulation (i.e., other achievement tests). Other researchers can also consider different assessment programs with different distributions of examinees such as in the certification/licensure test where cut score is set at the 20^{th} percentile of the examinee population (Luecht, 2006a). A researcher can conduct a simulation in the context of extreme placement test in which only the top five percentage of examinees will be admitted (e.g., the case study by Prometric, Inc. (2011)). Additionally, more complex testing environments, such as those which exist in computer adaptive test settings, could be simulated.

Another direction for future researchers conducting a simulation study is to compare the performance of the bias-induced (e1) residual correlations that I

employed. Its performance can be compared with another correlational-based LID index such as the Yen's Q3 (Yen, 1984, 1993) since the Yen's Q3 index is known to be negatively biased (Yen's 1984) due to the part-whole contamination between the observed and expected response data.

A researcher who uses real data, especially from the operational setting, could demonstrate the effectiveness of the ERF-based residuals approach (Luecht & Ackerman, 2017). Other real data studies can also examine the residual covariance structures at various places along the latent scale (e.g. at various cut points, or for groups of different abilities) (see Goodman, Luecht, & Zhang, 2009; Reese, 1995; Taylor et al., 1999).

REFERENCES

Abedi, J. (2006) Language issues in item development. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 377–399). Rahway, NJ: Erlbaum.

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment*, 8(3), 231-257.

Abedi, J., Courtney, M., & Goldberg, J. (2000). *Language modification of reading, science, and mathematics test items*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Courtney, M., Mirocha, J., Leon, S., & Goldberg, J. (2005). *Language accommodations for English language learners in large-scale assessments: Bilingual dictionaries and linguistic modification (CSE Tech. Report. No. 666)*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data (CSE Tech. Report. No. 603)*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., & Linquanti, R. (2012). Issues and Opportunities in Improving the Quality of Large Scale Assessment Systems for English Language Learners. *Understanding Language: Language, Literacy, and Learning in the Content Areas*. Stanford University. Retrieved from http://ell.stanford.edu/sites/default/files/pdf/academic-papers/07-Abedi%20Linquanti%20Issues%20and%20Opportunities%20FINAL.pdf

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*(3), 219–234.

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math* performance (CSE Tech. Rep. No. 478). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Lord, C., Kim-Boscardin, C., & Miyoshi, J. (2000). The effects of accommodations on the assessment of LEP students in NAEP (CSE Tech. rep. No. 537). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., & Plummer, J. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CRESST Technical Report No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Gandara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice*, 36-46.

Ackerman, T. A. (1987). *The robustness of LOGIST and BILOG IRT estimation programs to violations of Local Independence*. Paper presented at the Annual Meeting of the American Educational Research Association.

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, *13*(2), 113-127.

Ackerman, T. A. (1994a). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement*, 1*8*(3), 257-275.

Ackerman, T. A. (1994b). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement In Education*, *7*(4), 255-278.

Ackerman, T. A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, *20*(4), 311-329

Ackerman, T. A. (2005) Multidimensional Item response theory modeling. In J. McArdle & A. Maydeu-Olivares (Eds.). *Contemporary psychometrics: Festschrift for Roderick P. McDonald*. Hillsdale,NJ: Erlbaum.

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay,multiple-choice, and free-response writing tests. *Applied Psychological Measurement, 12,* 117-128.

Adams, R. J., Wilson, M. R., & Wang, W. C. (1997). The multidimensional random coefficientsmultinomial logit. *Applied Psychological Measurement, 21,* 1-23.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Ames, A. J., & Penfield, R. D. (2015). An NCME instructional module on item-fit statistics for item response theory models. Educational Measurement: Issues and Practice, 34(3), 39-48.

Andrich, D. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*, 581-594.

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Anstrom, K., DiCerbo, P., Butler, F., Katz, A., Millet, J., & Rivera, C. (2010). *A review of the literature on academic English: Implications for K-12 English language learners*. Arlington, VA: The George Washington University Center for Equity and Excellence in Education. Retrieved from http://www.vidyablog.com/LitReviewAcademicEnglish.pdf

Bachman, L. F. (1990) *Fundamental Considerations in Language Testing.* New York: Oxford University Press.

Bachman, L. F. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues & Practice, 21*(3), 5-18.

Bachman, L. F. and Palmer, A. (1996) *Language Testing in Practice.* New York: Oxford University Press.

Bachman, L. F. and Palmer, A. (2010) *Language Assessment in Practice.* New York: Oxford University Press.

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple- choice and short-answer questions in a marketing context. *Journal of Marketing Education, 25*, 31-36.

Baghaei, P., & Ravand, H. (2016). Modeling local item dependence in cloze and reading comprehension test items using testlet response theory. *Psicológica*, *37*, 85-104.

Bailey, A. L., & Wolf, M, K. (2012). The challenge of assessing language proficiency *aligned* to the Common Core State Standards and some possible solutions. *Understanding Language: Language, Literacy, and Learning in the Content Areas*. Stanford University. Retrieved from http://ell.stanford.edu/sites/default/files/pdf/academic-papers/08-Bailey%20Wolf%20Challenges%20of%20Assessment%20Language%20Proficiency%20FINAL_0.pdf

Baker, F., & Kim, S. (2004). Item Response Theory: Parameter Estimation Techniques (2nd ed.). New York, NY: Marcel Dekker, Inc.

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Soloway, E. (1990). The relationship of expert-system scored constrained free-response items to multiple-choice and open-ended items. *Applied Psychological Measurement, 14,* 151-162.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice responseformats – it does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11,* 385-395.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1983). The discrete Bayesian. In H. Wainer & S. Messick (Eds.), Principals of modern psychological measurement. A festschrift for Frederick M. Lord (pp. 103-115), NJ: Lawrence Erlbaum.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, *46*, 443-459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261-280.

Bock, R. D., & Jones, L. V. (1968). *The measurement and prediction of judgment and choice*. San Francisco: Holden-Day.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment, *Applied Psychological Measurement*, *6*(4), 431-444.

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice*, *14*, 9-27.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

Browne, C., Culligan, B., & Phillips, J. (August, 2016). *New general service list: The most important words for second language learners of English*. Retrieved from http://www.newgeneralservicelist.org/about/

Bunch, M. (2010). Testing English language learners under No Child Left Behind. *Language Testing*, *28*(3), 323-341.

Bukhari, N., Boughton, K., & Kim, D. I. (2016). *Psychometric characteristics of technology enhanced items from a computer-based interim assessment program*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington DC.

Bunch, G. C., Kibler, A., & Pimentel, S. (2012). Realizing opportunities for English learners in the Common Core English language arts and disciplinary literacy standards. *Understanding Language: Language, Literacy, and Learning in the Content Areas*. Stanford University. Retrieved from http://ell.stanford.edu/sites/default/files/pdf/academic-papers/01_Bunch_Kibler_Pimentel_RealizingOpp%20in%20ELA_FINAL_0.pdf

Bunch, M. B. (2011). Testing English language learners under No Child Left Behind. *Language Testing*, *28*(3), 323-341.

Burge, S. S. (2007). An investigation of dimensionality across grade levels and effects on vertical linking for elementary grade mathematics achievement tests (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations. (UMI No. 3289093).

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. Psychometrika, 75(1), 33-57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. Journal of Educational and Behavioral Statistics, 35(3), 307-337.

Cai, L. (2010c). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75,* 581–612. doi:10.1007/s11336-010-9178-0

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*(3), 221–248.

Cawthon, S. (2007). Hidden benefits and unintended consequences of No Child Left Behind policies for students who are Deaf or hard of hearing. *American Educational Research Journal, 44*(3), 460–492.

Cawthon, S. (2011). Test item linguistic complexity and assessments for Deaf students. *American Annals of the Deaf*, *156*(3), 255-269.

Celce-Murcia, M., Dornyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics*, *6*(2), 5–35.

Center for Applied Special Technology (CAST) (2017). *Universal design for learning guidelines: An online handout.* Peabody, MA: CAST. Retrieved January, 2016, from the World Wide Web: www.cast.org

Center for Universal Design. (1997). What is universal design? North Carolina State University. Retrieved October 10 2015, from the World Wide Web: http://www.design.ncsu.edu.

Chalhoub-Deville, M. (2001). Task-based assessment: Characteristics and validity evidence. In P. Skehan, M. Swain, & M. Bygate (Eds.), *Applied language studies: Task based research.* (pp. 210-228). NY: Longman.

Chalhoub-Deville, M. (2009). Content validity considerations in language testing contexts. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 241-263). Charlotte, NC: Information Age Publishing, Inc.

Chalhoub-Deville, M. and Deville, C. (2008). Nationally mandated testing for accountability: English language learners in the US. In B. Spolsky and F. Hult (Eds.), *The handbook of educational linguistics*, (pp. 510-522). Malden, MA: Blackwell Publishing.

Chalmers, P. R. (2012). mirt: A multidimensional item response theory package for the R Environment. *Journal of Statistical Software*, *48*(6), 1-29. URL http://www.jstatsoft.org/v48/i06/.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. UK: Cambridge University Press.

Chen, F. (2010). *Differential language influence on math achievement* (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations. (UMI No. 3434131)

Chen, W.-H. (1993). IRT LD: A computer program for the detection of pairwise local dependence between test items. Research Memorandum 93-2. Chapel Hill, NC: L.L.Thurstone Laboratory, University of North Carolina at Chapel Hill.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. Journal of Educational and Behavioral Statistics, 22(3), 265-289.

Chiappe, P., Siegel, L. S., & Wade-Woolley, L. (2002). Linguistic diversity and the development of reading skills: A Longitudinal Study. *Scientific Studies of Reading*, *6*(4), 369-400, DOI: 10.1207/S1532799XSSR0604_04

Clauser, B. E., Margolis, M. J., & Clauser, C. J. (2016). Issues in simulation-based assessment. In F. Drasgow (Eds). *Technology and testing: Improving educational and psychological measurement* (pp. 49-78). New York: Routledge.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31-44.

Cohen, A., & Farley, F. H. (1977). The common-item problem in measurement: Effects on cross-cultural invariance of personality inventory structure. *Educational and Psychological Measurement, 37*(3),757-760.

College Board (2015). *The official SAT® study guide: Your first look at the redesigned SAT® direct from the maker of the test*. College Board: New York.

Collis, K. F., Romberg, T. A., & Jurdak, M. E. (1986). A technique for assessing mathematical problem-solving ability. *Journal of Research in Mathematics Education,* 17,206-211.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238.

Criteria Corp. (2017). *Criteria pre-employment testing*. Downloaded January 24 2017, from the World Wide Web: https://www.criteriacorp.com/solution/test_portfolio.php

Crocker, L. M., Miller, D., & Franks, E. A. (1989). Quantitative methods for assessing The fit between test and curriculum. *Applied Measurement in Education*, *2*, 179-194.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests, *Psychological Bulletin*, *52*(4), 281-302.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, *19*, 121–129.

Cummins, J. (1981). Age on arrival and immigrant second language learning in Canada: a reassessment. *Applied Linguistics*, *2*, 132–149.

Cummins, J. (1994). Primary language instruction and the education of language minority students. In C. F. Leyba(Ed.), *Schooling and language minority students: A theoretical framework*. Los Angeles, CA: Evaluation Dissemination and Assessment Center, California State University, Los Angeles.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*(4), 481-489. de Ayala, R. J. (2009). *Theory and practice of item response theory*. New York: Guilford Press.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Publications, Inc.

De Champlain, A. F. (1996). The effect of multidimensionality on IRT True-Score equating for subgroups of examinees. *Journal of Educational Measurement, 33*(2), 181-201.

De Champlain, A. F., & Gessaroli, M. E. (1998). Assessing the dimensionality of item response matrices with small sampe sizes and short test lengths. *Applied Measurement in Education*, *11*(3), 231-253.

Dessoff, A. (2012). Are you ready for Common Core Math? *District Administration*, *48*(3), 53-60.

Deville, C. and Chalhoub-Deville, M. (2006). Old and new thoughts on test score variability: Implications for reliability and validity. In M. Chalhoub-Deville, C. Chapelle, and P. Duff (Eds.), *Inference and Generalizability in applied linguistics: multiple perspectives* (pp. 9-26). Philadelphia, PA: John Benjamins Publishing Company.

Donoghue, J. R. (1994). An empirical investigation of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal* of *Educational Measurement,* 31, 295-3 11.

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. In R. E. Bennett & W. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessments* (pp. 135-165). Hillsdale, NJ: Lawrence Erlbaum Associates

Downing, S. M. (2006a). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 3–25). Mahwah, NJ: Lawrence Erlbaum Associates.

Downing, S. M. (2006b). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 207-301). Mahwah, NJ: Lawrence Erlbaum Associates.

Downing, S. M., Baranowski, R. A., Grosso, L. J., & Norcini, J. J. (1995). Item type and cognitive ability measured: The validity evidence for multiple true-false items in medical specialty certification. *Applied Measurement in Education, 8*, 187-197.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Drasgow, F. (2016). *Technology and testing: Improving educational and psychological measurement*. New York: Routledge.

Dudley, A. (2006). Multiple dichotomous-scored items in second language testing: Investigating the multiple true-false item type under norm-referenced conditions. *Language Testing, 23*, 198-228.

Duran, R. P., & Szymanski, M. (1995). Cooperative learning interaction and constructionof activity. *Discourse Processes*, *19*, 149–164. doi:10.1080/01638539109544909

Dutro, S., & C. Moran. (2003). Rethinking English language instruction: an architectural approach. In G. G. Garcia (Ed.), *English learners: reaching the highest level of English literacy*, (pp. 227–58). Newark, DE: International Reading Association. Retrieved from http://www.wou.edu/~aalthumayri13/REthinking_ESL_instruction_Article.pdf

Edwards, M. C., & Cai, L. (2008, September). A new diagnostic procedure to detect departures from local independence in item response models. Paper presented at the Quantitative Forum at the L. L. Thurstone Psychometric Laboratory, Chapel Hill, NC.

Edwards, M. C., & Cai, L. (2011, July). *A new procedure for detecting departures from local independence in item response models.* Paper presented at the annual meeting of American Psychological Association, Washington, D.C. Retrieved from http://faculty.psy.ohio-state.edu/edwards/documents/APA8.2.11.pdf

Embretson (Whitely), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179-197.

Embretson S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher, 36*(8)*,* 449–455.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New J ersey: Erlbaum.Ercikan, K., Schwarz, R., Julian, M., Burket, G., Weber, M., & Link, V. (1998). Calibrationand scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 35,* 1 37-155.

Erickson, F. (2004). *Talk and social theory: Ecologies of speaking and listening in everyday life*. Cambridge, England: Polity Press.

Farley, F. H., & Cohen, A. (1980). Common items and reliability in personality measurement. *Journal of Research in Personality,14*(2), 207-211.

Fillmore, L. W., & Fillmore, C. J. (2012). What does text complexity mean for English language learners and language minority students? Understanding Language: Language, Literacy, and Learning in the Content Areas. Stanford University. Retrieved from http://ell.stanford.edu/sites/default/files/pdf/academic-papers/06-LWF%20CJF%20Text%20Complexity%20FINAL_0.pdf

Finch, H., & Habing, B. (2007). Performance of DIMTEST- and NOHARM-based statistics for testing unidimensionality. *Applied Psychological Measurement*, *31*(4), 292-307.

Finney, D. J. (1952). *Probit analysis: A statistical treatment of the sigmoid response curve*. London: Cambridge University Press.

Flanagan, J. C. (1939). General considerations in the selection of test items and a short method of estimating the product-moment coefficient from data at the tails of the distribution. *Journal of Educational Psychology, 30,* 674–80.

Florida Department of Education. (2016). Understanding NGSSS Reports Grades 5 & 8 Science and End-of-Course Assessments Spring 2016. Retrieved from http://www.fldoe.org/core/fileparse.php/5662/urlt/UNGSSSSEOCRSpring16.pdf

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23,* 267-269.

Garcia, O., Kleifgen, J. A., & Falchi, L. (2008). From English language learners to emergent bilinguals (Equity Matters: Research Review No. 1). New York: Teachers College, Columbia University, A Research Initiative of the Campaign for Educational Equity. Retrieved from http://www.equitycampaign.org/i/a/document/6532_Ofelia_ELL_Final.pdf

Gessaroli, M. E., & De Champlain, A. F. (1996). Using the approximate Chi-Square statisticto test the number of dimensions underlying the repsonses to a set of items. *Journal of Educational Measurement, 33*, 157-179.

Geva, E., Yaghoub-Zadeh, Z., & Schuster, B. (2000). Understanding differences in word recognition skills of ESL children. *Annals of Dyslexia*, *50*, 123–154.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423-436.

Gibbons, R. D., Bock, D. R., Hedeker, D. R., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-information item bi-factor analysis of graded response data. *Applied Psychological Measurement*, *31*(1), 4-19.

Gonulates, E. (2004). *Technical report: Dimensionality analysis of the uniform CPA examination*. American Institute of Certified Public Accountants & Michigan State University.

Goodman, J. (2008). *An examination of the residual covariance structures of complex performance assessments under various scaling and scoring methods* (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations. (UMI No. 3316319).

Goodman, J. T., Luecht, R. M., & Zhang, Y. O. (2009). *Technical report: An examination of the magnitude of residual covariance for complex performance assessments under various scoring and scaling methods*. American Institute of Certified Public Accountants.

Gorin, Joanna S. (2007). Reconsidering issues in validity theory. *Educational Researcher, 36*(8)*,* 456–462.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). New Jersey: Lawrence Erlbaum Associates, Inc.

Gulliksen, H. O. (1950). *Theory of mental tests*. New York: Wiley. (Reprinted in 1987 by Lawrence Erlbaum Associates: Hillsdale, NJ).

Gutierrez, K. D. (2008). Developing a sociocritical literacy in the third space. Reading Research Quarterly, 43, 148–164. doi:10.1598/RRQ.43.2.3

Haberkorn, K., Pohl, S., & Carstensen, C. H. (2016). Incorporating different response formats of competence tests in an IRT model. *Psychological Test and Assessment Modeling*, *58*(2), 223-252.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. Psychometrika, 75(2), 209-227.

Habing, B., Finch, H., Roberts, J. S. (2005). A Q3 statistic for unfolding item response theory models: Assessment of unidimensionality with two factors and simple structure. *Applied Psychological Measurement*, *29*(6), 457–471.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 65-110). Westport, CT: Praeger.

Hakuta, K., Butler, Y., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* University of California Linguistic Minority Research Institute Policy Report 2000-1. Santa Barbara: UC-LMRI.

Haladyna, T. M. (1992). Context-dependent item set. *Educational Measurement: Issues and Practice*, *11*(1), 21-25.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27.

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation & The Health Professions*, *27*(4), 349–368.

Haley, D. C. 91952). *Estimation of the dosage mortality relationship when the dose is subject to error*. Stanford: Applied Mathematics and Statistics Laboratory, Stanford University, Technical Report 15.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory.* Boston: Kluwer-Nijhoff.

Hambleton, R. K., & Swaminathan, H. (2010). *Item response theory: Principles and applications*. Boston: Kluwer.

Harris, J., & Bamford, C. (2001). The uphill struggle: Services for Deaf and hard of hearing people: Issues of equality, participation, and access. *Disability and Society, 16*(7), 969–980.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement, 20,* 1–14.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2,* 41–54. doi:10.1007/BF02287965

Houts, C. R., & Edwards, M. C. (2013). The performance of local dependence measures with psychological data. *Applied Psychological Measurement*, *37*(7), 541-562.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice*, *20*(3), 16-25.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. Language Testing, 26, 219-244.

International Test Commission (2005a). *International Guidelines on Computer-Based and Internet Delivered Testing.* Retrieved from https://www.intestcom.org/files/guideline_computer_based_testing.pdf

International Test Commission (2005b). *ITC Guidelines for Adapting Tests.* Retrieved from https://www.intestcom.org/files/guideline_test_adaptation.pdf

International Test Commission (2013a). *International Guidelines for Test Use.* Retrieved from https://www.intestcom.org/files/guideline_test_use.pdf

International Test Commission (2013b). *ITC Guidelines for Quality Control in Scoring, Test Analysis, and Reporting of Test Scores.* Retrieved from https://www.intestcom.org/files/guideline_quality_control.pdf

International Test Commission (July, 2014). *ITC Guidelines for Security of Tests, Examinations, and Other Assessments.* Retrieved from https://www.intestcom.org/files/guideline_test_security.pdf

International Test Commission (2015). *ITC Guidelines for Practitioner Use of Test Revisions, Obsolete Tests, & Test Disposal.* Retrieved from https://www.intestcom.org/files/guideline_test_disposal.pdf

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. Journal of Educational Measurement, 40(1), 1-15.

Johnson, D. F, & White, C. B. (1980). Effects of training on computerized test performance in the elderly. *Journal of Applied Psychology*, *65,* 357–358.

Johnson, M. (1987). The body in the mind: The bodily basis of meaning, imagination, and reason. Chicago: University of Chicago Press.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, *2*(3), 135-170.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport, CT: Praeger.

Kane, M. T. (2009). Validating the interpretations and uses if test scores. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 39-64). Charlotte, NC: Information Age Publishing, Inc.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 5*(1)*,* 1-73.

Kane, M. T., Crooks, T.,&Cohen, A. (1999). Validating measures of performance. Educational Measurement: Issues and Practice, 18(2), 5–17.

Ketterlin-Geller, L. R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice*, *27*(3), 3-16.

Kim, H. R. (1994). New techniques for the dimensionality assessment of standardized test data (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations. (UMI No. 9512427).

Goodman, J. (2008). *An examination of the residual covariance structures of complex performance assessments under various scaling and scoring methods* (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations. (UMI No. 3316319).

Kim, J. S., & Bolt, D. M. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, *26*(4), 38-51.

Kim, S. H., Cohen, A. S., & Lin, Y.-H. (2006). LDIP: A computer program for local dependence indices for polytomous. *Applied Psychological Measurement*, *30*(6), 509–510.

Kim, D., De Ayala, R. J., Ferdous, A. A., & Nering, M. L. (2011).The comparative performance of conditional independence indices. *Applied Psychological Measurement*, *35*(6), 447-471.

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, *19*(4), 357-381.

Kim. S., Walker, M. E., & McHale, F. (2010). Comparisons among designs for equating mixed-format test in large-scale assessments. *Journal of Educational Measurement*, *47*(1), 36-53.

Kingston, N. M. (2009) Comparability of Computer- and Paper- Administered Multiple-Choice Tests for K–12 Populations: A Synthesis, *Applied Measurement in Education*, *22*(1), 22-37.

Kingston, N. M., & Dorans, N. J. (1982). The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test. *GRE Board Tech. Rep*. No. 79-12. Princeton, NJ: Educational Testing Service.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: The Guilford Press.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 155-186). Westport, CT: Praeger.

Kolen, M. J., & Lee, W. (2011). Psychometric properties of raw and scale scores on mixed-format tests. *Educational Measurement: Issues and Practice, 30*(2), 15–24.

Lakin, J. M. (2014). Test directions as a critical component of test design: Best practices and the impact of examinee characteristics. *Educational Assessment*, *19*(1), 17-34, DOI: 10.1080/10627197.2014.869448

Lane, S. & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 387-431). Westport, CT: Praeger.

Lee, J. (1986). The effects of past computer experience on computerized aptitude test performance. *Educational and Psychological Measurement, 46,* 727–733.

Lee, O., Quinn, H., & Valdés, G. (2013). Science and language for English language learners in relation to Next Generation Science Standards and with implications for Common Core State Standards for English Language Arts and mathematics. *Educational Researcher*, *42*(4), 223–233.

Lee, Y.-W. (2004). Examining passage-related local item dependence (LID) and measurement construct using *Q3* statistics in an EFL reading comprehension test. *Language Testing*, *21*(1), 74–100.

Lent, R. W. (2004). Toward a unifying theoretical and practical perspective on wellbeing and psychosocial adjustment. *Journal of Counseling Psychology, 51*(4), 482-509.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 1-55.

Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 119–158). Washington, DC: American Council on Education.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15,* 1-16.

Linn, R. L.& Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice*, *13*, 5-15.

Linquanti, R. & George, C. (2007). Establishing and utilizing an NCLB Title III accountability system: California's approach and findings to date. In J. Abedi (Ed.), *English language proficiency assessment in the nation: Current status and future practice* (pp. 105-118). Davis: University of California Press. Linn,

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice*, *13*, 5-15.

Lissitz, R., & Jiao, H. (2012). Computers and their impact on state assessments. USA: Information Age Publishing Inc.

Lissitz, R. W., & Samuelsen, K. (2007a). A Suggested Change in Terminology and Emphasis regarding Validity and Education. *Educational Researcher*, *36*(8), 437-448

Lissitz, R. W., & Samuelsen, K. (2007b). Further Clarification Regarding Validity and Education. Educational Researcher, *36*(8), 482–484.

Lollis, J., & LaSasso, C. (2009). The appropriateness of the NC state-mandated reading competency test for Deaf students as a criterion for high school graduation. *Journal of Deaf Studies and Deaf Education, 14*(1), 76–98.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lorie, W. (2014). *Application of a scoring framework for technology-enhanced items*. Paper presented at the 2014 annual meeting of the National Council for Measurement in Education, Philadelphia, PA.

Luecht, R. M. (2001). Capturing, codifying and scoring complex data for innovative, computer-based items. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Luecth, R. M. (2003). Applications of multidimensional diagnostic scoring for certification and licensure tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 21-25, 2003).

Luecht, R. M. (2006a). Designing tests for pass-fail decisions using item response theory. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 575–596). Mahwah, NJ: Lawrence Erlbaum Associates.

Luecht, R. M. (2006b, May). Engineering the test: From principled item design to automated test assembly. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

Luecht, R. M. (2006c, September). Assessment engineering: An emerging discipline. Paper presented in the Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB, Canada.

Luecht, R. M. (2007, February). Assessment engineering workshop. Presented at Association of Test Publishers Conference. Palm Spring, CA.

Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (2006, April). Scalability and the development of useful diagnostic scales. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Luecht, R. M. (2013). Assessment Engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology*, *14*, 1-38, Retrieved from http://www.jtla.org

Luecht, R. M., Miller, T. R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement*, *16*(3), 279-293.

Luecht, R. M., & Ackerman, T. (2017). A technical note on IRT simulation studies: Dealing with truth, estimates, observed data, and residuals. *Educational Measurement: Issues and Practice*. Manuscript submitted for publication.

Martiniello, M. (2008). Language and the performance of English language learners in math word problems. *Harvard Educational Review*, *78*, 333–368.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differentialitem functioning for English language learners in math tests. *Educational Assessment*, *14*(3-4), 160-179.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Masters, J., Famularo, L., & King, K. (2015). *Conducting research on technology-enhanced assessment: lessons learned from the field*. Paper presented at the Annual Meeting of the American Educational Research Association Chicago, IL.

Mazzeo, J., Druesne, B., Raffeld, P., Checketts, K., & Muhlstein, A. (1991). *Comparability of computer and paper-and-pencil scores for two CLEP general examinations* (College Board Report No. 91-5). Princeton, NJ: Educational Testing Service.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, *19*, 91-100.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13-100). Washington, D.C.: American Council on Education.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(3), 13–23.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*(1), 100-117.

McDonald, R. P. (1982). Linear versus non-linear models in item response theory. *Applied Psychological Measurement*, *6(4)*, 379-396.

McDonald, R. P., & Mok, M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30*, 23-40.

McKinley, R. L., & Reckase, M. D. (1983). An application of a multidimensional extension of the two-parameter logistic latent trait model (ONR-83-3). (ERIC Document Reproduction Service No. ED 240 168).

McKinley, R. L., & Reckase, M. D. (1983a). An extension of the two-parameter logistic model to the multidimensional latent space (Research Report ONR 83-2). Iowa City IA: The American College Testing Program.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 335-355). New York: American Council on Education & MacMillan.

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36*(8), 463–469.

Mislevy, R. J., & Duran, R. P. (2014). A sociocognitive perspective on assessing EL students in the age of Common Core and Next Generation Science Standards. *TESOL Quarterly*, *48*(3), 560-585.

Mitchell, R. (2008). Academic achievement of Deaf students. In R. Johnson & R. Mitchell (Eds.), *Testing Deaf students in an age of accountability* (pp. 38–50). Washington, DC: Gallaudet University Press.

Mosier, C. I. (1943). On the reliability of a weighted composite. *Psychometrika*, *8*, 161-168.

Moschkovich, J. (2012). Mathematics, the Common Core, and language: Recommendations for mathematics instruction for ELs aligned with the Common Core. *Understanding Language: Language, Literacy, and Learning in the Content Areas*. Stanford University. Retrieved from http://ell.stanford.edu/sites/default/files/pdf/academic-papers/02-JMoschkovich%20Math%20FINAL_bound%20with%20appendix.pdf

Moss, P. A. (2007). Reconstructing validity. *Educational Researcher, 36*(8), 470–476.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, *17*(4), 351-363.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, *19*(1), 73-90.

Mutua, N. K., & Elhoweris, H. (2002). Parents' expectations about the postschool outcomes of children with hearing impairments. *Exceptionality, 10*(3), 189–201.

Nandakumar, R., & Stout, W. E (1993). Refinement of Stout's procedure for assessing latent trait essential unidimensionality. *Journal of Educational Statistics, 18*(1), 41-68.

National Center for Education Statistics. (May, 2016). *English language learners in public schools*. Retrieved August 22, 2016 http://nces.ed.gov/programs/coe/indicator_cgf.asp

National Governors Association Center for Best Practices, Council of Chief State School Officers (CCSSO). (2010a). Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects. Washington, DC: Author.

National Governors Association Center for Best Practices, Council of Chief State School Officers (CCSSO). (2010b). Common Core State Standards for Mathematics. Washington, DC: Author.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: The National Academies Press.

Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York: Routledge, Taylor and Francis Group.

North Carolina Department of Public Instruction (2016). Assessment briefs: North Carolina READY End-of-Grade assessment data reporting. Retrieved from http://www.dpi.state.nc.us/docs/accountability/policyoperations/assessbriefs/eogdatareport16.pdf

Next Generation Science Standards (NGSS) Lead States. (2013). Next Generation Science Standards: For states, by states. Washington, DC: National Academies Press.

North Carolina Department of Public Instruction (2014a). Achievement Levels for End-of-Grade Mathematics. Retrieved from http://www.ncpublicschools.org/docs/accountability/testing/achievelevels/eogmathald14.pdf

North Carolina Department of Public Instruction (2014b). Technical Brief. Retrieved from http://www.ncpublicschools.org/docs/accountability/testing/eoceogtechbrief14.pdf

No Child Left Behind (2001). Public Law 107–110, January, 2002. Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/107-110.pdf

No Child Left Behind Act (2001a). Public Law 107-110, Title I, January, 2002. Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/pg1.html

No Child Left Behind Act (2001b). Public Law 107-110, Title III, January, 2002. Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/pg39.html

Organisation for Economic Co-operation and Development. (2012). *PISA 2009 technical report*. Paris, France: OECD.

Ohio Department of Education. (2016). Ohio's state tests 2015-2016 subscore definitions. Retrieved from https://education.ohio.gov/getattachment/Topics/Testing/Testing-Results/Results-for-Ohios-State-Tests/OSTSubscoreChart.pdf.aspx

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*(3), 200-219.

Pearson Educational Measurement. (2002). *Research report on recent trends in comparability studies.* Austin, TX: Author.

PARCC (2016a). Proposed sources of cognitive complexity in PARCC items and tasks: ELA/Literacy (summary). Retrieved from http://www.parcconline.org/assessments/test-design/ela-literacy/ela-performance-level-descriptors

PARCC (2016b). Proposed sources of cognitive complexity in PARCC items and tasks: Mathematics (summary). Retrieved from http://www.parcconline.org/assessments/test-design/mathematics/math-performance-level-descriptors

PARCC (2016c). Score results: Understanding the 2015-16 score report. Retrieved from http://www.parcconline.org/assessments/score-results

Parshall, C. G., & Harmes, J. C. (2009). Improving the quality of innovative item types: Fourtasks for design and development. Journal of Applied Testing Technology, 10(1), 1-20.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. (2010). Innovative items for computerized testing. In Van der Linden, W.J. & Glas, C.E.W. (eds.). Elements of Adaptive Testing. (2010). New York: Springer.

Patz, R., & Junker, B. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.

Penfield, R. D. (2010). Item analysis. In K. Geisinger (Ed.), *APA handbook of testing and assessment in Psychology*. American Psychological Association.

Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*, 33(1), 36-48.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, *19*, 5-15. doi: 10.1111/j.1745-3992.2000.tb00033.x

Peyton, V., Kingston, N.M, Skorupski, W., Glasnapp, D., & Poggio, J. (2009). Kansas English Language Proficiency Assessment (KELPA) Technical Manual. Center for Educational Testing and Evaluation, University of Kansas. https://cete.ku.edu/sites/cete.drupal.ku.edu/files/docs/Technical_Reports/2009/peyton2009_KELPA.pdf

Poggio, J. & McJunkin, L. (2012). History, current practice, perspectives, and what the future holds for computer based assessment in K-12 education. In R. Lissitz, & H. Jiao (Eds.). *Computers and their impact on state assessments* (pp. 25-53). USA: Information Age Publishing Inc.

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report – Scaling the data of the competence tests.* (NEPS Working Paper No. 14). Bamberg, Germany: University of Bamberg, National Educational Panel Study.

Popp, E. C., Tuzinski, K., & Fetzer, M. (2016). Actor or avatar? Considerations in selecting appropriate formats for assessment content. In F. Drasgow (Eds). *Technology and testing: Improving educational and psychological measurement* (pp. 79-103). New York: Routledge.

Powers, D., & O'Neill, K. (1993). Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills. *Educational Assessment, 1,* 153–173.

Prometric, Inc. (2011). The right approach to for Indian Institute of Management (IIM) Test Delivery. *Case Sudy*. Downloaded January 24 2017, from the World Wide Web: https://www.prometric.com/en-us/news-and-resources/case-studies/documents/TestDelivery-IIM.pdf

Rabinowitz, S., & Brandt. S. (2001). Computer-based assessment: Can it deliver on its promise? *WestEd Knowledge Brief*. Downloaded October 10 2015, from the World Wide Web: http://www.wested.org/online_pubs/kn-01-05.pdf

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhager, Denmark: Danish Institute for Educational Research.

R Core Team (2016). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL* http://www.R-project.org/

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*(4), 401-412.

Reckase, M. D. (2009). *Multidimensional item response theory*. East Lansing MI: Springer.

Reckase, M. D., & McKinley, R. L. (1991) The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*(4), 361-373.

Reese, L. M., & Pashley, P. J. (1999). Impact of local item dependence on true-score equating. *LSAC Research Report Series*.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking cur- riculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: Kluwer.

Roberts, J. S., Donoghue, J. R. & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24*(1), 3-32.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40,* 163-184.

Romhild, A., Kenyon, D., & MacGregor, D. (2011). Exploring domain general and domain specific linguistic knowledge in the assessment of academic English anguage proficiency. *Language Assessment Quarterly*, *8*, 213–228.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*(3), 425-435.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika,* 53,349-359.

Roussos, L. A., Stout, W. E., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement*, *35*(1), 1-30.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychological Monograph No. 17). Richmond, VA: Psychometric Society.

Samejima, F. (1980). *Is Bayesian estimation proper for estimating the individual's ability* (Research Rep.80-3). Knoxville, TN: University of Tennessee, Department of Psychology.

San Martin, E., Gonzalez, J., & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3PL model. *Psychometrika*, *80*(2), 450-467.

Sato, E., Rabinowitz, S., Gallagher, C., & Huang, C.-W. (2010). *Accommodations for English Language Learner students: The effect of linguistic modification of math test item sets.* Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from U.S. Department of Education website: http://ies.ed.gov/ncee/edlabs/regions/west/pdf/REL_20094079.pdf

Scalise, K. (2012). Creating innovative assessment items and test forms. In R. Lissitz, & H. iao (Eds.). *Computers and their impact on state assessments* (pp. 133-155). USA: Information Age Publishing Inc.

Scalise, K. (2009). *Computer-based assessment: "Intermediate constraint" questions and tasks for technology platforms*. Downloaded October10 2015, from the World Wide Web: http://pages.uoregon.edu/kscalise/taxonomy/taxonomy.html

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 4*(6), 1-45. Retrieved from http://www.jtla.org

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 307-354). Westport, CT: Praeger.

Secada, W. G. (1992). Race, ethnicity, social class, language, and achievement in mathematics. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 623–660). New York: Macmillan.

Shaftel, J., Belton-Kocher, E., Glasnapp, D. & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of test items of English language learners and students with disabilities. *Educational Assessment*, *11*(2), 105-126.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*(3), 215-232.

Sick, J. (2010). Rasch measurement in language education: Assumptions and requirement of Rasch measurement. *JALT Testing & Evaluation SIG Newsletter*, *14*(2), 23-29.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477–481.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 39-64). Charlotte, NC: Information Age Publishing, Inc.

Sireci, S. G., Thissen, D., Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: Inpursuit of improved construct representation. In S. M. Downing & T. M. Haladyna, (Eds.), *Handbook of test development* (pp. 329-347). Mahwah, NJ: Lawrence Erlbaum Associates.

Smarter Balanced News. (May/June 2014). The field test: A teacher perspective. Retrieved from http://smarterbalanced.createsend1.com/t/ViewEmail/r/49C8B264EC25CCBF2540EF23F30FEDED/CECFC392B969A1D8B3138EAD4DECE712

Smarter Balanced Assessment Consortium (SBAC). (2015). End of Grant Report. Retrieved on February 2016 https://www.smarterbalanced.org/

Smarter Balanced Assessment Consortium (SBAC). (2016a). Smarter Balanced Assessment Consortium: 2014-15 Technical Report. Retrieved on February 2016 https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf

Smarter Balanced Assessment Consortium's (SBAC) (2016b). Training Modules. Retrieved on February 2016 https://www.smarterbalanced.org/

Snow, C. E. (2008). What is the vocabulary of science? In A. S. Rosebery, & B. Warren (Eds.), *Teaching science to English language learners: Building on students' strengths* (pp. 71-83). National Science Teachers Association.

Snow, C. E. (2010). Academic language and the challenge of reading for learning. *Science*, *328*(5977), 450–452.

Snow, R. E. (1994). Abilities in academic tasks. In R. J. Sternberg & R. K. Wagner (Eds.), *Mind in context* (pp. 3-37). Cambridge: Cambridge University Press.

Stark, S., Chernyshenko, O., & Drasgow, F. (2002). *Technical report: Investigating the effects of local dependence on the accuracy of irt ability estimation*. American Institute of Certified Public Accountants.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of Experimental Psychology*. New York: John Wiley & Sons.

Stone, E., Laitusis, C. C., & Cook, L. L. (2016). Increasing the accessibility of assessments through technology. In F. Drasgow (Eds). *Technology and testing: Improving educational and psychological measurement* (pp. 217-234). New York: Routledge.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality.

sychometrika, *52*, 589-617.

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67*, 485 518.

Stout, W. E., Habing, B., Douglas, J., Kim, H. R., Roussos, L. A., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. Applied Psychological Measurement, *20*(4), 331-354.

Stufflebeam, D. L. (2001). Evaluation models. *New Directions for Evaluation*, *89*, 7-98.

Spanos, G., Rhodes, N. C., Dale, T. C., & Crandall, J. (1988). *Linguistic features of mathematical problem solving: Insights and applications*. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 221–240). Hillsdale, NJ: Erlbaum.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure ofresponses to test items. *Applied Psychological Measurement, 27*, 159-203.

Taylor, C., Kirsch, I., Eignor, D., & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning, 49,* 219-274.

Thissen, D., Bender, R., Chen, W., Hayashi, K., & Wiesen, C. A. (1992). *Item response theory and local dependence: A preliminary report* (Research Memorandum 92-2). Chapel Hill: L. L. Thurstone Laboratory, University of North Carolina at Chapel Hill.

Thissen, D. J., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567-577.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple categorical-response models. *Journal of Educational Measurement,* 26,247-260.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Wainer, H., & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31*, 113-123.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). Universal design applied to large scale assessments (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved October 10 2015, from the World Wide Web: https://osepideasthatwork.org/Toolkit/pdf/Universal_Design_LSA.pdf

Thompson, S., Thurlow, M., & Malouf, D. B. (2004). Creating better tests for everyone through universally designed assessments. *Journal of Applied testing technology.* Retrieved October 10, 2015, from the World Wide Web: http://www.testpublishers.org/assets/documents/volume%206%20issue%201%20Creating%20%20better%20tests.pdf

Thompson, T. D., & Pommerich, M. (1996). E*xamining the sources and effects of local dependence*. Paper presented at the American Educational Research Association. Retrieved October 16, 2016 from the ERIC database http://files.eric.ed.gov/fulltext/ED400311.pdf

Thorndike, R. L. (1971). *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16, 433-449.

Traub, R. E. (1993). On the equivalence of traits assessed by multiple-choice and constructed response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1-27). Hillsdale, NJ: Lawrence Erlbaum Associates.

Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiplechoice tests. *Applied Psychological Measurement*, *1*(3), 355-369.

Traxler, C. (2000). The Stanford Achievement Test, ninth edition: National norming and performance standards for deaf and hard of hearing students. *Journal of Deaf Studies and Deaf Education, 5*(4), 337–348.

Tsai, T. H., & Hsu, Y. C. (2005, April). *The use of information entropy as a local item dependence assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics, 5,* 99-114.

Turner, J. D., & Danridge, J. C. (2014). Accelerating the College and Career Readiness of diverse K-5 literacy learners. *Theory Into Practice*, *53*, 212-219.

University of Nevada of Las Vegas. (2015). Michigan Test of English Language Proficiency. Retrieved October 26, 2016 from https://www.unlv.edu/elc/mtelp

Urry, V. W. (1974). Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement*, *34*, 253-269.

Valdes, G., Menken, K., & Castro, M. (2015). *Common Core bilingual and English language learners: A resource for educators.* Philadelphia: Caslon Publishing.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer.

Vector Psychometric Group. (2017). flexMIRT® version 3.0 RC: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group, LLC.

Venables, W. N., & Ripley, B. D. (2001). *Modern applied statistics with S*. (4th ed.). New York: Springer.

Wainer, H., & Feinberg, R. (2015). For want of a nail: Why unnecessarily long tests may be impeding the progress of Western civilization. *Significance*, *12*(1), 16-21.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement,* 24, 185-202.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement,* 27(1), 1-14.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, *12*, 339-368.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*, 103-118.

Wainer, H., & Thissen, D. (1993). How is reliability realted to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, *15*(1), 22-29.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*(3), 203-220.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensionalproficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40,* 255-275. doi: 10.1111/j.1745-3984.2003.tb01107.x.

Wan, L., & Henley, G. A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education*, *25*(1), 58-78.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement*, *26*, 109-128.

Wang, W., & Wilson, M. (2005). The Rasch Testlet Model. *Applied Psychological Measurement*, *29*(2), 126-149.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement*, *6*, 1-11.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-responseand machine-scorable forms of a test. *Journal of Educational Measurement, 17,* 11-29.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427-450.

Whitely, S. E., & Dawis, R. V. (1976). The influence of test context on item difficulty. *Educational and Psychological Measurement*, *36*(2), 329-337.

Wilson, M. R., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, *60*,181-198.

Wolf, M., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment*, *14*, 139-159.

Wolf, M. K., & Faulkner-Bond, M. (2016). Validating English language proficiency assessment uses for English learners: Academic language proficiency and content assessment performance. *Educational Measurement: Issues and Practice*, *35*(2), 6-18.

Wolf, M. K., Wang, Y., Huang, B. H., Blood, I. (2014). Investigating the language demands in the Common Core State Standards for English language learners: A comparison study of standards. *Middle Grades Research Journal*, *9*(1), 2014, 35–52.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement, 47*(3), 339-360.

Yao, L., & Boughton, K. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31*(2), 83-105, DOI:10.1177/0146621606291559.

Yao, L., & Boughton, K. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement*, *46*(2), 177-197.

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, *30*(6), 469-492. DOI: 10.1177/0146621605284537

Yan, J. (1997). Examining local item dependence effects in a large-scale science assessment by a Rasch Partial Credit Model. Paper presented at the Annual Meeting of the American Educational Research Association. Retrieved October 13, 2016 from the ERIC database http://files.eric.ed.gov/fulltext/ED412219.pdf

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125-145.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 155-186). Westport, CT: Praeger.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement*, *39*(4), 291-309.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2003). Effects of local item dependence on the validity of IRT item, test, and ability statistics, Medical College Admission Test, AAMC. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.5458&rep=rep1&type=pdf

Zenisky, A. L., & Sireci, S. G. (2001). *Technical report: Feasibility review of selected performance assessment item types for the computerized uniform CPA exam*. American Institute of Certified Public Accountants.

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education*, *15*, 337-362

Zhang, J., & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika*, *64*(2), 129-152.

Zhang, J., & Stout, W. (1999b). The theoretical DETECT index of dimensionality and itsapplication to approximate simple structure. *Psychometrika, 64*(2), 213-249.

APPENDIX A

SELECTED-RESPONSE ITEM FORMAT FROM SMARTER BALANCED ASSESSMENT CORPORATION ELA ITEM DESIGN TRAINING MODULE (RETRIEVED ON DECEMBER, 2015)

Grade	: 4	**DOK**	: 2
Claim 1	: Students can read closely and analytically to comprehend a range of increasingly complex literary and informational texts		
Target 14 LANGUAGE USE	: Determine or interpret figurative language/literary devices or connotative meanings of words and phrases used in context and the impact of those word choices on meaning and tone.		

Amelia Earhart

Amelia Earhart Learns to Fly

Amelia Earhart was born in Atchison, Kansas, on July 24, 1897. In those days, airplanes were not nearly as common as they are today. Earhart was 12 years old before she ever saw an airplane, and she did not take her first flight until 1920. Amelia Earhart was so thrilled by her first airplane ride that she quickly began to take flying lessons. She wrote, "As soon as I left the ground, I knew I myself had to fly."

Earhart excelled as a pilot. Her first instructor was Neta Snook, one of the first women to graduate from the Curtiss School of Aviation. Earhart borrowed money from her mother to buy a two-seat plane. She got her U.S. flying license in December 1921, and by October 1922, she set an altitude record for women of 14,000 feet. In 1923, Earhart received her international pilot's license - only the 16th woman to do so. At the same time, she was becoming famous for her aviation achievements.

Amelia Earhart Flies Across the Atlantic

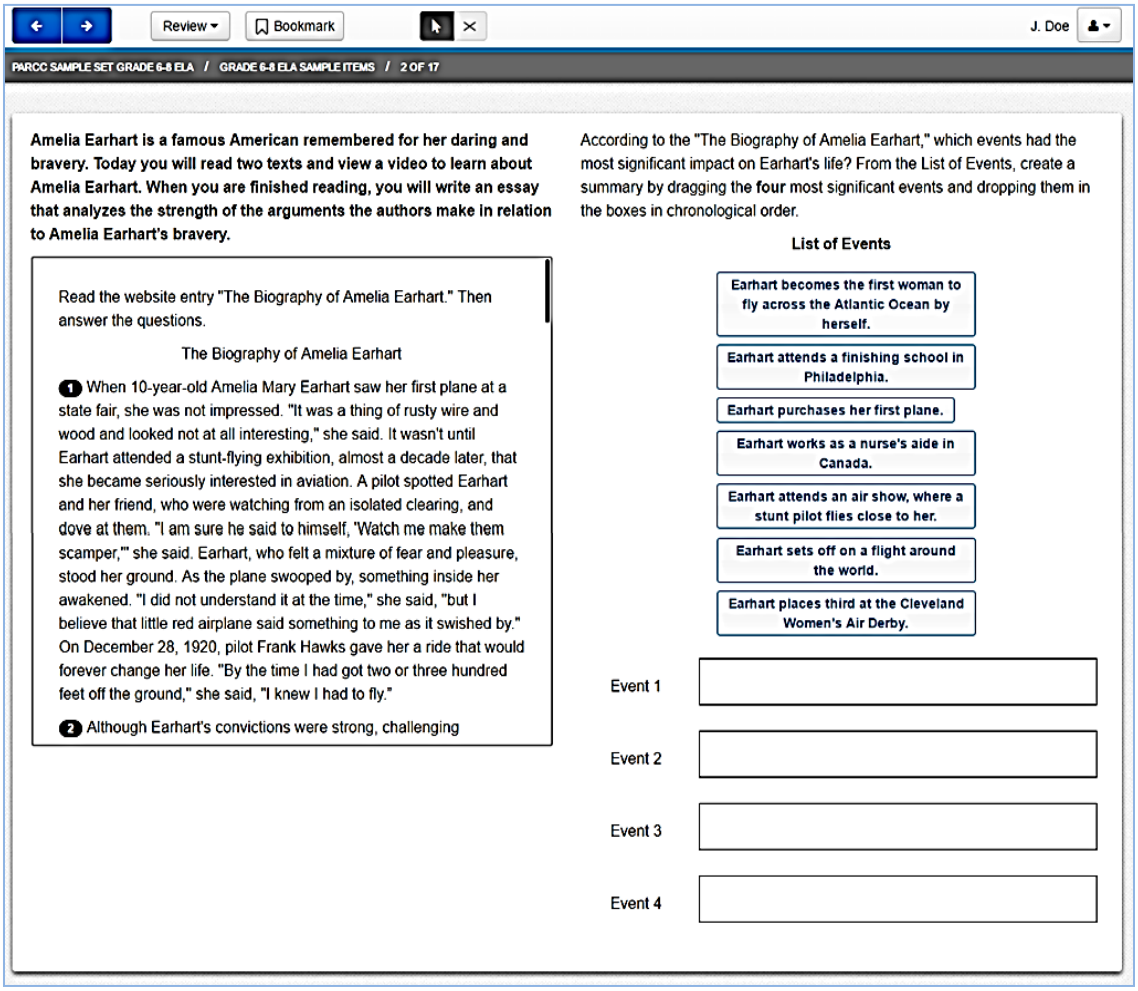
In 1928, Amelia Earhart received a phone call that would change her life. She was invited to become the first woman passenger to cross the Atlantic Ocean in a plane. "The idea of just going as 'extra weight' did not appeal to me at all," she said, but she accepted the offer nonetheless. On June 17, after several delays due to bad weather, Amelia Earhart flew in a plane named Friendship with co-pilots Wilmer "Bill" Stultz and Louis "Slim" Gordon. The plane landed at Burry Port, South Wales, with just a small amount of fuel left.

Amelia said, "The idea of just going as 'extra weight' did not appeal to me at all." What does the phrase 'extra weight' refer to?

A. Her fame as an international pilot
B. Her role as a passenger on the plane
C. Her understanding of how heavy she was
D. Her awareness of how she was making history

APPENDIX B

TECHNOLOGY-ENHANCED ITEM FORMAT FOR ELA. PEARSON EDUCATION: PARTNERSHIP FOR ASSESSMENT OF READINESS FOR COLLEGE AND CAREERS (PARCC) ASSESSMENT (2015)



Source: https://parcctrng.testnav.com/client/index.html#login?username=LGN102605442&password=3JETPVHA

APPENDIX C

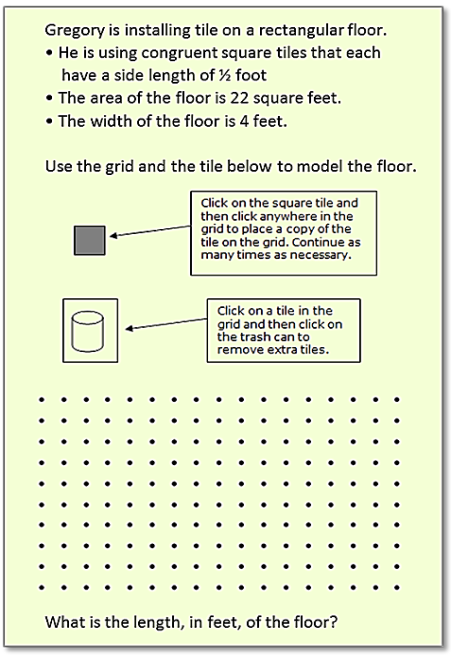
TWO TYPES OF SBAC ITEM FORMATS:

(A) TECHNOLOGY-ENABLED ITEM FORMAT,

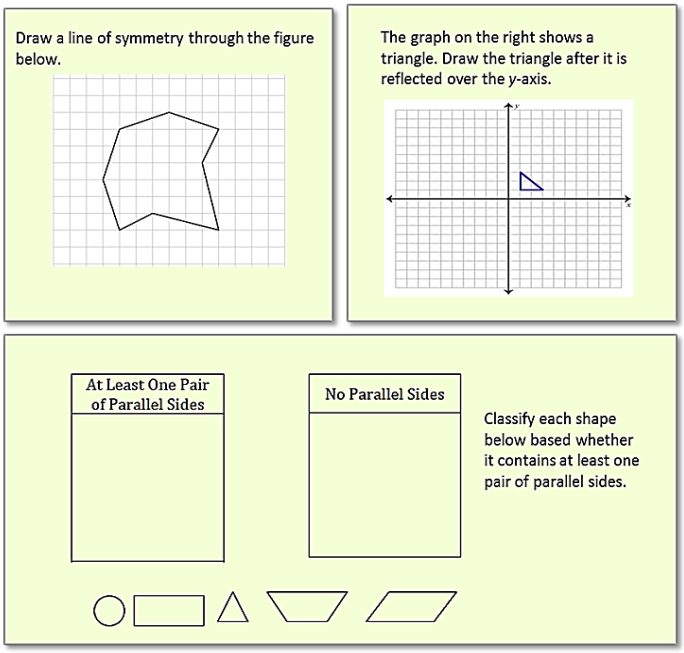
(B) TECHNOLOGY-ENHANCED ITEM FORMAT

FROM SBAC MATHEMATICS ITEM DESIGN TRAINING MODULE

(RETRIEVED ON DECEMBER, 2015)



(A)



(B)

APPENDIX D

GRIDDED RESPONSE ITEM FORMAT

(STATE OF FLORIDA DEPARTMENT OF EDUCATION, 2013)

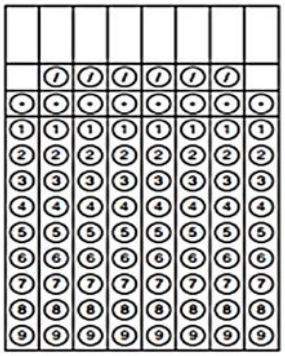
Sample Item 2

Grade/Course	Item Type	DOK	NGSSS Benchmark	CCSS Benchmark	Point Value
912/ Chemistry	GR	2	SC.912.P.8.9: Apply the mole concept and the law of conservation of mass to calculate quantities of chemicals participating in reactions.	MACC.912.N-Q.1.1: Use units as a way to understand problems and to guide the solution of multi-step problems; choose and interpret units consistently in formulas; choose and interpret the scale and the origin in graphs and data displays.	1

Use the following chemical equation to answer the question:

$2CH_4(g) + S_8(s) \rightarrow 2CS_2(l) + 4H_2S(g)$

Calculate the mass of CS_2 when 10.0 g of CH_4 react with 80.0 g of S_8 to produce 47.5 g of H_2S.



Source: http://www.fldoe.org/core/fileparse.php/5423/urlt/FL-Item-Spec-SCI-Chemistry-WT-r2g.pdf

APPENDIX E

DESCRIPTIVE STATISTICS FOR CONDITIONAL e0 (BASED ON PERCENTAGE SCORES)

1000 examinees																	
			10 items per dimension								20 items per dimension							
nuisance	vcor	ncor	low item discrimination				high item discrimination				low item discrimination				high item discrimination			
			mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max
0	0.40	NA	-0.68	4.77	-11.58	9.20	0.07	6.88	-15.85	23.49	-0.26	3.87	-10.1	10.54	-0.02	1.10	-3.38	3.23
	0.80		-0.15	5.17	-12.12	11.77	-0.64	8.14	-21.5	21.46	0.00	4.73	-14.16	13.18	0.02	1.13	-4.87	3.46
1	0.40	NA	-0.37	6.69	-18.4	17.37	0.13	7.55	-20.77	20.30	-0.18	5.43	-12.72	14.22	-0.01	1.05	-2.97	4.69
	0.80		-0.20	6.55	-18.37	18.84	-0.16	9.00	-29.77	26.44	-0.09	6.57	-20.88	18.33	-0.02	1.10	-3.66	3.96
2	0.40	0.00	-0.37	7.22	-17.72	22.13	-0.83	8.27	-21.75	18.62	0.16	7.19	-20.21	20.54	0.05	1.11	-2.85	3.98
		0.40	-0.17	7.98	-21.82	23.24	0.65	8.89	-25.11	27.59	0.08	7.50	-20.10	24.76	0.07	1.11	-3.76	3.53
		0.70	-0.19	9.03	-28.23	24.08	0.23	9.36	-26.89	24.59	0.50	8.01	-23.66	28.25	-0.01	1.09	-3.60	4.48
	0.80	0.00	-0.03	8.22	-24.66	18.96	0.32	9.61	-31.91	24.57	0.02	7.09	-17.69	20.88	-0.01	1.04	-3.97	3.13
		0.40	-0.31	8.41	-23.20	27.54	-0.02	10.33	-30.49	27.48	0.01	7.52	-21.54	23.26	0.03	1.04	-3.67	2.57
		0.70	0.09	9.12	-23.76	25.39	-0.39	10.59	-30.61	26.62	-0.22	8.38	-22.62	25.93	-0.03	1.09	-4.25	3.05

5000 examinees																		
			10 items per dimension								20 items per dimension							
nuisance	vcor	ncor	low item discrimination				high item discrimination				low item discrimination				high item discrimination			
			mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max	mean	*SD*	min	max
0	0.40	NA	-0.16	5.98	-14.5	12.16	-0.19	8.04	-21.39	19.87	0.16	4.77	-12.16	14.07	0.06	1.08	-3.51	4.88
	0.80		0.14	6.41	-16.64	18.27	-0.13	9.85	-29.79	23.40	0.05	5.38	-13.71	14.62	-0.02	0.98	-4.58	4.89
1	0.40	NA	-0.27	7.61	-20.67	19.69	-0.16	9.17	-23.09	26.53	0.16	6.52	-20.11	18.17	0.02	0.95	-4.08	3.55
	0.80		0.01	7.94	-25.19	17.03	-0.18	10.57	-34.59	27.38	0.21	7.02	-20.83	19.21	0.00	0.93	-5.25	3.55
2	0.40	0.00	0.11	9.01	-31.88	21.28	0.19	10.12	-28.38	26.20	0.15	7.97	-22.70	23.17	0.00	0.89	-3.97	3.35
		0.40	-0.58	9.86	-29.29	22.68	0.37	10.19	-25.75	24.03	0.22	8.02	-22.74	22.33	0.00	0.92	-4.09	3.33
		0.70	0.02	9.97	-26.14	26.61	0.00	10.73	-26.69	31.17	-0.02	9.12	-27.99	33.08	0.00	0.88	-3.76	3.73
	0.80	0.00	0.06	8.90	-24.57	21.55	-0.19	10.56	-42.12	27.90	0.41	8.00	-22.86	20.57	-0.02	0.86	-3.82	3.97
		0.40	0.01	9.74	-30.00	26.25	-0.76	11.46	-39.68	24.67	0.20	8.63	-22.73	22.07	-0.01	0.92	-3.57	4.43
		0.70	0.35	10.42	-31.36	32.65	-0.11	11.15	-37.04	28.39	0.34	9.87	-26.41	26.97	0.01	0.86	-3.18	3.21

Note. nuisance:number of nuisance dimension(s) present; vcor: correlations between primary dimensions; ncor: correlation between two nuisance dimensions