

May 2014

# The Performance of the Linear Logistic Test Model When the Q-Matrix is Misspecified: A Simulation Study

George T. Macdonald

*University of South Florida*, [gmacdona@mail.usf.edu](mailto:gmacdona@mail.usf.edu)

Follow this and additional works at: <http://scholarcommons.usf.edu/etd>

 Part of the [Educational Assessment, Evaluation, and Research Commons](#)

---

## Scholar Commons Citation

Macdonald, George T., "The Performance of the Linear Logistic Test Model When the Q-Matrix is Misspecified: A Simulation Study" (2014). *Graduate Theses and Dissertations*.  
<http://scholarcommons.usf.edu/etd/5065>

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact [scholarcommons@usf.edu](mailto:scholarcommons@usf.edu).

The Performance of the Linear Logistic Test Model When the Q-Matrix is  
Misspecified: A Simulation Study

by

George T. MacDonald

A dissertation submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
in Curriculum and Instruction  
with an emphasis in Measurement and Evaluation  
Department of Educational Measurement and Research  
College of Education  
University of South Florida

Co-Major Professor: Jeffrey D. Kromrey, Ph.D.  
Co-Major Professor: Yi-Hsin Chen, Ph.D.  
Robert F. Dedrick, Ph.D.  
Gladis Kersaint, Ph.D.

Date of Approval:  
November 14, 2013

Keywords: Cognitive Diagnostic Assessment, LLTM

Copyright © 2013, George T. MacDonald

## DEDICATION

I dedicate this dissertation to Paula J. Baxendale-MacDonald, my wife, without whom this would not have been possible.

I re-dedicate this dissertation to Stephen, Chloe, Emily, Michael, Robert and Ashleigh the light of my life.

Additionally, I re-dedicate this dissertation to Rev. Donald J. MacDonald, and Marjorie A. MacDonald, my parents.

## ACKNOWLEDGMENTS

Committee Members: I chose my committee members because I knew them to be demanding mentors. This dissertation would not have been possible without (a) Dr. Yi-Hsin Chen who issued my first research invitation in the summer of 2008, and introduced me to LLTM; (b) Dr. Jeffrey Kromrey who guided all aspects of my work; (c) Dr. Robert Dedrick who asked important questions; and (d) Dr. Gladis Kersaint who showed me the world of mathematics education.

Mentors: Additionally, I was supported and guided by Dr. Kevin Kieffer, Dr. John Ferron, and Dr. Patrick Draves (deceased).

Colleagues: I have spent countless hours in consultation with Theodore Dwyer, and Reginald Lee.

Directors and Staff of the David C. Anchin Center: I have been supported by a Graduate Research Assistantship, the inaugural Tampa Bay Educational Partnership (TBEP) Fellowship, and an Assistant Directorship in Research and Grant Development. I express my appreciation to Dr. Bruce Jones, Dr. Gladis Kersaint, and Dr. Brian Mann.

Hillsborough County Public Schools (HCPS): My Fellowship was jointly supported through the Anchin Center and HCPS and I express my appreciation to Dr. Pansy Houghton.

A special thank you to Dr. Douglas Lunsford who recommended I pursue a degree in educational research and measurement.

## TABLE OF CONTENTS

LIST OF TABLES.....	iv
LIST OF FIGURES .....	vi
ABSTRACT.....	viii
CHAPTER ONE: INTRODUCTION.....	1
Background.....	1
Cognitive Diagnostic Assessment .....	1
Linear Logistic Test Model .....	2
Adoption and Use of the LLTM.....	4
Model of Interest and Statistical Program.....	4
The Weight Matrix or More Commonly the Q-matrix .....	4
Summary of Issues of Interest .....	5
Purpose of the Study .....	6
Research Questions .....	7
Research Hypothesis and Research Questions.....	7
Study Significance .....	8
Definitions .....	8
CHAPTER TWO: LITERATURE REVIEW.....	11
Educational Uses of LLTM.....	12
Mathematical Models .....	12
Geometric and Spatial Models .....	21
Paragraph Comprehension and Reading Models.....	23
Other Approaches.....	27
Fischer’s Contribution.....	28
Normalization Constant .....	29
Cognitive Component Coefficient Estimation.....	30
Item Difficulty Computation.....	32
Person Ability .....	35
Estimation Approaches.....	35
Model of Interest.....	37
SAS 9.3 and Proc NLMIXED .....	37
General Comments .....	37
Q-matrix Simulation Literature .....	38
Embretson’s Contribution .....	44
Other Q-matrix Simulation Studies in the CDA Family .....	44
The MacDonald LLTM and Simulation Studies .....	47
Summary.....	50
CHAPTER THREE: METHOD .....	51
Study Purpose.....	51
Research Questions .....	51

SAS 9.3.....	52
NLMIXED.....	52
Estimation Technique .....	52
Replications.....	53
Cognitive Components .....	53
SAS Code .....	53
Code Validation .....	53
The Simulation .....	54
Q-matrix .....	55
Percentage of Q-Matrix Misspecification.....	55
Form of Misspecification.....	56
Misspecification .....	56
Under-Specification.....	56
Balanced Misspecification.....	56
Over-specification .....	57
Sample Size .....	57
Q-Matrix Density.....	57
Dense Matrix.....	57
Sparse Matrix.....	58
Test Length .....	59
Skewness of Person Ability Distribution.....	60
Parameters of Interest .....	60
Evaluation Criteria .....	61
Statistical Bias .....	61
Root Mean Square Error (RMSE) .....	62
Confidence Interval Coverage and Confidence Interval Width.....	63
Cognitive Components Pairwise Phi Correlation Coefficients .....	64
Data Analysis .....	66
Major Steps in the SAS Simulation Code.....	67
CHAPTER FOUR: RESULTS .....	68
Estimated Bias .....	71
Percent of Misspecification .....	74
Form of Misspecification .....	76
Percent of Misspecification by Form of Misspecification .....	78
Density of the Q-matrix .....	79
Percent of Misspecification by Density of the Q-matrix .....	80
Number of Items .....	82
Percent of Misspecification by Number of Items .....	83
Summary for Estimated Mean Bias Results .....	84
Root Mean Squared Error.....	86
Percentage of Misspecification.....	89
Form of Misspecification .....	91
Percent of Misspecification by Form of Misspecification .....	92
Sample Size.....	94
Density of the Q-matrix .....	95
Number of Items .....	96
Percent of Misspecification by Number of Items .....	98
Summary for Mean Estimated Root Mean Squared Error .....	99
Estimated Confidence Interval Coverage.....	101
Percentage of Misspecification.....	104

Form of Misspecification .....	106
Sample Size.....	108
Percent of Misspecification by Sample Size .....	109
Number of Items .....	111
Summary of Estimated Confidence Interval Coverage .....	112
Confidence Interval Width.....	113
Sample Size.....	116
Number of Items .....	117
Summary of Confidence Interval Width .....	118
Phi Pairwise Cognitive Components Coefficients .....	119
Performance of SAS PROC NLMIXED .....	122
CHAPTER FIVE: DISCUSSION.....	123
The Impact on Parameters of Misspecifying the Q-matrix.....	125
Under-specifying, Over-specifying or Balanced Misspecifying the Q-Matrix.....	128
Misspecification Interacting with Sample Size.....	129
Misspecification Interacting with Dense or Sparse Q-matrices.....	130
Misspecification Interacting with Number of Items .....	131
Misspecification Interacting with Skewness of the Person Distribution.....	132
Conclusions.....	132
Limitations of the Study and Future Research .....	133
REFERENCES .....	135
APPENDIX A: PROGRAMMING CODE.....	142
The Item Matrix, Cognitive Components, Beta LLTM and Theta.....	142
Impose Misspecification on the A Matrix.....	145
Send Generated Data to SAS for Analysis .....	146
LLTM Specified using Proc NLMIXED .....	147
Cognitive Components Analysis .....	148
Beta LLTM Analysis.....	150
Theta Analysis.....	150
ABOUT THE AUTHOR .....	End Page

## LIST OF TABLES

Table 1	Educational Uses of LLTM from 1972-2011 .....	13
Table 2	Q-matrix for Cognitive Components Transformation (CC1) and Division (CC2) .....	31
Table 3	A 20 by 5 True Q-matrix that Describes the Relationship between Items and Cognitive Components .....	55
Table 4	Dense 5 by 20 True Q-matrix (60% 1s).....	58
Table 5	Sparse 20 by 5 Q-matrix (46% 1S) .....	59
Table 6	Phi Pairwise Cognitive Component Correlation Coefficients Exemplifying Negative, Positive, and Neutral Phi Correlations.....	65
Table 7	Major Steps in SAS Code .....	67
Table 8	Means, Standard deviations, Minimum, and Maximum Bias Values for Beta, CC1-CC5, and Theta (N=1890) .....	71
Table 9	Eta-Squared Values for the Association of Design Factors and 1st Level Interaction Effects with the Average Estimated Overall Bias for CCs, Beta, and Theta .....	73
Table 10	Summary of Estimated Mean Bias Results .....	85
Table 11	Mean, Standard Deviation, Minimum, and Maximum values for the Estimated RMSE for CC1 - CC5, Beta, and Theta.....	87
Table 12	Eta-Squared Values for the Association of Design Factors and 1st Level Interaction Effects with the Average Estimated Overall Root Mean Squared Error for CCs, Beta, and Theta .....	88
Table 13	Summary of Estimated Mean RMSE Results.....	100
Table 14	Mean, Standard Deviation, Minimum, and Maximum Values for the Estimated Confidence Interval Coverage for CC1-CC5 and Theta .....	102
Table 15	Eta-Squared Values for the Association of Design Factors and 1st Level Interaction Effects with the Average Estimated Overall Confidence Interval Coverage for CCs and Theta .....	103
Table 16	Summary of Mean Confidence Interval Coverage.....	112
Table 17	Mean, Standard Deviation, Minimum, and Maximum Values for Confidence Interval width for CC1-CC5 and theta .....	114



Table 18	Eta-Squared Values for the Association of Design Factors and 1st Level Interaction Effects with the Average Overall Confidence Interval Width for CCs and Theta .....	115
Table 19	Summary of Mean Confidence Interval Width .....	118
Table 20	Pairwise Cognitive Component Phi Correlation Coefficients for Truth, Under, Over, and Balanced Misspecification by Percent of Misspecification in the Sparse Q-matrix .....	119
Table 21	Pairwise Cognitive Component Phi Correlation Coefficients for Truth, Under, Over, and Balanced Misspecification by Percent of Misspecification in the Dense Q-matrix .....	120

## LIST OF FIGURES

Figure 1. Item 5 from 23 Item Taiwanese fraction test. ....	33
Figure 2. Item 13 from a 23 Item Taiwanese Fraction Test. ....	34
Figure 3. Distribution of average overall statistical bias estimates across beta, CC1-CC5, and theta. ....	72
Figure 4. Average estimated bias for beta, CC1-CC5, and theta by percent of misspecification. ....	75
Figure 5. Average estimated bias for beta, CC1-CC5, and theta by form of misspecification. ....	77
Figure 6. Average estimated bias for CC1, CC4, and CC5 for percent of misspecification by form of misspecification.....	79
Figure 7. Average estimated bias for beta, CC1-CC5, and theta by density of the Q-matrix.....	80
Figure 8. Average estimated bias for CC1 & CC3 for percent of misspecification by density of the Q-matrix.....	81
Figure 9. Average estimated bias for beta, CC1-CC5, and theta by number of items.....	82
Figure 10. Average estimated bias for beta, and theta for percent of misspecification by number of items.....	84
Figure 11. Distribution of estimated RMSE for CC1-CC5, beta, and theta. ....	87
Figure 12. Average estimated root mean squared error for beta, CC1-CC5, and theta by percent of misspecification.....	90
Figure 13. Average estimated root mean squared error for beta, CC1-CC5, and theta by form of misspecification. ....	92
Figure 14. Average estimated root mean squared error for CC5 for percent of misspecification by form of misspecification.....	93
Figure 15. Average estimated root mean squared error for beta, CC1-CC5, and theta by sample size.....	94
Figure 16. Average estimated root mean squared error for beta, CC1-CC5, and theta by density of the Q-matrix.....	96

Figure 17. Average estimated root mean squared error beta, CC1-CC5, and theta by number of items.....	97
Figure 18. Average root mean squared error for theta for percent of misspecification by number of items.....	98
Figure 19. Distribution of estimated confidence interval coverage for CC1-CC5 and theta. ....	102
Figure 20. Average estimated confidence interval coverage for CC1-CC5 and theta by percent of misspecification.....	105
Figure 21. Average estimated confidence interval coverage for CC1-CC5 and theta by form of misspecification. ....	107
Figure 22. Average estimated confidence interval coverage for CC1-CC5 and theta by sample size.....	109
Figure 23. Average estimated confidence interval coverage CC1, CC3, and CC4 for percent of misspecification by sample size.. ....	110
Figure 24. Average estimated confidence interval coverage for CC1-CC5 and theta.....	111
Figure 25. Distribution of the confidence interval width for CC1 - CC5 and theta. ....	113
Figure 26. Average confidence interval width for CC1-CC5 and theta by sample size. ....	116
Figure 27. Average estimated confidence interval width for CC1-CC5 and theta. ....	117
Figure 28. Average pairwise phi correlation coefficients by percent of misspecification for dense and sparse Q-matrices in all form of misspecification.....	121

## ABSTRACT

A simulation study was conducted to explore the performance of the linear logistic test model (LLTM) when the relationships between items and cognitive components were misspecified. Factors manipulated included percent of misspecification (0%, 1%, 5%, 10%, and 15%), form of misspecification (under-specification, balanced misspecification, and over-specification), sample size (20, 40, 80, 160, 320, 640, and 1280), Q-matrix density (60% and 46%), number of items (20, 40, and 60 items), and skewness of person ability distribution (-0.5, 0, and 0.5). Statistical bias, root mean squared error, confidence interval coverage, confidence interval width, and pairwise cognitive components correlations were computed. The impact of the design factors were interpreted for cognitive components, item difficulty, and person ability parameter estimates.

The simulation provided rich results and selected key conclusions include (a) SAS works superbly when estimating LLTM using a marginal maximum likelihood approach for cognitive components and an empirical Bayes estimation for person ability, (b) parameter estimates are sensitive to misspecification, (c) under-specification is preferred to over-specification of the Q-matrix, (d) when properly specified the cognitive components parameter estimates often have tolerable amounts of root mean squared error when the sample size is greater than 80, (e) LLTM is robust to the density of Q-matrix specification, (f) the LLTM works well when the number of items is 40 or greater, and (g) LLTM is robust to a slight skewness of the person ability distribution. In sum, the LLTM is capable of identifying conceptual knowledge when the Q-matrix is properly specified, which is a rich area for applied empirical research.

## CHAPTER ONE: INTRODUCTION

### **Background**

In the report, *Adding it Up: Helping Children Learn Mathematics*, the Mathematics Learning Committee from the National Research Council (Kilpatrick, Swafford, & Findell, 2001) suggested that student learning and performance could be enhanced if conceptual understanding was taught at the same time as procedural fluency. This major new direction placed a great emphasis on the idea of conceptual understanding as a main strand within mathematics education. It would help to define the term concept. A concept can be considered a mental representation of something (Neath & Surprenant, 2003). In that, a concept can be understood as a thought, notion, or mental representation originating in the mind. It builds upon the input of the senses; it builds upon the perceptual knowledge derived from the inputs of the senses; this perceptual knowledge is commonly understood to be encoded, and stored in human memory available for later retrieval; concepts are an abstraction of what is common to a plurality of all possible instances (Solso, Maclin, & Maclin, 2008).

### **Cognitive Diagnostic Assessment**

Cognitive diagnostic assessment (CDA) is an important thrust in measurement designed to assess students' cognitive knowledge structures and processing skills in relation to item difficulty (Leighton & Gierl, 2007). There has been an increasing demand in education to develop assessments that map and measure the psychological processes involved in conceptual understanding. In fact, the Indiana Department of Education issued a Request for Proposal (RFP) in June, 2012 on behalf of *The Partnership for Assessment of Readiness for College and Careers (PARCC)*. PARCC is defined on their website as:

...a consortium of 23 states plus the U.S. Virgin Islands working together to develop a common set of K-12 assessments in English and math anchored in what it

takes to be ready for college and careers. These new K-12 assessments will build a pathway to college and career readiness by the end of high school, mark students' progress toward this goal from 3rd grade up, and provide teachers with timely information to inform instruction and provide student support. The PARCC assessments will be ready for states to administer during the 2014-15 school year.

A major goal of the RFP *Solicitation for: PARCC Item Tryout, Field Testing, Operational Form Construction, and Embedded Research* is diagnostic assessment which they define as:

...designed to measure students' strengths and weaknesses in terms of specific standards or knowledge and skills. These assessments are developed to inform instructional strategies, remediation, and intervention. Development of diagnostic assessments requires a theoretical framework that explains how students learn different skills and knowledge, and how such skills and knowledge are interrelated. Results of such assessments inform what particular standards students need to master in order to master the next set of knowledge and skills within their reach.

If the goal of assessing students' strengths and weaknesses is to be accomplished, it will be important to develop standardized assessments that measure the psychological processes involved in conceptual understanding. The field of CDA in general and the linear logistic test model in particular can be thought of as a response to these emerging educational needs.

### **Linear Logistic Test Model**

Fischer (1973) introduced a model, called the linear logistic test model (LLTM) that is capable of bridging cognitive processing models and psychometric models. He believed that an essential requirement for instruction is the teacher understands the parametric difficulties of curricular material being taught and the teacher's capacity to parse that curriculum into smaller learning steps, which are easy for the student to master. In other words, it is necessary for the teacher to describe the curricular units using quantified parametric descriptions of learning intertwined with modern test theory. He further held that the systematic analysis of subject

matter and the parsing of instructional units by means of quantitative parameters are simplest when dealing with an isolated subject area with a limited number of cognitive operations.

In his mathematics study, Fischer found that differentiating calculus items could be explained by eight basic cognitive operations that the examinee must implement. He postulated that item difficulty could be re-parameterized to express these operations. In that, the linear logistic test model could be used to predict the probability that the person passes a particular item from that person's ability and a cognitive processing model of item difficulty.

Fischer showed that his model is appropriate for assessments which are solved by a linear combination of cognitive operations or rules. His study was conducted on the psychological complexity of problems in elementary differential calculus, as taught in secondary school mathematics. The psychological contribution of this analysis was not limited to a statistical description of item difficulties but included validating the psychological units or cognitive operations used to solve problem. For example, he successfully demonstrated that differentiation of a polynomial is to be considered a single psychological operation, which is mastered and can be combined with other operations. He also argued that the complexity of a task is primarily determined by the combination of different operations and is not increased significantly when the same operation occurs repeatedly within the problem.

The statistical analysis is made up of (a) the estimation of item difficulty (Rasch) for each task (item) according to the dichotomous Rasch' model, (b) the estimation of the cognitive components (cognitive operations), (c) the linear reconstruction of item difficulty (LLTM) from cognitive components, and (d) the statistical comparison between item difficulty-Rasch and item difficulty-LLTM.

Since Fischer built his theories on the Rasch model, LLTM is often understood to be an extension of the Rasch Item Response Theory (IRT) model (Embretson & Reise, 2000), but formally speaking, the LLTM multiplicatively joins cognitive component vectors to student's raw scores.

## **Adoption and Use of the LLTM**

Educational and psychological use of the linear logistic test model from 1973 to 2012 can be thought of as falling into three main categories (a) mathematical models, (b) geometric and spatial models, and (c) paragraph comprehension and reading models. Contributions to the mathematical modeling include: (a) Fischer (1973); (b) Dimitrov (2007); (c) Spada and Kluwe (1980); (d) Embretson and Daniels (2008); (e) Xie and Wilson (2008); (f) Medina-Diaz (2009); and (g) Chen, MacDonald and Leu (2011). Contributions to the geometric and spatial modeling have been made by: (a) Whitley and Schneider (1981); (b) Embretson (1992); and (c) Tanzer, Gitler and Ellis (1995). Contributions to the paragraph comprehension and reading modeling have been made by: (a) Embretson and Wetzel (1987); (b) Sheehan and Mislevy (1990); (c) Cisse (1994); (d) Embretson and Gorin (2001); (e) Sonnleitner (2008); and (f) Holling, Blank, Kuchenbaker, and Kuhn (2008).

## **Model of Interest and Statistical Program**

Fischer's explication of the LLTM is important; however, it was discovered (Chen, MacDonald, & Leu, 2011) that Fischer's LLTM software program (i.e., LCPMWin 1.0) did not operate consistently. About half of the time the software would not run the requested analysis despite following manual instructions exactly. As a result, MacDonald sought a replacement software program to continue working with the LLTM. A reasonable solution was offered by De Boeck and Wilson (De Boeck & Wilson, 2004; Embretson & Yang, 2006) who used the Statistical Analysis System (SAS) to implement LLTM. Their variant of the LLTM as implemented in PROC NLMIXED will be the model of interest; SAS 9.3 (SAS Institute Inc., 2008) will be employed to compute the needed analysis. The SAS code that will be used to analyze the LLTM is found in Appendix A.

## **The Weight Matrix or More Commonly the Q-matrix**

To assess cognitive components a unique feature known as a weight matrix is constructed. Latent attributes or cognitive components can be quantified in the weight matrix



commonly known as a Q-matrix (Tatsuoka, 1983). The Q-matrix can, therefore, be understood to hold the cognitive specifications for test design, test construction, and the decomposition of cognitive components (Leighton, Gierl, & Hunka, 2004). To specify a Q-matrix, a group of content experts can be gathered who determine, based on theory and their experience, what psychological units or cognitive components are required of examinees when they are responding to specific items on a specific assessment. After the content experts group has agreed on the cognitive components contained in the assessment, they are then asked to independently indicate if each cognitive component is required to solve each item on the assessment. If a cognitive component is required to solve an item its weight is 1 and if it is not required to solve an item its weight is 0. Once the content experts have independently created their weight matrix, the whole group is gathered to compare results and reach consensus on a final weight matrix. The gathering of the cognitive components weights into an item by cognitive components matrix has been named a weight matrix and a Q-matrix. For the purposes of this study the term Q-matrix will be used.

### **Summary of Issues of Interest**

In Chapter 4, *The Strands of Mathematical Proficiency in Adding it Up: Helping Children Learn Mathematics*, (Kilpatrick, Swafford, & Findell, 2001) a strong argument is made for conceptual understanding and procedural fluency (strategic competence, adaptive reasoning, and positive disposition round out the five intertwining strands comprising mathematical competency). The field of educational measurement has long been able to measure procedural fluency. The general consensus in classical test theory (CTT) is that computing a mathematical problem correctly indicates mastery of procedural fluency. This capacity to measure procedural fluency has been available to teachers, administrators, and districts since the early part of the 20<sup>th</sup> century. The call for an intertwining of conceptual understanding and procedural flexibility leading to mathematical competency is recent. The capacity to measure the conceptual understanding of students is basically in its nascent period. Without good measurement tools it

is difficult to understand how teachers, administrators, parents, and students could assess students' conceptual understanding. Further, it is hard to understand how a nationalized public school system could intertwine procedural knowledge with conceptual understanding without such tools.

While Fischer began work in the 1970s, interest in statistical models permitting the profiling of respondents according to competencies, traits, or skills began to grow in the 1990s. Models of these types offer diagnostic feedback to respondents as well as criterion-referenced interpretations of multiple proficiencies (Rupp & Templin, 2008). These models (a) must be sound theoretically, (b) require item level response data from the assessment, and (c) require the latent attributes to be measured by the assessment.

Most model fit analysis does not address the correctness of the constructed Q-matrix (de la Torre, 2008). However, the amount of model misspecification that the Q-matrix can tolerate and still function adequately is unknown. Little simulation work examining the sensitivity of the LLTM to misspecification of the Q-matrix has been conducted (Baker, 1993). This is the issue of interest in this research study, to provide some evidence of the functioning of the LLTM when the Q-matrix or Weight Matrix is misspecified.

### **Purpose of the Study**

The purpose of this study is to provide some evidence to help determine if the LLTM functions well and is robust when (a) the Q-matrix is progressively more misspecified, (b) the Q-matrix is properly specified; under-specified, balanced misspecified, and over-specified.; (c) the sample size is varied; (d) the Q-matrix is densely and sparsely populated; (e) test length varies; and (f) the distribution of person responses is normally distributed, negatively skewed, and positively skewed.

## **Research Questions**

Questions 1 through 6 will examine:

- 1.) To what extent does the LLTM function well when the Q-matrix is progressively more misspecified?
- 2.) To what extent does the LLTM function well when the Q-matrix is properly specified, under specified, balanced misspecified, and over specified?
- 3.) To what extent does the LLTM function well under different conditions of model misspecification when the sample size varies?
- 4.) To what extent does the LLTM function well under different conditions of model misspecification when the Q-matrix is densely or sparsely populated?
- 5.) To what extent does the LLTM function well under different conditions of model misspecification when test length varies?
- 6.) To what extent does the LLTM function well under different conditions of model misspecification when the population distribution is normally distributed, negatively skewed, and positively skewed?

## **Research Hypothesis and Research Questions**

Results (MacDonald & Kromrey, 2011, 2012) from two earlier simulation studies have situated these research questions. It was concluded that the LLTM model works very, very well when the Q-matrix is correct and that PROC NLMIXED is superb at estimating these models (MacDonald & Kromrey, 2011). The results have shown that (a) as the Q-Matrix moves away from the truth cognitive components parameter estimates become progressively more biased, (b) interval estimates of the cognitive components lose accuracy with incorrect Q matrices, (c) small amounts of misspecification are notable, and (d) sampling error decreased rapidly with increasing sample size. It is expected these results will be confirmed for the CCPE reflected in questions one through three. Questions for item difficulty and person ability are exploratory. Questions four through six are exploratory for cognitive components, beta, and theta

## Study Significance

LLTM can be a bridge that combines psychometric models and cognitive processing operations in the human mind allowing researchers to decompose item difficulty into cognitive components. These cognitive components when properly explicated can provide important information for item development, test design, and standardized assessments in education (Rupp & Templin, 2008). In fact, PARCC intends to include what it describes as *Diagnostic Assessment* in its assessments which will be ready for 2014-15. It is no small matter that these assessments will impact twenty-five million students across twenty-three states.

The implications for mathematics education are important. If the LLTM performs well and is robust to misspecification, then a new set of measurement tools will be available to mathematical educators as they respond to the National Council of Teachers of Mathematics (NCTM) challenge to measure conceptual understanding. Further, present attempts to write items and develop large scale standardized assessments aligned to the Common Core State Standards Initiative (CCSSI) mathematical benchmarks could be greatly enhanced by such modeling.

## Definitions

*A Priori* Specification of a Matrix: Is the qualitative specification of a Weight Matrix commonly known as a Q-matrix, by a content experts group based on theory and domain specific content knowledge.

Item Difficulty ( $\hat{\beta}_{LLTM}; \beta_i = \sum_{j=1}^p \omega_{ij} \alpha_j + c$ ): An independent variable in IRT modeling

representing item difficulty. Item difficulty in the LLTM model is a linear combination of the estimated cognitive components coefficients. Note this is a major difference between item difficulty in IRT and LLTM; in IRT item difficulty is an estimated parameter, whereas in LLTM item difficulty is computed as a linear combination of the cognitive components coefficients parameter estimates. LLTM can be thought of as closely

related to principal component analysis (PCA) and factor analysis in that all these techniques look for linear combinations of variables.

$\alpha_j, j = 1, \dots, p, :$  represents the contribution of the cognitive components of the model to item difficulty

$\omega_{ij} :$  is a 0 or 1 indicator of the association between component  $j$  and  $I_i$ . The  $\omega_{ij}$ 's are assembled in a weight matrix some have called Q.

Cognitive Component Coefficient Estimate (CCCE;  $\alpha_j$ ): Quantitatively estimated parameters capable of bridging cognitive processing models and psychometric models. The cognitive components coefficients are estimated via a marginal maximum likelihood approach that fits the non-linear models by maximizing an approximation to the likelihood over the random effect of person ability. The cognitive components are valid latent psychological units (cognitive operations).

Conditional Maximum Likelihood (CML): An estimation technique employed in RASCH modeling that uses item distribution as a sufficient statistic to compute item difficulty.

This technique is used by Fischer (1973) but not by De Boeck and Wilson (2004).

LCPMWin 1.0: Software developed by Fischer and Forman (1972) used to estimate the LLTM models.

Linear Logistic Test Model (LLTM): A statistical analysis in the general field of CDA introduced by Fischer (1973). The linear logistic test model (LLTM) is capable of bridging cognitive processing models and psychometric models. In the LLTM item difficulty is re-expressed as a summative composite of the cognitive components coefficients parameter estimates associated with the item difficulty.

Marginal Maximum Likelihood: An estimation technique available in SAS PROC NLMIXED (SAS Institute Inc., 2008) used to estimate the LLTM. To estimate cognitive components, Proc

NLMIXED employs a marginal maximum likelihood approach that fits the non-linear models by maximizing an approximation to the likelihood over the random effect of person ability. To estimate person ability Proc NLMIXED employs an empirical Bayes approach which randomly draw from a density defined over the person population.

Q-matrix: A matrix which gathers the weights for cognitive components often referred to in the literature as a weight matrix (Baker, 1993).

Rasch Model: A popular item response theory model used to estimate item difficulty parameters from outcome scores.

Person Ability (Theta;  $\theta_p$ ): Is an independent variable in IRT modeling representing a person's ability level. In Fischer's (1973) original LLTM, person ability could only be estimated once item difficulty was known. In the De Boeck and Wilson's (2004) variant of LLTM, person ability is estimated as a random effect over the person population using an empirical Bayes approach. In SAS (SAS Institute Inc., 2008) an output data set can be generated containing empirical Bayes estimates of these random effects.

## CHAPTER TWO: LITERATURE REVIEW

A number of European psychometricians (e.g., Scheiblechner, Kluwe, Fischer, and Spada) developed and began the process of applying the LLTM to real world applications (Embretson, 1985). The initial work was done in Europe in large part by German researchers. However, the work began to be more readily available in the English-speaking world when Kluwe and Spada (1980) published an edited volume from a conference entitled *Developmental Models of Thinking* which was held at the University of Kiel, Germany. The contributed papers addressed issues related to modeling cognitive development and included an early paper employing the LLTM (Spada & Kluwe, 1980).

Susan Embretson (1985), while at the University of Kansas, edited the volume *Test Design: Developments in Psychology and Psychometrics*. This volume includes contributions from David Andrich, Isaac Bejar, Earl Butterfield, Hartmann Scheiblechner, Richard Snow, Hans Spada, and Robert Sternberg. The goal of this volume was to present recent developments in psychology and psychometrics contributing to the design of psychological tests. The LLTM figured prominently throughout the volume along with other latent trait models such as the multicomponent latent-trait model (MLTM: Whitley, 1980), general component latent-trait model (GLTM: Embretson, 1985), linear exponential model (LEM: Scheiblechner, 1985), and the dispersion location model (DLM: Andrich, 1985).

The LLTM modeling continued to be associated with psychological testing (e.g., Embretson, 1981; Green & Smith, 1987; Embretson & Wetzel, 1987; Sheehan & Mislevy, 1990) as the method spread. As will be demonstrated in this literature review, Susan Embretson (formerly Whitley) can be thought of as a major contributor in the development and evolution of LLTM.

This literature review will examine: (a) the LLTM from its beginnings in Germany to its present usage in education and psychology; (b) educational and psychological uses for the LLTM method breaking this down into mathematical models, geometric and spatial models, paragraph comprehension and reading models, and other approaches; (c) Fischer's contribution; (d) cognitive component coefficients estimation; (e) item difficulty coefficient computation; (f) the LLTM estimation approach taken by De Boeck and Wilson (2004); (g) the Q-matrix simulation literature; (h) Embretson's contribution; (i) Q-matrix simulation literature within cognitive diagnostic assessment family; and (j) MacDonald LLTM and simulation research.

## **Educational Uses of LLTM**

### **Mathematical Models**

LLTM has enjoyed a varied but limited application beginning with Fischer (1973). Table 1 offers a sample of major LLTM applications by domain including a listing of the cognitive components (CCs) which were validated. Fischer (1973) showed that his model was appropriate for students solving items by a combination of cognitive operations, or what he called 'rules'. In particular, he would say, the psychological complexity of differential calculus taught in secondary school mathematics is explainable through seven psychologically meaningful operations. He found and validated (a) differentiation of a polynomial, (b) product rule, (c) quotient rule, (d) compound functions, (e)  $\sin x$ , (f)  $\cos x$ , and (g)  $\exp x$ . His contribution did not lie only in the decomposition of Rasch item difficulty into these cognitive components, but also the demonstration that operations in solving problems are psychological units. In other words, the differentiation of a polynomial is considered to be a single psychological



Table 1

*Educational Uses of LLTM from 1972-2011*

Author	Year	Domain	Cognitive Components
Fischer	1973	Differential Calculus	Differentiation of a polynomial, product rule, quotient rule, compound functions, sin x, cos x, exp x, ln x
Spada, Kluwe	1980	Balance Scale problem solving	Deduction from different amounts of weights, deduction from different lengths, compensation for weight change, compensation for change on same side, compensation for change on opposite side, equilibrium, equality of weights and levers, inequality of modality
Embretson	1981	Geometric analogies	Elements, number, shading, shape, size, reflection, exchanges, rotation
Green, Smith	1987	Visual Attention, Short term memory	Tapping sequence: no. of taps, no. of reversals, and distance covered
Embretson, Wetzel	1987	Paragraph Comprehension	Text Model (modifier Propositional Density, Predicate Propositional Density, Connective Propositional Density, Argument Density, Text Content Word Frequency, Percent Content Words); Decision Model (Percent Relevant Text, Falsification, Confirmation, Word Frequency- Distractors, Word Frequency-Correct, Reasoning-Distractors, Reasoning-Correct)
Sheehan, Mislevy	1990	Document Literacy Tasks	No. of OCs, No. of OCs embedded, Levels of OC Embeddings, No. of SPEs, No. of SPEs embedded, Levels of SPE Embeddings, No. of OCCs, Levels of OC Embeddings, No. of SPEs, Degrees of Correspondence, Type of Information, Plausibility of Distractors

Table 1 (Continued)

*Educational Uses of LLTM from 1972-2011*

Author	Year	Domain	Cognitive Components
Embretson	1992, 2002	Spatial Ability	Distractor type, problem type, spatial problems (linear, quadratic), surfaces (linear, quadratic), position problems (linear, quadratic)
Cisse	1994	Math word Problems	Part-whole, double-role counters, re-representation, comparative terms, action cues, language consistency
Embretson (Multi-dimensional Rasch Model for Learning and Change)	1995	Mathematical Problem Solving	Factual Linguistic (Number of Words, Flesch-Kincaid reading level, conversion of units); Schematic (Number of Equations, Relative definition of Variables , Problem Type); Strategic (Transformation required to isolate unknown)
Tanzer, Gittler, Ellis	1995	Spatial Ability Test	3-turn item, 1 center pattern, 2 center patterns, attractive distracter, extreme position, cumulative practice items 2-7
Dimitrov	1996	Cognitive Operations – Statistics	Concept, rule or principal (CRP) identified from a set of given options, CRP identified from verbal interpretation, CRP inference from implicit contextual information, use of CRP for inference, justification or explanation, simple routine procedure, familiar algorithm problem, familiar non-algorithm problem, unfamiliar “jointing type” problem, unfamiliar “analysis type” problem
Embretson (2-PL Constrained)	1999	Abstract Reasoning	Cognitive (Number of rules, abstract correspondence); Perception (Overlay, Fusion, Distortion)
Embretson & Gorin	2001, 2005	Paragraph Comprehension	Modifier density, predicate density, connective density, argument density, content words frequency,

Table 1 (Continued)

*Educational Uses of LLTM from 1972-2011*

Author	Year	Domain	Cognitive Components
			content words %, relevant text, falsification, confirmation, distractor word, key word frequency, distractor reasoning, key reasoning
De Boeck, Wilson	2004, 2008, 2011	Verbal Aggression	Do vs Want, Others-to-blame, Blaming (curse & scold vs Shout), Expressing (Curse and Shout vs Scold), intercept
Gorin	2005	Reading Comprehension	(Embretson & Wetzel, 1987) (Gorin & Embretson, 2001, 2005)
Dimitrov	2007	Fischer's differential calculus data	Differentiation of a polynomial, product rule, quotient rule, compound functions, $\sin(x)$ , $\cos(x)$ , $\exp(x)$ , $\ln(x)$
Embretson & Daniels	2008	Mathematical Problem Solving	Constant, Discrimination Constant, Encoding, Equation Needed, Translate Equation, Generate New Equation, Visualization, Maximum Knowledge, Equation Recall Count, Subgoals Count, Relative Definition, Procedural Level, Computation Count, Decision Processing
Sonnleitner	2008	Reading Comprehension	Propositional complexity, inference of causality, inference of the emotional reaction of a character, inference of a subordinate noun category, inference of a used instrument, inference of general conditions, degree of coherence, number of response options, number of correct response options, temporal dependency, ambiguity, text
Draney, Wilson	2008	Teacher rating of classroom assessment (BEAR assessment)	Teacher rating severity across variables (Evidence and tradeoffs, designing and conducting investigations, understanding scientific concepts, communicating

Table 1 (Continued)

*Educational Uses of LLTM from 1972-2011*

Author	Year	Domain system)	Cognitive Components
Xie, Wilson	2008	PISA Mathematics Items	scientific information, group interaction) Intercept; Content (space and shape, change and relationship, uncertainty); Process (connection, reflection); Situation (educational/occupational, public, scientific)
Kubinger	2008	Various approaches using LLTM	Item calibration consecutively in time, measuring content specific- learning effects, measuring position effects of item presentation, measuring warming-up effects, measuring effects of speeded item presentation, measuring effects of different item response formats
Holling, Blank, Kuchenbacker , Kuhn	2008	Statistical Word Problems	Context, Number format, or, and, complementary, rearrange, irrelevant, unknown, Grade 11, Grade 12, order, intercept
Hohensinn, Kubinger, Reif, Holocher-Ertl, Khorramdel, Frebort	2008	Large Scale Testing	Item position effects
Hahne	2008	Position effects within reasoning items	Vietnamese matrices
Medina-Diaz	2009	Solving Linear Equations with one variable	Production Rules representing an individual's knowledge: Collecting (P1-P3); Balancing (P4-P5); Removing Parenthesis (P6-P7); Solve for the Variable (P8)
Chen, MacDonald, & Leu	2011	Rational Numbers	Using Illustrations, Providing Interpretations, Applying Judgment, Computation, Checking Distracters, Solving Routine Problems

operation which must be mastered and correctly combined with other operations. He also demonstrated that the complexity of a task is primarily determined by the combination of different operations and is not increased significantly by the repetition of the same cognitive operation within items.

Dimitrov (2007) replicated the rules as identified by Fischer while noting that the cognitive structure of a test is typically defined as a set of cognitive operations and processes required to produce correct answers for items on a test. Knowledge about cognitive structures can help (a) construct test items with tolerable measurement and cognitive characteristics (b) develop teaching strategies that target cognitive and processing abilities, (c) operationalize constructs, and (d) better understand cognitive processes.

Similarly, Spada and Kluwe (1980) produced modeling based on the hypotheses that children of different ages solve individual problems employing cognitive operations measurable via cognitive diagnostic assessment. In particular, this modeling allows specification and testing of cognitive processes employed by students during problem solving. In two studies, students aged nine to sixteen solved balance scale problems. Spada and Kluwe specified *a priori* a cognitive processing model based on the work of Inhelder and Piaget (1958), and their observations of students solving similar problems. They produced two deterministic models of cognitive development and a probabilistic model, which came to be named the linear logistic model of thinking. They indicated it is feasible to test assumptions about cognitive development if the modeling has been built upon the specific solution to the specific items on the assessment.

They validated the following cognitive operations (a) deduction from different amounts of weights, (b) deduction from different lengths, (c) compensation for weight change, (d) compensation for change on same side, (e) compensation for change on opposite side, (f) equilibrium, (g) equality of weights and levers, and (h) inequality of modality.

Embretson and Daniels (2008) suggested that when the LLTM uses items that have complex variables that are theoretically interesting and empirically supported the researcher gains several advantages. These advantages include (a) elaborating construct validity at the item level, (b) defining variables for test design, (c) predicting parameters for new items, (d) banking items by complexity, and (e) offering a solid basis for item design and item generation. They note that LLTM has not been used for large-scale standardized testing. In this study they compare the results from a regression model and the LLTM, as applied to mathematical problem solving from a wide variety of readily available tests. In their work they validated an interesting model for cognitive processing (a) constant, (b) discrimination constant, (c) encoding, (d) equation needed, (e) translate equation, (f) generate new equation, (g) visualization, (h) maximum knowledge, and (i) equation recall count. This model included the sub goals of count, and relative definition and a procedural level comprising computation count, and decision processing.

Chen, MacDonald, and Leu (2011) used LLTM to validate a set of cognitive components that explain cognitive processes employed by Grade 5 and 6 Taiwanese students when they solve a twenty-three item assessment. The participants were 2,612 students drawn from across 42 schools in Taiwan. Six cognitive components were identified for the fraction conceptual items including (a) using illustrations, (b) providing interpretations, (c) applying judgment, (d) computation, (e) checking distractors, and (f) solving routine problems. This model has similarities to the cognitive processing models proposed by Spada and Kluwe and Fischer with the exception it does not try to exhaustively explicate all processes required by a student who solves the fractional items. It validates six cognitive processes employed by the students without suggesting the model is complete.

Medina-Diaz (2009) defined the cognitive structure of an algebra test using LLTM but she also employed another method as a comparison known as quadratic assignment (QA). A

full explanation of this model is found in her paper. She indicated the cognitive structure of a test is specified using the Q-matrix. The cognitive structure she defined is based on a set of eight production rules that represented the mathematical procedures employed in solving linear equations with one variable. To accomplish these goals, she constructed a 29-item test and administered it to 235 ninth-graders. She used Fischer and Forman's (1972) LCPMWin 1.0 computer program. The structure provided a cognitive processing model with four rules for solving Linear Equations with one variable (a) collecting, (b) balancing, (c) removing parenthesis, and (d) solving for the variable.

Embretson (1995) linked student learning to changes in cognitive processes and knowledge structures. She employed a multidimensional Rasch model for learning and change. The model is of interest because she employed a linear combination of what she calls 'complexity factors'. The linear combination is the same construct under consideration in the LLTM. The knowledge structures and cognitive processes are involved in initial learner goal states and throughout the stages of learning. The initial learner states need to be assessed for the mastery of cognitive processing capabilities and acquired knowledge. However, the acquisition and transition from initial learner to master also requires iterative measurement which is directly linked to targeted instructional delivery. Teachers who are quantitatively informed about the cognitive processing or knowledge states of students can target the development of that psychological unit at a course and grade appropriate level.

Embretson's knowledge types were defined as (a) factual linguistic (number of words, Flesch-Kincaid reading level, conversion of units), (b) schematic (number of equations, relative definition of variables, problem type), and (c) strategic (transformation required to isolate unknown). Embretson concluded that this model can be applied to link competency changes to substantive aspects of mathematical problem solving.

Dimitrov (2007) followed up his doctoral dissertation work by offering a cognitive operations model for statistics. Rather than adopt a rule processing cognitive components model he validated a more conceptual model comprised of (a) concept, (b) rule or principal (CRP) identified from a set of given options, (c) CRP identified from verbal interpretation, (d) CRP inference from implicit contextual information, (e) use of CRP for inference, (f) justification or explanation, (g) simple routine procedure, (h) familiar algorithm problem, (i) familiar non-algorithm problem, (j) unfamiliar jointing type problem, and (k) unfamiliar analysis type problem.

Xie and Wilson (2008) studied differential item functioning (DIF) using the LLTM along with three other approaches. They examined mathematical items from the *Program for International Student Assessment (PISA)* database (2003). Their analysis was conducted using NLMIXED in SAS (SAS Institute Inc., 2008). Their study focused on item fairness for underlying dimensions such as group membership, gender, and ethnicity. They pulled sample sizes ranging from 481 to 534 students from Canada, Taiwan, the United States and Japan. The items they chose ranged from 19 to 40 per domain. They validated a content and process model parameterized as (a) content (space and shape, change and relationship, quantity, and uncertainty), (b) process (reproduction, connections, and reflection), and (c) situation (personal, educational or occupational, public and scientific).

They created the Q-matrix according to the domains and looked at the results of the parameter estimates comparatively across countries to determine if there was differential facet functioning. They concluded the response data conformed to the structure of the test design and that SAS easily handled the necessary analysis.

In this section mathematical models validated using LLTM were explored. These include (a) seven rules for solving differential calculus (Fischer, 1973; Dimitrov, 2007), (b) cognitive operations required to solve balance scale problems (Spada & Kluwe, 1980), (c) a mathematics cognitive processing model (Embretson & Daniels, 2008), (d) cognitive processing model for



fractions (Chen, MacDonald, & Leu, 2011), (e) rules for solving linear equations (Medina-Diaz, 2009), (f) a model for assessing mastery of cognitive processing capabilities and acquired knowledge (Embretson, 1995), (g) a cognitive operations model for statistics (Dimitrov, 2007), and (h) a content and process model for differential item functioning (Xie & Wilson, 2008).

In sum these studies showed that (a) LLTM is capable of building reliable and valid mathematical models to assess cognitive components, (b) CCs allow a better understanding of the cognitive processes used by examinees, (c) CCs allow teachers to target development of cognitive processes at a course and grade level, (d) LLTM is useful for item banking, (e) LLTM is useful when elaborating construct validity at the item level, (f) LLTM is useful for the design and generation of test items, (g) LLTM models can be directly related to rule production or cognitive processes (e.g., the fractions model built by Chen, MacDonald, & Leu, 2011) or more abstract models (e.g., mastery of cognitive processing capabilities and acquired knowledge, Embretson, 1995), and (h) SAS easily handles the LLTM statistical analysis.

### **Geometric and Spatial Models**

Whitely and Schneider (1981) used LLTM to describe the processing contributions to item difficulty. They administered 30 geometric analogies to 211 undergraduate students. The response time was fixed at 25 seconds to control for the test position effect and to allow the researchers to have more certainty that item differences would depend on item complexity. They validated eight cognitive operations (a) elements, (b) number, (c) shading, (d) shape, (e) size, (f) reflection, (g) exchanges, and (h) rotation. The authors also looked at the complexity of the parameters by gender analyzing three different models. Interestingly, they did not find gender differences in the information structure parameters which indicate there is no DIF in the items when considering male and female participants. They concluded the data indicated that two types of transformations did not involve different abilities, yet test developers can design items for difficulty by controlling their information structure.

Embretson (1992) studied cognitive modifiability from the responsiveness of an individual's performance to intervention using the Spatial Learning Ability Test. Of interest in this study was her use of LLTM which she applied to the MRMLC results. Using a program (LINLOG, 1998) developed by Whitely and Neich, LLTM parameters were estimated. Embretson examined the difficulty of item responses according to distractor type, problem type, spatial problems (degrees – linear, quadratic), surfaces (linear, and quadratic), and position problems (degrees—linear and quadratic). She examined the similarity of the LLTM models across occasions by coding product interactions for each of the cognitive components listed above. Embretson primarily studied cognitive modifiability of spatial ability over time, and her paper is of interest because it demonstrated the wide applicability of LLTM in varying types of applications.

Tanzer, Gittler, and Ellis (1995) examined spatial ability using the Three-Dimensional Cube Comparison Test. This test was administered to 384 students in the United States and 307 students in Austria. It is worth noting, Xie and Wilson (2008) followed a similar pattern. When they analyzed their results they validated a six operation cognitive processing model (a) 3-turn item, (b) 1 center pattern, (c) 2 center patterns, (d) attractive distracter, (e) extreme position, and (f) cumulative practice items.

They discovered that the students in the United States worked more quickly than students in Austria but had lower scores. The authors suggest motivation may have caused the test taking difference. After controlling for motivation, the responses were analyzed using the LLTM and no cross-cultural differences were found. They conclude that the performance of the LLTM demonstrates it is a useful tool for identifying the components that make up item complexity in a domain. They further point out that LLTM is a useful cross cultural tool to identify item bias. In other words, LLTM can be employed to identify the latent constructs which differential item functioning represents.

In this section geometric and spatial models were validated using LLTM. These include (a) eight geometric and spatial operations and used the model to examine the impact items might have when in various positions on a test (Whitley & Schneider, 1981), (b) a model for cognitive modifiability using the Spatial Learning Ability Test (Embretson, 1992), and (c) validated a six operation spatial processing model capable of cross-cultural assessment (Tanzer, Gitler, & Ellis, 1996).

In addition to many of the conclusions already discussed, these studies demonstrate that (a) LLTM modeling can be used to validate geometric and spatial models of cognitive processing, (b) LLTM is useful when identifying item complexity associated with cognitive processes, and (c) LLTM is useful when identifying cross cultural item bias.

### **Paragraph Comprehension and Reading Models**

Close to the end of the 1980s Embretson began researching reading comprehension cognitive processing models. Embretson and Wetzel (1987) proposed a text model and decision model for paragraph comprehension. Their two processing stages include a text model (modifier, propositional density, predicate propositional density, connective propositional density, argument density, text content word frequency, and percent content words), and a decision model (percent relevant text, falsification, confirmation, word frequency- distractors, word frequency-correct, reasoning-distractors, and reasoning-correct). A useful outcome of the study was an explication of the cognitive design principles in the item bank. They suggested LLTM can be useful when selecting items while at the same time controlling the source of item complexity. In other words, the weights for the cognitive model can be useful to estimate the complexity of items cognitive components. The results indicate that (a) successful prediction of item difficulty is obtained from models with wide representation of both text and decision processing, (b) items can be screened for processing difficulty prior to being administered to

examinees, (c) the two processing stages involve two different ability dimensions, (d) and employing the LLTM to validate the model was a wise choice.

The National Assessment of Educational Progress (NAEP) conducted a review from 1984 to 1986 of the literacy skills of America's young adults ages 21 to 25. The assessment required the use of literacy tasks that simulated the diverse literacy demands of adult interactions in occupational, social, and educational settings. It included 63 items requiring ability in acquiring and using information from written documents. Sheehan and Mislavy (1990) at Educational Testing Services employed LLTM to analyze the data. They classified the variables into (a) materials variables, defined as the length and organizational complexity (OC) of the document in a task; (b) directive variables, defined as the length and organizational complexity of the task; and (c) process variables, defined as the difficulty of the task solution process. They employed LLTM to validate the following cognitive structure; number of organizing categories (OC), number of OCs embedded, levels of OC embedding's, Number of specific categories (SPE), number of SPEs embedded, levels of SPE embedding's, number of OCCs, levels of OC embedding's, number of SPEs, degrees of correspondence, type of information, and plausibility of distractors.

They concluded their article signaling that simply having a highly reliable assessment does not mean that you have a valid assessment. They point out the strong link found between IRT difficulty and document literacy is promising cognitively and psychometrically, and that using LLTM was an important step in that direction.

Cisse (1994) wrote his dissertation at the University of Alberta taking a psychometric approach to modeling the importance of six cognitive components' for arithmetic word problems. Forty students from Grades 1 through 3 were asked to solve 32 addition and subtraction word problems. Cisse posited that knowledge available to students determines outcomes; however, the age at which knowledge structures develop remains unknown. Fitting the LLTM to the data,

Cisse validated (a) part-whole, (b) double-role counters, (c) re-representation, (d) comparative terms, (e) action cues, and (f) language consistency. While this is an interesting study, MacDonald and Kromrey (2011) suggested LLTM parameters do not settle down until the sample size reaches somewhere between 600 and 1,000 participants. Given this study only employs about 40 students no evaluation of the LLTM will be made.

Embretson (Embretson & Gorin, 2001) worked with one of her graduate students, Joanna Gorin, in what is now being called Diagnostic Assessment. Embretson and Gorin, as a piece of their exploration of construct validity for cognitive psychology, conducted a LLTM analysis. They validated a cognitive component model for paragraph comprehension (a) for the text model (modifier density, predicate density, connective density, argument density, content words frequency, and content words percent), (b) for the decision model (relevant text, falsification, confirmation, and distractor word), and (c) for the frequency model (key word frequency, distractor reasoning, and key reasoning).

They concluded that, while there is great potential for cognitive psychology to improve construct validity, actual application has lagged. Not employing construct meaning when defining constructs (e.g., selecting item types, diagnosing sources of performance, and developing and evaluating scoring systems) represents a gap which should be closed.

Sonnleitner (2008), like Embretson, was interested in item-generation systems. They used the LEVE-E (Leseverständnistest für Erwachsene [Reading Comprehension Test for Adults]) which is a computer-based, multiple-choice test, consisting of two texts of equal length on two different topics. They tested the LEVE-E on 301 university students of different academic disciplines as well as clients taking a psychological driver's examination. The sample size is small and suggests their parameter estimates will be biased (MacDonald & Kromrey, 2011); however, the sample is large enough to warrant comment.

They offered a 12 parameter model and validated 11 of the 12 parameters. Their 12 component processing model for reading comprehension is (a) Propositional complexity, (b) inference of causality, (c) inference of the emotional reaction of a character, (d) inference of a subordinate noun category, (e) inference of a used instrument, (f) inference of general conditions, (g) degree of coherence, (h) number of response options, (i) number of correct response options, (j) temporal dependency, (k) ambiguity, and (l) text. Further, the researchers compared Rasch item difficulty to the linearly recomposed LLTM item difficulty, which is a return to the graphical model check first proposed by Fischer (1973). Sonnleitner notes this comparison can offer valuable information useful when modifying or reformulating cognitive components. Their conclusions of interest suggest LLTM successfully explained 21 item parameters in its 12 elementary operations. While they feel this is a good result, they also call for more research to replicate and validate these findings.

Mispelkamp (1985) was the first to use the LLTM to test an item-generating system for reading comprehension but this is an unpublished dissertation available only in German and is, therefore, excluded from consideration.

Holling, Blank, Kuchenbacker, and Kuhn (2008) examined cognitive processing models for statistical word problems. They report results from a pilot study with systematically-designed statistical word problems using the LLTM. Their sample size (192 students from five different German grammar schools) of study participants was small and the parameter estimate will likely be biased. However, the authors were clear that this research is a pilot study and exploratory in nature. They attempted to validate a statistical word model comprising (a) context, (b) number format/complementary, (c) rearrange, (d) irrelevant, (e) unknown, (f) Grade 11, (g) Grade 12, and (h) order. They found that the cognitive model fit their data. They suggest that statistical word problems, similar to recommendations from Embretson for paragraph comprehension, can

be designed and analyzed in a systematic way, and that their cognitive model for solving statistical word problems could be used in future assessments.

In this section LLTM was used to validate paragraph comprehension and reading models. These include (a) a processing model for text and decision making (Embretson & Wetzel, 1987), (b) a model to assess adult's ability to acquire and use information from written documents (Sheehan & Mislevy, 1990), (c) a thirteen cognitive component model for paragraph comprehension (Embretson & Gorin, 2001 ), (d) an eleven cognitive component model for reading comprehension (Sonnleitner, 2008), and (e) a model for statistical word problems (Holling, Blank, Kutchenbacker, & Kuhn, 2008).

In addition to the conclusions found in the first two sections, additional conclusions include: (a) the link between IRT item difficulty and document literacy is promising and that LLTM modeling is an important step in building reliable and valid models; (b) that paragraph reading and comprehension models can be built using the LLTM; and (c) LLTM offers the potential to improve construct validity; however, actual application has lagged. Not employing construct meaning when defining constructs (e.g., selecting item types, diagnosing sources of performance, and developing and evaluating scoring systems) represents a gap which should be closed.

### **Other Approaches**

Not all of the uses of LLTM fall easily within the groupings already suggested. Kubinger (2008) made a spirited defense of LLTM; Hahne (2008) found there was no evidence for position effects within Viennese Matrices using LLTM; and Hohensinn, Kubinger, Reif, Holocher-Ertl, Khorramdel, and Frebort (2008) used LLTM to validate item position effects for large scale testing. Item position effects are defined as several test booklets with the same items presented at different test positions. If item effects are established, it would mean that the

estimated item parameters do not depend exclusively on the items' difficulties due to content but also on their presentation positions. As a consequence, item calibration could be biased.

Kubinger, Hohensinn, Holocher-Ertl, and Heuberger (2011) employed LLTM for modeling children's cognitive age-acceleration function. Embretson (1982) offered an abstract reasoning model parsing cognition and perception parameterized as cognitive (number of rules, abstract correspondence), and perception (overlay, fusion, distortion).

In sum (a) LLTM can be used to validate mathematical, geometric and spatial, and paragraph comprehension and reading models; (b) LLTM is useful when identifying item complexity associated with cognitive processes; (c) LLTM is useful when identifying cross cultural item bias; (d) LLTM offers the potential to improve construct validity; (e) LLTM permits a better understanding of the cognitive processes used by examinees; (c) LLTM allows teachers to target development of cognitive processes at a course and grade level (d) LLTM is useful for item banking; (e) LLTM is useful when elaborating construct validity at the item level; (d) LLTM is useful for the design and generation of test items (g) LLTM models can be directly related to rule production or more abstract models; (h) SAS easily handles the LLTM statistical analysis; and (i) LLTM can be useful in diagnosing the position effect.

### **Fischer's Contribution**

Fischer employed a dichotomous Rasch (1960) model known as the LLTM. The Rasch model, a well-known IRT model, can be represented as:

$$P(\theta) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)} \quad (2.1)$$

where  $P(\theta)$  is the probability of an examinee responding correctly to an item;  
 $\theta_p$  is the ability of the examinee; and  
 $\beta_i$  is the difficulty of the item.



The Rasch model is most often presented as a 1-parameter logistic model because it computes item difficulty using Conditional Maximum Likelihood (CML). In the Rasch model the total score is a sufficient statistic for estimating item difficulty (Embretson & Reise, 2000).

Item difficulty was of interest to Fischer because he believed that re-parameterized item difficulty was capable of bridging cognitive processing models and psychometric models. It is helpful to follow the re-parameterized item difficulty to provide the reader a working knowledge of the model.

Fischer's explication of the LLTM is an extension of the Rasch Model where the item difficulty ( $\beta_i, i = 1, \dots, k$ ) is linearly decomposed as follows:

$$\beta_i = \sum_{j=1}^p \omega_{ij} \alpha_j + c \quad (2.2)$$

where  $\alpha_j$  is the parameter estimate for the cognitive component  $j$ ;

$\omega_{ij}$  is the given weight of  $\alpha_j$ , with respect to the difficulty of item  $I_i$ , and

$c$  is an arbitrary normalization constant.

Item difficulty is not estimated directly from the item-response matrix using standard IRT methods. Rather, the LLTM approach estimates cognitive components coefficients ( $\alpha_j$ ) from the item response matrix based on the input of the Q-matrix. The coefficients of the cognitive components are summed to create item difficulty estimates ( $\sum_{j=1}^p \omega_{ij} \alpha_j$ 's).

### Normalization Constant

Fischer (1973) specified LLTM where  $C$  is a normalization constant which is given by:

$$C = \frac{-\sum_i^n \sum_{j=1}^p \omega_{ij} \alpha_j}{n} \quad (2.3)$$

For the purposes of a simulation study the constant is not needed and will be omitted.

## Cognitive Component Coefficient Estimation

It is the cognitive component coefficients that are estimated in LLTM modeling. As an example, 14 math items were extracted from the released items from the Trends in International Mathematics and Science Study (TIMSS) 2007 User Guide, for Grade 4 Mathematics. The items were taken from the Math section for Grade 4, Blocks M01 and M02 in the Number domain. Blocks M01 and M02 were administered together in Booklet 1.

In Rasch, the Item-Response Matrix would be used to estimate item difficulty for each of the fourteen items. Fischer's contribution was to multiplicatively join the cognitive components to the Item-Response Matrix. Table 2 presents two of three cognitive components. The full information from the Item-Response matrix is used when estimating item difficulty for the Rasch model. In LLTM the Q-matrix restricts the amount of information available in the Item-Response matrix when the cognitive components coefficients are estimated.

For the purposes of this example, two of three identified cognitive components will be used. The first cognitive component found in these TIMMS items was transformation defined as the examinee must use transformation to solve this item (e.g.,  $a+b = ( )$  becomes  $a+ ( ) = c$ ; or  $a-b = ( )$  becomes  $a- ( ) = c$ ). The second cognitive component is division defined as the examinee must use division to solve the item.

From Table 2 notice transformation (CC1) has 1s for items 2, 3, 4, 6, and 9 and has 0s for item 1, 5, 7,8,10, 11, 12, 13, and 14. Students must use transformation to solve items 2,3,4,6 and 9. When the cognitive component coefficient for transformation is estimated, information will only be available in the Item-Response matrix from items 2, 3, 4, 6, and 9. In this way the Q-matrix constrains the estimate of the cognitive component coefficient to items that contain student outcome scores for transformation in the Item-Response matrix.

Table 2

*Q-matrix for Cognitive Components Transformation (CC1) and Division (CC2)*

Item	CC1	CC2
1	0	1
2	1	0
3	1	0
4	1	0
5	0	0
6	1	0
7	0	0
8	0	0
9	1	0
10	0	0
11	0	0
12	0	1
13	0	0
14	0	0

From Table 2 notice division (CC2) has 1s for items 1 and 12 and 0s for all of the remaining 14 items. Items 1 and 12 require the student to use division to solve the fraction item and the other twelve items do not require the student to use division. When the cognitive component coefficient for division is estimated information will only be available in the Item-Response matrix from items 1 and 12. In this way the Q-matrix constrains the estimate of the cognitive component coefficient to items that contain student outcome scores for division in the Item-Response matrix.

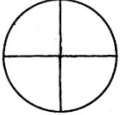
Notice, when comparing transformation (CC1) and division (CC2) that transformation is built from five items and division is built from two items. Transformation can be thought of as denser than division.

### Item Difficulty Computation

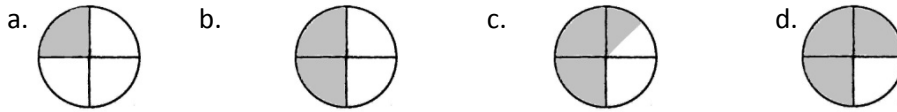
To obtain item difficulty coefficients  $\left( \beta_i = \sum_{j=1}^p \omega_{ij} \alpha_j \right)$ , first the cognitive component coefficients are estimated as indicated in the section above. Then the estimated cognitive components coefficients ( $\alpha_j$ ) are summed to produce item difficulty coefficients. Note that Fischer did not add an error term to his equation. Some researchers have added an error term to the LLTM model (De Boeck & Wilson, 2004).

For an operationalized exemplar, we look at two questions (Chen, MacDonald, & Leu, 2011) in which the authors validated a cognitive processing model for solving fractions. Grades 5 and 6 students in Taiwan completed a 23 item test. Item 5 is chosen as an exemplar of an easy item to solve; it requires the student to use illustrations, provide an interpretation of the item, apply judgment, and compute the fraction. Item 5 is presented in Figure 1.

In the equation below using illustrations, computation, and solving routine problems, enter the equation, but providing interpretations, applying judgment, and checking distractors do not. Summing the cognitive components coefficients produces an item difficulty for Item 5 of -2.36. An examination of the original formula indicates that the two easiest cognitive components (i.e., computation, and solving routine problems) are included. Therefore, it is not surprising the item is an easy item with an item difficulty more than two standard deviations below zero in the logit unit for the Taiwanese students who took this test.

5. Grandmother made a cheesecake and cut it into slices as  shown in the right figure. Stephen took  $\frac{5}{8}$  of the cheesecake. Which of the following

gray areas represent the amount of the cheesecake taken by Stephen?



- e. If none of the above is correct, paint in gray the correct amount of the cheesecake taken by Stephen below.

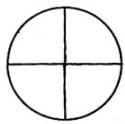
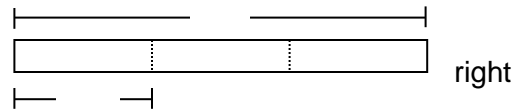


Figure 1. Item 5 from 23 Item Taiwanese fraction test.

$$\begin{aligned} \beta_5' &= \omega_{ij} \alpha_{using-illustrations} + \omega_{ij} \alpha_{providing-interpretations} + \omega_{ij} \alpha_{applying-judgement} + \omega_{ij} \alpha_{computation} \\ &+ \omega_{ij} \alpha_{checking-distractors} + \omega_{ij} \alpha_{solving-routine-problems} \\ \beta_5' &= (1 * .07) + (0 * 1.11) + (0 * 1.81) + (1 * -.67) + (0 * .57) + (1 * -1.76) \\ \beta_5' &= .07 + 0 + 0 - .67 + 0 - 1.76 \\ \beta_5' &= -2.36 \end{aligned}$$

Item 13 is chosen as an exemplar of a difficult item to solve; it requires the student to use illustrations, provide an interpretation of the item, apply judgment, and compute the fraction. It may be the case that a student who is procedurally fluent in computing fractions may be able to solve the item without reference to the illustration. Item 13 appears in figure 2.

13. A ribbon of 2 meters long was folded into 3 segments of equal lengths as shown in the figure. How long is each segment in meters?



- a.  $\frac{1}{3}$  m b.  $\frac{2}{3}$  m c.  $\frac{3}{2}$  m d. 0.66m  
 e. If none of the above is correct, give your answer:

Figure 2. Item 13 from a 23 Item Taiwanese Fraction Test.

This item requires the Taiwanese student to use the illustration, provide an interpretation, apply judgment, and compute the item. The item does not require the student to check the distractor or solve routine parts of the item in order to answer the question correctly.

Item difficulty for item 13 is computed below:

$$\beta'_{13} = \omega_{ij}\alpha_{using-illustrations} + \omega_{ij}\alpha_{providing-interpretations} + \omega_{ij}\alpha_{applying-judgement}$$

$$+ \omega_{ij}\alpha_{computation} + \omega_{ij}\alpha_{checking-distractors} + \omega_{ij}\alpha_{solving-routine-problems}$$

$$\beta'_{13} = (1 * .07) + (1 * 1.11) + (1 * 1.81) + (1 * -.67) + (0 * .57) + (0 * -1.76)$$

$$\beta'_{13} = .07 + 1.11 + 1.81 - .67 + 0 + 0$$

$$\beta'_{13} = 2.32$$

An examination of the original formula indicates that the two hardest cognitive components (i.e., providing-interpretations, applying-judgment) are included. Therefore, it is not surprising that item 13 has an estimated cognitive component coefficient value of 2.32, or is more than two standard deviations above zero in the logit unit for the Taiwanese students who took the exam.

For another operationalized exemplar of the LLTM see Embretson (Embretson & Reise, 2000, p. 280).

## **Person Ability**

In LPCM-Win, person ability coefficients are estimated after item difficulty coefficients have been estimated and are dependent on item difficulty coefficient estimates. In SAS (SAS Institute Inc., 2008), the empirical Bayes method of estimating person ability is random and computed independently of item difficulty as a random draw from a density defined over the population of persons. (De Boeck & Wilson, 2004)

## **Estimation Approaches**

De Boeck and Wilson (2004) have done extensive work with LLTM. In addition, they provide a website (<http://bear.soe.berkeley.edu/EIRM>) which contains data sets, command files for analyses, sample output, and sample answers to exercises. Their model uses sample data for a verbal aggression cognitive processing model. Their parameterized models included (a) do vs. want, (b) others-to-blame, (c) blaming (curse & scold vs. shout), and (d) expressing (curse and shout vs. scold).

The estimation of the cognitive components in LPCM-WIN 1.0 was accomplished using conditional maximum likelihood. This approach used (a) a summation algorithm, (b) an improved version of the difference algorithm, (c) the Quasi-Newton technique, and (d) the Armijo step-length rule. The interested reader is referred to Fischer (1973) for a detailed discussion of these algorithms. Once the cognitive component coefficients have been estimated and item difficulty values have been computed, then person ability parameters estimates can be maximized by means of the Newton method.

The variant of LLTM employed by De Boeck and Wilson (2004) in SAS 9.3 has an important difference when estimating the coefficients of the cognitive components. Recall that conditional maximum likelihood estimation does not involve the person parameter (Fischer, 1973) when computing cognitive components coefficients because the total score is a sufficient statistic for estimating trait levels (Embretson & Reise, 2000). Therefore, in LPCM-Win the

cognitive components coefficients are estimated without including the variance from person ability. In SAS (SAS Institute Inc., 2008) the cognitive components coefficients are estimated by maximizing an approximation to the likelihood over the random effect of person ability. This method of estimating cognitive components coefficients includes the variance from person ability in the estimation of cognitive components coefficients. De Boeck and Wilson (2004) point out that making inferences about cognitive component coefficients independent of the person ability distribution is undesirable because not all the information available in the person parameters is part of the cognitive components coefficients likelihood estimation.

Given that SAS (SAS Institute Inc., 2008) via PROC NLMIXED computes cognitive components coefficients as an approximation to the likelihood over the random effect of person ability, the De Boeck and Wilson (2004) variant of LLTM cannot be thought of as a 1 parameter solution. In SAS (SAS Institute Inc., 2008) cognitive components coefficients are not sufficient statistics because person ability variance is included in the estimation of the cognitive component coefficients. The interested reader is directed to De Boeck's and Wilson's 12<sup>th</sup> chapter, The SAS 9.2 Users Guide for NLMIXED (SAS Institute Inc., 2008), Fischer's LpcM-Win 1.0 Users Guide, and Fischer's publications including his 1995b article.

The syntax provided by De Boeck and Wilson (2004) has important omissions. Their LLTM code recovers the coefficient estimates of the cognitive components ( $\alpha_j$ ) but the code does not recover item difficulty coefficients, and the coefficients for person ability ( $\theta_p$ ). The SAS code, found in appendix A, in this simulation study will recover the coefficient estimates for the cognitive components, the coefficients for item difficulty, and the coefficients for person ability.



### Model of Interest

This study will examine De Boeck and Wilson's variant of the LLTM which will be estimated using SAS (SAS Institute Inc., 2008). SAS PROC NLMIXED employs marginal maximum likelihood and offers an interesting array of estimation techniques.

$$P(\theta) = \frac{\exp(\theta_p - \sum \omega_{ij} \alpha_j)}{1 + \exp(\theta_p - \sum \omega_{ij} \alpha_j)} \quad (2.4)$$

$\theta_p$  is the ability of the examinee;

$\omega_{ij}$  is the given weight of  $\alpha_j$ , with respect to the difficulty of item  $I_i$ ; and

$\alpha_j$  is the parameter estimate for the cognitive component  $j$ .

### SAS 9.3 and Proc NLMIXED

Fischer introduced LpcM-Win 1.0, a software program capable of producing parameter estimates for the LLTM. Given the problems noted earlier (i.e., Model of Interest and Statistical Program) this study will employ SAS (SAS Institute Inc., 2008) to implement the LLTM model based on the work of De Boeck and Wilson (2004)

### General Comments

The equations in the LLTM models are anchored on the item distribution with person ability free to vary. Person ability is normally distributed,  $\theta_p \sim N(0, \sigma_\theta^2)$ , and is estimated using an empirical Bayes algorithm as a random draw over the person ability distribution. The parameter estimates for the cognitive components coefficients and item difficulty coefficients are fixed effects. Person ability estimates are randomly estimated with zero in the logit unit and a standard deviation of 1.

## Q-matrix Simulation Literature

A critical step when estimating the LLTM model is specification of the cognitive components which are to be measured by items on the cognitive diagnostic assessment. A weight is assigned to each item,  $i$ , for each cognitive component,  $k$ . These weights are specified *a priori* by content experts. There are a number of steps recommended when specifying a Q-matrix (a) convene a panel of content knowledge with expertise in item specification, item development, content teaching, and domain expertise; (b) task the content expert group determining which cognitive components are required of examinees who respond to the items on the assessment; (c) conduct cognitive *think-aloud's* with a representative sample of examinees; (d) conduct a qualitative analysis of the *think-aloud's* to determine if the cognitive components the content experts group has defined are consistent with student thinking; (e) task the content knowledge group to determine if each cognitive component is required for each item on the assessment; and (f) re-convene the content knowledge group to discuss variances in their Q matrices with the goal of arriving at a mutually agreed upon Q-matrix. The importance of developing competing Q matrices has been noticed in the literature (Rupp, 2012)

Simulation research focusing on the Q-matrix is sparse. In 1987, Green and Smith did a simulation study concerning the Q-matrix. The first true study of the sensitivity of the Q-matrix was conducted by Baker (1993). Cassuto focused his dissertation work on the accuracy and practicality of the Q-matrix. He examined how the internal structure of the Q-matrix affected beta and theta estimation when test length and sample size varied. Further, he examined the effects on the estimation of theta and beta when the distribution of theta is skewed (Cassuto, 1996). A number of simulation studies (Chiu, 2013; de la Torre, 2008; Rupp & Templin, 2008; Liu, Xu, & Ying, 2012) have focused on validating the Q-matrix and the effects of misspecification on parameter estimates in the cognitive diagnostic model deterministic, inputs, noisy and gate (DINA). DINA can be thought of as discrete latent variable model that allows

inferences about the cognitive information of the items and cognitive attributes of examinees. Attribute misspecification was simulated for the Rule Space method (Im & Corter, 2011). A simulation study was conducted on the impact of model misspecification on parameter estimation (Kurnin-Habenicht, Rupp, & Wilhelm, 2012) for log-linear diagnostic classification models.

About twenty years after Fischer first introduced LLTM modeling, Green and Smith (1987) engaged interesting research to understand the mental processes or components which contribute to the difficulty of a task. In particular, they used *the Knox Cube Block Test* to measure visual attention and short term memory. They parameterized their results as (a) tapping sequence, (b) number of taps, (c) number of reversals, and (d) distance covered. They used simulated and real data. For one simulation, data samples of 30, 200, and 1,000 were used to respond to the 18 item test. The authors correlated the cognitive component coefficients and found they were highly correlated ranging from .89 to .97 within a normal distribution of -2.5 to +2.5 logits for cognitive components with estimated starting value of .2, .3 and .1. In a second simulation, the sample sizes remained constant but the original estimates for the cognitive components were set at .2, .4, and .8. When the resulting cognitive components coefficients correlations were computed, they were found to range between -.21 to -.25. A third simulation increased the cognitive components coefficients for item difficulty by one logit in an effort to simulate students having difficulty adjusting to the material. They used Fischer's (1973) original program to obtain parameter estimates but, because the standard error is not offered in this software, they also used Embretson's LINLOG.

As far as LLTM is concerned, the authors concluded (a) the model works well and is stable, (b) LLTM is sensitive to the presence of correlated coefficients measurement disturbance in misspecified component models, and (c) they recommended a regression approach over LLTM. Embretson would later answer this criticism by suggesting that LLTM

offered better control of standard error and was, therefore, preferable to the regression approach (Embretson, 2000).

Baker was the first to conduct simulation work looking at the sensitivity of the misspecification of the Q-matrix. Baker conducted his work with sparse Q matrices, dense Q matrices, eight cognitive components, and 21 items. The sparse matrix was 20% (34/168) filled with 1s and the dense matrix was 57% (96/168) filled with 1s. The coefficients of the cognitive components ranged between -.75 and 1.2 and were considered to be the truth and error free. He varied the sample sizes from 20, 50, 100, and 1,000. To conduct his analysis, 0s were transformed into 1s and vice versa at the following percentages: 1%, 2%, 3%, 5%, 7.5% and 10%.

Baker concluded that a small degree of misspecification in the Q-matrix had a large impact on the parameter estimates. When the misspecification was between 1%-3%, the sparse Q-matrix yielded a larger root mean square (RMS) average than did the dense Q-matrix. This was also true for higher levels of misspecification between 5%-10%. Therefore, the density of the Q-matrix is an important factor affecting the estimates of cognitive components. He posits a dense Q-matrix with a low level of misspecification and a large number of items the effects of misspecification are masked. Therefore, for dense Q matrices, the effects of misspecification are not quite as serious. Sample size had a minimal effect on root mean square error. Therefore, to obtain proper estimates of the cognitive components there must be a minimum number of examinees responding correctly to the items for each cognitive component and the cognitive components must appear in a sufficient number of items. He does not suggest what those levels might be.

Baker found, once the estimation of cognitive components becomes stable, further increases in sample size have little additional effect on parameter estimates. He suggested further research needed to be conducted on sparse Q matrices especially when the sample size

or number of respondents is small. He points out that due to the linear addition of cognitive components item difficulty in LLTM should have more error compared with item difficulty in Rasch modeling.

Cassuto (1996) studied the performance of LLTM under various testing conditions. His doctoral dissertation addressed issues pertinent to the accuracy and practicality of the LLTM and its Q-matrix. In particular, he researched the effect the internal structure of the Q-matrix has on the estimation of cognitive components, item difficulty, and person ability. He simulated (a) different test lengths, (b) sample size, and (c) negative skew of the person ability distribution. Cassuto examined four different types of Q matrices (a) orthogonal-dense, (b) correlated-dense, (c) orthogonal-sparse, and (d) correlated-sparse.

He found three influential factors in the recovery of person ability (a) test length, (b) sample size, and (c) Q-matrix structure. Specifically longer tests with larger sample sizes and orthogonal-sparse Q matrices are preferred. Cassuto discussed the influential factors associated with item difficulty. He found recovery of item difficulty coefficients was very good across all conditions, which is a finding replicated by recent simulation studies (MacDonald & Kromrey, 2011, 2012). Further he found that Q-matrix density, correlation, and skew did not result in attenuated item difficulty parameter estimates.

Cassuto's conclusions about cognitive components coefficients are: (a) the recovery of item difficulty was generally good across all conditions but not as good as the recovery of cognitive components; (b) correlated sparse Q-matrix did not recover cognitive components as well when there were 50 or fewer examinees. He suggests there may not be enough examinees in the sample size to stabilize the estimation of the cognitive components in the correlated sparse Q-matrix and that the sharing of variance across a correlated Q-matrix causes attenuation in the recovery of cognitive components coefficients. He suggests that estimation of cognitive component coefficients should settle down somewhere between 50 to 250 examinees.

Recent research (MacDonald & Kromrey, 2011, 2012) suggests that cognitive components coefficients estimation recovery truly settles down around 600 to 1,000 examinees; (c) an orthogonal Q-matrix is optimal compared to a correlated Q-matrix because each cognitive component will account for different portions of item difficulty variance; (d) Sparse Q matrices are more tolerable than dense Q matrices because each cognitive component will have more weight in determining the actual cognitive components coefficients and will yield items that are less complex in structure; and (e) differential performance of sparse and dense Q matrices decreases as sample size and test length increase; and (e) for smaller sample sizes sparse Q matrices are preferred.

Overall, he concludes that sample size influences the recovery of item difficulty coefficients and cognitive components coefficients, which is consistent with what Baker (1993) and MacDonald and Kromrey (2011; 2012) found in their simulation studies. Further, in all conditions item difficulty in LLTM is more sensitive to measurement error than item difficulty for the Rasch model.

Cassuto recommends: (a) simple items with cognitive components coefficients that are independent. These conditions yield the best estimation of cognitive components coefficients and item difficulty coefficients; (b) sparse Q matrices are preferred over dense Q matrices, the theory being that sparse Q matrices allows the cognitive components coefficients to have greater influence.

In general: (a) cognitive components coefficients, followed by item difficulty, and then person ability coefficients had the best parameter estimation. Person ability is estimated only after item difficulty coefficients have been computed which depends on cognitive components coefficients estimation. Therefore, person ability can accumulate more error during estimation; (b) The estimation of item difficulty was not as accurate as the estimation of the cognitive components coefficients because item difficulty estimation is a composite of the estimation of

the cognitive components coefficients and, therefore, accumulates more error as the number of cognitive components increases; and (c) LLTM generated items may be useful in the computer adaptive environment.

Green and Smith (1987), Baker (1993) and Cassuto (1996) represent the sum of the LLTM literature on Q-matrix simulation. Their studies are foundational and informed decision making about research questions, and methods in this present study.

In sum their results are (a) the model works well and is stable; (b) a small degree of misspecification in the Q-matrix had a large impact on the parameter estimates; (c) in dense Q matrices the effects of misspecification are not quite as serious; (d) once the estimation of cognitive components coefficients have becomes stable, further increases in sample size have little additional effect on parameter estimates; (e) the recovery of item difficulty was generally good across all conditions but not as good as the recovery of cognitive components coefficients; (f) correlated sparse Q-matrix did not recover cognitive components as well when there were 50 or fewer examinees; (g) an orthogonal Q-matrix is optimal compared to a correlated Q-matrix; (h) Sparse Q matrices are more tolerable than dense Q matrices because each cognitive component will have more weight in determining the actual cognitive components coefficients and will yield items that are less complex in structure; (i) differential performance of sparse and dense Q matrices decreases as sample size and test length increase; (j) for smaller sample sizes sparse Q matrices are preferred; (k) sample size influences the recovery of item difficulty coefficients and cognitive components coefficients; (l) simple items with cognitive components coefficients that are independent yield the best estimation of cognitive components coefficients and item difficulty coefficients; (m) cognitive components coefficients, followed by item difficulty, and then person ability coefficients had the best parameter estimation in that order; and (n) LLTM generated items may be useful in the computer adaptive environment.

### **Embretson's Contribution**

Of the thinkers employing Fischer's LLTM since 1973, Embretson's work is the deepest and broadest. She has approached the use of LLTM from many different perspectives, explicating a variety of new and interesting models (e.g., 2-PL constrained, Saltus, and MRMLC). The interested reader is referred to Embretson's articles in the reference section of this dissertation and her published works on the subject. Embretson can be thought of as a pioneer in the field of Cognitive Diagnostic Assessment.

### **Other Q-matrix Simulation Studies in the CDA Family**

Given the scarcity of the LLTM Q-matrix simulation studies, and to deepen this portion of the literature review other simulation studies within the cognitive diagnostic family of modeling that have focused on the Q-matrix will be evaluated. These studies include (a) The DINA model (Rupp & Templin, 2008; de la Torre, 2008), (b) the Rule Space Method (Im & Corter, 2011), and (c) Log linear diagnostic classification model (Kunia-Habenicht, Rupp, and Wilhelm, 2012)

Jimmy de la Torre (2008) simulated Q matrices for the DINA model. He concluded that, when the Q-matrix is misspecified, parameter estimates for items show large biases and one of the two required parameters in this model, slip, is shrunken.

Rupp and Templin (2008) did an interesting simulation study examining the effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. In this simulation (a) the Q-matrix was under-specified, over-specified, and balanced misspecified; (b) an average number of items measuring an attribute and the average number of attributes measuring an item was maintained; and (c) the Q-matrix was under-specified, over-specified, and balanced misspecified for blocks of items that required a fixed number of attributes. They conducted misspecification for incorrect dependency. The latter condition can be thought of as analogous to the work of Cassuto on oblique correlations.



They found that (a) when a 0 is changed to a 1 in the Q-matrix (i.e., over-specification) the slipping parameter is overestimated but the guessing parameter is not affected, (b) when a 1 is changed to a 0 (i.e., under-specification) the slipping parameter is accurate but the guessing parameter is overestimated, and (c) when the Q-matrix was balanced misspecified the mean absolute deviation increased for the slipping and guessing parameter.

Lie, Xu, and Ying (2012) simulated a data driven approach to the specification of the Q-matrix. While not precisely on topic for this present study, it does, however, raise the possibility of the quantitative specification of the Q-matrix. They maintain, and it is a compelling argument, that misspecification of the Q-matrix leads to erroneous attribute identification. This may be overcome with the utilization of a Q-matrix estimator. They suggest a T matrix which establishes a connection between the observed response distribution and the model structure. In effect, it sets up a linear dependence between the attribute distribution, item difficulty, and the response distribution, person ability. For an interesting explanation of the simulation they conducted, the reader is referred to their original article. Q-matrix validation will continue to be an important challenge for researchers who employ cognitive diagnostic assessments and the development of quantitative measures to validate the Q-matrix could be a significant contribution to the field.

Im and Corter (2011) examined the consequences of attribute misspecification using the rule space method. The rule space method is a cognitive diagnostic assessment model which aims at explaining an examinees problem solving behavior via knowledge, processes, and skills which explain an individual's performance on an assessment. The study examines two types of attribute misspecification exclusion of an essential attribute needed for solving a test item, and inclusion of superfluous attributes that are not necessary for solving a test item. A Q-matrix from an actual test comprising 20 items was used. There were a total of eight Q matrices created for the study. The claim of the authors is that the successful classification of examinees

was 100%, no matter which superfluous attributes were added. When essential attributes were excluded, the classification rate dropped to approximately 90%.

The results of the simulation study show consistent bias from attribute misspecification but, in the authors' opinion, the amount of bias is not large. They concluded that when (a) an essential attribute is excluded the classification consistency of examinee attribute mastery was lower than when a superfluous attribute was added, (b) exclusion of an attribute may result in a reclassification of examinees who belonged to the excluded attribute, and (c) inclusion of an attribute may result in an over classification of an examinee which is a less deleterious outcome. When an essential attribute was excluded, the mean biases were negative and, when superfluous attributes are included, the mean biases are positive. They explain this phenomenon by suggesting that in attribute exclusion more examinees move from 1 to 0 and in attribute inclusion more examinees move from 0 to 1. Not surprisingly, the authors conclude that excluding an attribute in Rule Space is more detrimental than including a superfluous attribute. The result of excluding an essential attribute was the elimination of an essential knowledge state and the inclusion of a superfluous attribute introduces superfluous knowledge states.

Kunina-Habenicht, Rupp, and Wilhelm (2012) conducted a simulation of Log-Linear Diagnostic Classification Models. As in the case of Cassuto, the study is thorough and complex. The results are rich and we will touch briefly on them. The simulation varied (a) the number of respondents or examinees between 1,000 and 10,000; (b) the number of attributes, three and five; (c) the number of items, 25 and 50; (d) attribute correlations, .50 and .80; and (e) marginal attribute difficulties, equal versus different. These variables were studied for under-specification Q matrices, over-specification Q matrices and Balance misspecification Q Matrices. They also examined the utility of the AIC and BIC information based fit-indices.

The results of the study suggest that (a) with sample sizes of 200 or more relative model fit indices were able to point to the correct generating model with a high level of consistency; (b) under-specification, omitting attributes from the Q-matrix, had a detrimental effect on parameter recovery and examinee classification; (c) once sample size surpassed 500 intercepts and main effects were stable; and (d) to examine interaction effects sample sizes in the order of 4,000 were required.

The studies in this section included (a) the DINA model (Rupp & Templin, 2008; de la Torre, 2008), (b) the rule space method (Im & Corter, 2011), and (c) the log linear diagnostic classification model (Kunia-Habenicht, Rupp, and Wilhelm, 2012). These studies are foundational and informed decision making about research questions, and methods in this present study. These studies held that (a) when the Q-matrix is misspecified parameter estimates show large biases, (b) excluding an attribute in Rule Space eliminates an essential knowledge state and is more detrimental than including a superfluous attribute, (c) with sample sizes of 200 or more relative model fit indices were able to point to the correct generating model with a high level of consistency, (d) when sample size surpassed 500 intercepts and main effects were stable, and (e) research is being conducted to develop a quantitative measures of validating the Q-matrix.

### **The MacDonald LLTM and Simulation Studies**

MacDonald (Chen, MacDonald, & Leu, 2011) employed LLTM to validate a set of cognitive components that explain cognitive processes employed by Grade 5 and 6 Taiwanese students when they solve a twenty-three item fraction assessment. During this research MacDonald discovered that Fischer's software (LCPMWin 1.0) worked about half of the time. When the software worked it worked well but about half the time when the submit button was pushed the software simply did not run. For this reason and because LCPMWin 1.0 is not

easily commercially available, MacDonald began using SAS which De Boeck and Wilson (2004) had demonstrated could be used for the specification of LLTM modeling.

This decision raised questions about the functionality of SAS given the differences between CML used in LCPMWin 1.0 and Marginal Maximum Likelihood used in SAS. MacDonald (MacDonald & Kromrey, 2011) examined the parameter estimates produced by SAS for the cognitive components coefficient estimates when using (a) the Trust Region Optimization (TRUREG) algorithm, (b) the Newton-Raphson Optimization algorithm with Line Search (NEWRAP), (c) the Newton-Raphson Ridge Optimization algorithm (NRRIDG), (d) the Quasi-Newton Optimization (QUANEW) algorithm, (e) the Double-Dogleg Optimization (DBLDOG) algorithm, (f) the Conjugate Gradient Optimization (CONGRA) algorithm, and (g) the Nelder-Mead Simplex Optimization (NMSIMP) algorithm.

From within these optimization algorithms:

- QUANEW offers (a) DBFGS which performs the dual Broyden, Fletcher, Goldfarb; (b) Shanno (BFGS) update of the Cholesky factor of the Hessian matrix; (c) DDFP which performs the dual Davidon, Fletcher, and Powell (DFP) update of the Cholesky factor of the Hessian matrix; (d) BFGS which performs the original BFGS update of the inverse Hessian matrix; and (e) DFP which performs the original DFP update of the inverse Hessian matrix.
- DBLDOG has two updates available: (a) DBFGS performs the dual Broyden, Fletcher, Goldfarb, and Shanno update; and (b) DDFP performs the dual Davidon, Fletcher, and Powell update.
- CONGRA offers: (a) PB which performs the automatic restart update method of Powell (1977) and Beale (1972); (b) FR which performs the Fletcher-Reeves update (Fletcher 1987); (c) PR which performs the Polak-Ribiere update (Fletcher 1987); and (d) CD which performs a conjugate-descent update of Fletcher (1987).

This simulation work was conducted for the various estimation techniques when sample size, and number of cognitive components varied. The differences in parameter estimates obtained using the various algorithms were negligible and there was no non-convergence in any samples with any algorithm. Bias was negligible and confidence intervals were very accurate. Confidence intervals, as expected, were very wide with small sample sizes and the RMSE was large in these conditions. The simulation study provided evidence that the De Boeck and Wilson (2004) variant of the LLTM employing marginal maximum likelihood works very, very well. They concluded that PROC NLMIXED provides accurate parameter estimates for the LLTM models.

In what can be thought of as a pilot of this larger study, MacDonald working with Kromrey (MacDonald & Kromrey, 2012) conducted a simulation examining the amount of model misspecification the Q-Matrix can tolerate and still function adequately. This study was intended to provide some evidence of the functioning of these models when the Q-Matrix is properly specified, under specified, balanced misspecified, and over specified. All Q matrices were randomly misspecified at 1%, 5%, 10%, 15%, and 20% according to various sample sizes.

The authors concluded that (a) Q-Matrix under-specification results in cognitive components that are progressively more positively biased as the Q-Matrix moves away from the truth; (b) balanced misspecification and over-specification of the Q-Matrix results in cognitive components that are progressively more negatively biased as the Q-Matrix moves away from the truth; (c) confidence interval estimates of the parameter estimates lose accuracy with incorrect Q matrices, small amounts of misspecification are notable, and larger sample sizes are more inaccurate; (d) confidence Interval width is stable when the Q-Matrix is under-specified, balanced misspecified, and over-specified; (e) consistent with our earlier simulation study (MacDonald & Kromrey, 2011), LLTM model works very, very well when the Q-Matrix is under-specified, balanced misspecified, and over-specified; and (f) SAS PROC NLMIXED is superb at

estimating these models given LLTM converged in all conditions of misspecification in every replication in every sample size.

### **Summary**

In this chapter, a review of the educational uses of LLTM was conducted and a wide variety of applications were examined in three major groupings (a) mathematical models, (b) geometric and spatial models, (c) and paragraph comprehension and reading models. Fischer's modeling, which built on the work of German scholars like Scheiblechner, Kluwe and Spada, was examined in detail. There was a close examination of the estimation of the cognitive component coefficients and the computation of item difficulty. A number of major themes fell out of the literature (a) a wide variety of models have been validated using the LLTM, (b) LLTM has been used successfully to identify psychological units also known as cognitive components, (c) LLTM has not yet been adopted into the mainstream of quantitative analysis, and (d) Susan Embretson is a pioneer in the use and extension of LLTM modeling. The decision of De Boeck and Wilson to use marginal maximum likelihood and the implications were introduced and discussed. It was noted that SAS's implementation of LLTM in PROC NLMIXED offered a full modeling of both item difficulty and person ability and unlike LLTM explicated by Fischer item difficulty is not estimated as a sufficient statistic. The Q-matrix simulation research, which is sparse, was examined in detail with a view to informing the methods and discussion in this dissertation. Finally, a discussion of recent LLTM studies conducted by MacDonald with Chen and Kromrey was presented.

## CHAPTER THREE: METHOD

### **Study Purpose**

The purpose of this study was to provide some evidence to help determine if the Linear Logistic Test Model (LLTM) functions well using the SAS NLMIXED procedure and is robust when (a) the Q-matrix is progressively more misspecified; (b) the Q-matrix is properly specified, under-specified, balanced misspecified, and over-specified; (c) the sample size is varied; (d) the Q-matrix is densely and sparsely populated; (e) test length varies; and (f) the distribution of person ability is normally distributed, negatively skewed, and positively skewed.

### **Research Questions**

Questions 1 through 6 examined:

- 1.) To what extent does the LLTM function well when the Q-matrix is progressively more misspecified?
- 2.) To what extent does the LLTM function well when the Q-matrix is properly specified, under specified, balanced misspecified, and over specified?
- 3.) To what extent does the LLTM function well under different conditions of model misspecification when the sample size varies?
- 4.) To what extent does the LLTM function well under different conditions of model misspecification when the Q-matrix is densely or sparsely populated?
- 5.) To what extent does the LLTM function well under different conditions of model misspecification when test length varies?
- 6.) To what extent does the LLTM function well under different conditions of model misspecification when the population distribution is normally distributed, negatively skewed, and positively skewed?

## **SAS 9.3**

### **NLMIXED**

The simulations in this study were conducted in SAS 9.3 (SAS Institute Inc., 2008). SAS is a powerful and flexible statistical environment which offers the researcher PROC IML. This procedure is an Interactive Matrix Language (IML) in which the data were simulated. All the various conditions of percent of misspecification, form of misspecification, test length, sample size, and distribution of person responses were controlled in PROC IML SAS code.

The SAS coder should note, the data need to be imported into SAS and organized in a vector or vertical string and sorted by the number of persons by the items per person before being submitted to PROC NLMIXED for analysis.

### **Estimation Technique**

SAS is a powerful environment which offers the researcher PROC NLMIXED. This SAS procedure fits nonlinear mixed models that can estimate fixed effects and random effects. PROC NLMIXED fits the non-linear mixed models by maximizing an approximation to the likelihood integrated over the random effects. For the interested reader the technical aspects can be found in the SAS/STAT User's Guide (SAS/STAT® 9.2, 2008)

When employing PROC NLMIXED in SAS the programmer has a variety of algorithms for parameter estimation. A recent simulation study found that all of the estimations techniques work well with the exception of the Nelder-Mead Simplex Optimization (NMSIMP) algorithm when seven or more cognitive components are estimated (MacDonald & Kromrey, 2011). Given the choices, the double dog leg estimation technique (tech=DBLDOG) was selected because it performs well and executes quickly.



## **Replications**

One thousand replications are commonly used when conducting simulation studies. This study used 1,085 replications based on table values (see Robey & Barcikowski, 1992) for the liberal criterion of  $\alpha \pm \frac{1}{2}\alpha$ ,  $1-\beta=.8$ , and  $\omega=.01$ .

## **Cognitive Components**

Most research designs employ between four and eight cognitive components (Rupp & Templin, 2008). In a recent simulation study three and five cognitive components were simulated (Kunina-Habenicht, Rupp, & Wilhelm, 2012). MacDonald and Kromrey (2011) examined three, five and seven cognitive components. The final results of this study for five and seven cognitive components have not yet been published; however, the researchers found that, with the exception of time required for parameter estimation, there was no significant difference in parameter estimation when using three, five or seven cognitive components. With these facts a decision was made to examine five cognitive components.

## **SAS Code**

The SAS code for the simulation used to create the Q-matrix, simulate data via SAS/IML, impose misspecification, and conduct the LLTM analysis is found in appendix A.

## **Code Validation**

The author conducted multiple LLTM studies between 2008-2012 and constructed many LLTM models using SAS and LCPMWin 1.0. These results have been presented at various conferences and published in peer reviewed journals. The code provided by De Boeck and Wilson (2004) was replicated in SAS 9.3 and the results were calibrated to those provided by De Boeck and Wilson (2004). At all phases of code development the SAS log was closely examined for warnings, and errors. The completed code was submitted to the co-major professors of this dissertation for examination and evaluation at every phase of the development. On two occasions a 'trace' was conducted to check IML and NLMIXED

computations. All matrices were examined as they were created and where possible; results were estimated or calculated in Excel.

### **The Simulation**

This simulation consumed approximately 5,148 hours of computer time based on an I-7 computer processing unit. This researcher would like to acknowledge the use of the services provided by Research Computing at the University of South Florida in conducting the simulation for this dissertation using the high performance computing (HPC) technology available at the University of South Florida. HPC at USF is defined as computing consisting of hardware and software resources used for computational science that are beyond what is commonly found on the desktop machine. Research Computing, a department within Information Technology, was established to promote the availability of high performance computing resources essential to effective research at the University of South Florida. Research Computing supports software tools, high performance computer hardware, and training for both faculty and students. Research computing maintains the CIRCE cluster. This system is currently being upgraded to 6000 cores. It uses a Lustre parallel file system for fast IO, and Infiniband for a computational interconnect. The file systems on the cluster are available for remote mounting via several protocols to aid in data movement. (Information retrieved from USF website at [http://www.usf.edu/it/research-computing/services/.](http://www.usf.edu/it/research-computing/services/))

The code was broken into 270 batches and submitted to the HPC system. Within two hours all 270 were started. The simulation began on a Friday at 5:30 and after allowing for problems completed on Wednesday morning at approximately 2:00 am. The simulation produced approximately 5.4 terabytes of output of which approximately 200 gigabytes was transferred to permanent SAS data files for analysis and reporting. In total, 2,050,650 replications were conducted during the simulation.

### Q-matrix

A Q-matrix or Weight matrix is specified *a priori* and is required to conduct a LLTM analysis. In this study a Q-matrix representing the truth is found in Table 3.

Table 3

*A 20 by 5 True Q-matrix that Describes the Relationship between Items and Cognitive Components*

Item	CC1	CC2	CC3	CC4	CC5
1	1	0	0	0	1
2	1	1	0	0	0
3	1	1	1	0	0
4	1	1	1	1	0
5	0	1	1	1	1
6	1	0	0	0	1
7	1	1	0	0	0
8	1	1	1	0	0
9	1	1	1	1	0
10	0	1	1	1	1
11	1	0	0	0	1
12	1	1	0	0	0
13	1	1	1	0	0
14	1	1	1	1	0
15	0	1	1	1	1
16	1	0	0	0	1
17	1	1	0	0	0
18	1	1	1	0	0
19	1	1	1	1	0
20	0	1	1	1	1

### Percentage of Q-Matrix Misspecification

Misspecification has been the key design employed when studying Q-matrices (Baker, 1993; Cassuto, 1996; Kunina-Habernicht, Rupp & Wilhelm, 2012; MacDonald & Kromrey, 2012; Rupp & Templin, 2008; Rupp et al., 2012). For the purposes of this simulation the Q-matrix was randomly misspecified at 1%, 5%, 10%, and 15% similar to the work of Baker (1993). In practical terms, when the Q-matrix is misspecified a 1 was transformed into 0 or a 0 was transformed into a 1.

## **Form of Misspecification**

The form of Q-matrix misspecification (i.e., under-specification, over-specification, and balanced misspecification) has been a key design factor in the Q-matrix simulation literature (Baker, 1993; Cassuto, 1996; Kunina-Habernicht, Rupp & Wilhelm, 2012; MacDonald & Kromrey, 2012; Rupp & Templin, 2008; Rupp et al., 2012).

### **Misspecification**

#### **Under-Specification**

Under-specification in this study has a particular definition. It refers to changing 1s to 0s in the true Q-matrix. In practical terms for a 100 cell matrix, when the Q-matrix is misspecified 1% a 1 was transformed into 0. Or again, when the Q-matrix is misspecified at 10%, ten 1's were transformed into 0's.

#### **Balanced Misspecification**

Balanced misspecification has a particular definition within this study. For every 0 which is transformed into a 1 in the Q-matrix a different 1 was transformed into a 0. In this way the Q-matrix is said to be misspecified but balanced. In practical terms for a 100 cell matrix, this means that when the Q-matrix is balanced misspecified at 5% for every five 1's transformed into 0's there were five other 0's transformed into 1's. The Q-matrix was randomly misspecified at 1%, 5%, 10%, and 15%.

There are at least two definitions of balanced misspecification. The full percentage of 1s are transformed into 0s and the full percentage of 0s are transformed into 1s (e.g, for a 100 cell matrix 10% ten 1s would be transformed into 0s and ten 0s would be transformed into 1s). An alternate definition of balanced misspecification is the transformation of half of the percentage of 1s into 0s and the transformation of half of the percentage of 0s into 1s (e.g., for a 100 cell matrix 10% five 1s would be transformed into 0s and five 0s would be transformed into 1s). For

the purposes of this study a decision was made to transform the full percentage of 1s into 0s and the full percentage of 0s into 1s.

### **Over-specification**

Over-specification has a particular definition within this study. A 0 was randomly transformed into a 1 in the Q-matrix. In practical terms for a 100 cell matrix, this means that for every Q-matrix over-specified 15% fifteen 0's were transformed into 1's in the Truth matrix. The Q-matrix was randomly misspecified at 1%, 5%, 10%, and 15

### **Sample Size**

Sample size has been a key design factor in the Q-matrix simulation literature (Baker, 1993; Cassuto, 1996; Kunina-Habenicht, Rupp, & Wilhelm, 2012; MacDonald & Kromrey, 2012; Yoes, 1990). Baker (1993) simulated sample sizes of 20, 100, 500 and 1000. Cassuto (1996) examined sample sizes of 50, 250, and 1000. Kunina-Habenicht, Rupp, and Wilhelm (2012) used 1,000 for main effects. MacDonald and Kromrey (2011) employed 20, 40, 80, 160, 320, 640, 1280, and 2560. Considering the sample sizes employed in previous simulation studies, and given that LLTM parameters tend to become stable somewhere between 500 and 1,000 examinees, (MacDonald & Kromrey, 2011) this research was simulated on samples of 20, 40, 80, 160, 320, 640, and 1,280.

### **Q-Matrix Density**

#### **Dense Matrix**

Researchers have consistently specified dense and sparse matrices when examining Q matrices (Baker, 1993; Cassuto, 1996; Kunina-Habenicht, Rupp, and Wilhelm, 2012; Rupp & Templin, 2008). A dense Q-matrix in the simulation literature used by Baker (Baker, 1993) specified 96 1's in a 168 cell matrix which is a 57.1% cell density. Cassuto calls for a dense Q-matrix to hold at least 70% 1s, but he also notes this is a purely arbitrary decision (Cassuto, 1996). For the purpose of this study, a dense Q-matrix is defined as a Q-matrix in which the

cognitive components are specified within the Q-matrix 60% of the time. In other words, given a 20 by 5 Q-matrix of 100 cells a dense Q-matrix is one in which 60 cells were specified with 1s. This decision was made, in part, to avoid saturating the Q-matrix during 15% over-specification. Table 4 presents an example of a dense Q-matrix which was used in this study.

Table 4

*Dense 5 by 20 True Q-matrix (60% 1s)*

Item	CC1	CC2	CC3	CC4	CC5
1	1	0	0	0	1
2	1	1	0	0	0
3	1	1	1	0	0
4	1	1	1	1	0
5	0	1	1	1	1
6	1	0	0	0	1
7	1	1	0	0	0
8	1	1	1	0	0
9	1	1	1	1	0
10	0	1	1	1	1
11	1	0	0	0	1
12	1	1	0	0	0
13	1	1	1	0	0
14	1	1	1	1	0
15	0	1	1	1	1
16	1	0	0	0	1
17	1	1	0	0	0
18	1	1	1	0	0
19	1	1	1	1	0
20	0	1	1	1	1

### **Sparse Matrix**

Researchers have consistently specified sparse matrices when examining Q matrices (Baker, 1993; Cassuto, 1996; Kunina-Habenicht, Rupp, & Wilhelm, 2012; Rupp & Templin, 2008). A sparse matrix in the simulation literature has specified 34 1s within a one hundred sixty-eight cell Q-matrix (Baker, 1993) which is a 20.2% density. Cassuto (1996) noted that there is no exact definition of sparse Q matrices; therefore, he arbitrarily set a sparse Q-matrix with 40% 1s. For the purpose of this study a sparse Q-matrix is defined as a Q-matrix in which

the cognitive components are specified, within the Q-matrix, 46% of the time. In other words, given a 20 by 5 Q-matrix of 100 cells a sparse Q-matrix is one in which 46 of the cells were specified with 1s. When CC1-CC5 is considered at least one of the five cognitive components must be specified with a 1 or the item does not have any of the cognitive components present. Given that each item requires at least the presence of one cognitive component and that underspecification of 15% will reduce the number of ones in the matrix to 31 it was decided that 46% specification was appropriate. An example of a sparse Q-matrix is found in Table 5.

Table 5

*Sparse 20 by 5 Q-matrix (46% 1s)*

Item	CC1	CC2	CC3	CC4	CC5
1	1	0	0	1	1
2	1	1	0	0	0
3	1	1	1	0	0
4	0	0	0	1	1
5	0	0	1	1	0
6	1	0	0	0	1
7	1	1	0	0	0
8	1	1	1	0	0
9	0	0	1	0	1
10	0	1	0	1	0
11	1	0	0	0	1
12	1	1	0	1	0
13	1	1	1	0	0
14	0	0	1	0	1
15	0	1	0	1	0
16	1	0	0	1	0
17	1	1	0	0	0
18	1	1	0	0	0
19	0	0	1	0	1
20	0	0	1	1	1

### Test Length

Test length is an important design factor affecting item parameter estimation and has been a variable used in simulation studies in the literature (Baker, 1993, Cassuto, 1996; Kunina-Habernicht, Rupp & Wilhelm, 2012; Rupp & Templin, 2009). Common test lengths range from

15 to 100 items (Yoes, 1990). Cassuto (1996) examined tests that were 20, and 60 items. Henson and Templin (2009) used 40 items. In a recent simulation study 25 and 50 items were simulated (Kunina-Habenicht, Rupp, and Wilhelm, 2012). For the purposes of this study test lengths of 20, 40, and 60 items were simulated.

### **Skewness of Person Ability Distribution**

The distribution of person ability has received some attention in the literature (Cassuto, 1996). This study examined the distribution of person ability when it is positively and negatively skewed. Cassuto suggested a negatively skewed distribution can represent advanced classes with individuals or participants in the upper level of the distribution. Cassuto employed a negative skew of -0.50 (Cassuto, 1996). A negative skew may also occur in educational data when high achievers are over-represented in a sample (e.g., students taking the GRE examination). For the purposes of this study a skewness of -0.5, a normal distribution, and a skewness of 0.5 was examined.

The method used to simulate non-normal distributions was explicated by Fleishman (1978). Fleishman's polynomial transformation formula used in this simulation for parameters b, c, and d was:

$$\text{theta} = (-1 * c) + (b * \text{theta}) + (c * \text{theta}^2) + (d * \text{theta}^3)$$

for a skewness of -0.50 and a kurtosis of 2 the code written in SAS was:

$$\text{theta} = (-1 * .06416925946524) + (.85011102914029 * \text{theta}) + (.06416925946524 * \text{theta}^2) + (.04641702467833 * \text{theta}^3)$$

### **Parameters of Interest**

In this study three parameters of interest were estimated: (a) the cognitive component ; (b) item difficulty; and (c) person ability.



## Evaluation Criteria

To evaluate the estimated parameters, statistical bias, the root mean square error (RMSE), estimated confidence interval coverage (CI coverage), and the mean confidence interval width (CI width) were computed and analyzed.

### Statistical Bias

The bias statistic informs the researchers (a) how close the average estimated cognitive component,  $\bar{\alpha}_j$ , is to true cognitive component,  $\alpha_j$  ( $Bias = \alpha_j - \bar{\alpha}_j$ ); (b) how close the average estimated item difficulty,  $\bar{\beta}_{LLTM}$ , is to true item difficulty,  $\beta_{LLTM}$ . ( $Bias = \beta_{LLTM} - \bar{\beta}_{LLTM}$ ); and (c) how close the average estimated examinee ability,  $\bar{\theta}_p$ , is to true examinee ability,  $\theta_p$  ( $Bias = \theta_p - \bar{\theta}_p$ ). Bias is a relative fit index related to the parameters of interest which are the cognitive components, item difficulty, and person ability in this study. The cognitive components coefficients produced by the LLTM analysis are standard deviations around zero in the logit unit. Cohen (1988) described one fifth of a standard deviation as the starting point for a small effect size (i.e., 0.20-0.49 is a small Cohen's d effect size). For the purposes of this research anything less than 0.20, or less than a small effect size, was considered tolerable. A practical example may help the reader evaluate this decision. Researchers often choose to rescale IRT parameter estimates with a mean of 300, a standard deviation of 50, and a range of  $\pm 4$  standard deviations creating an index of 100-500 points. In this environment a cognitive component whose parameter estimate is 1.2 would have a scale score of 360 ( $300 + (50 * 1.2)$ ). Tolerating bias of  $\pm 0.20$  the score could be 370 ( $300 + (50 * (1.2 + 0.2))$ ) or 350 ( $300 + (50 * (1.2 - 0.2))$ ). The decision reached for the purposes of this simulation is that bias quantified as less than a small effect size by Cohen (1988) is tolerable.

## Root Mean Square Error (RMSE)

RMSE has been used (Fischer, 1973; Green & Smith, 1987; Baker, 1993) to detect the magnitude of the estimation error. The mean squared error (MSE), whose introduction is usually credited to Carl Friedrich Gauss (1809), quantifies the difference between values implied by an estimated parameter and the truth. The MSE is the second moment around the origin of the error and combines the variance of the estimated parameter and its bias. The square root of the MSE returns the root-mean-square error which is used because it is expressed in the same units as the quantity being estimated. The RMSE helps to quantify (a) the typical difference between the true and estimated values of the cognitive components

$\left( RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{\alpha}_j - \alpha_j)^2}{n}} \right)$ ; (b) the typical difference between the true and estimated values of

the item difficulty  $\left( RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{\beta}_i - \beta_i)^2}{n}} \right)$ ; and (c) the typical difference between the true and

estimated value of the person ability  $\left( RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{\theta}_p - \theta_p)^2}{n}} \right)$ . An RMSE of zero is

interpreted as meaning the estimator predicts the observed latent parameter without error.

Practically speaking, no parameter estimate is free from error. The RMSE is a relative fit index related to the parameters of interest which are the cognitive components, item difficulty, and person ability in this study. For the same reasons presented in the bias section above, an RMSE of less than 0.20 was tolerable.

### Confidence Interval Coverage and Confidence Interval Width

The  $100(1-a)\%$  Wald confidence intervals (CIs) are based on the asymptotic normality of the parameter estimators. These CIs provide information about accuracy and precision and can be used to infer significance. CIs for  $\beta_j$  in SASs non-linear mixed is given by:

$$\hat{\beta}_j \pm Z_{1-a/2} \hat{\sigma}_j \quad (3.1)$$

where  $Z_p$  is the  $100_p$ th percentile of the standard normal distribution

$\hat{\beta}_j$  is the maximum likelihood estimate of  $\beta_j$  and

$\hat{\sigma}_j$  is the standard error estimate of  $\hat{\beta}_j$

CI coverage indicates the degree to which the parameter estimate is accurate. As a default SAS constructs CIs around cognitive components and person ability parameter estimates. CI coverage is calculated as the percentage of time the truth parameter, specified by the researcher in this study, falls within the 95% confidence interval constructed around the estimated cognitive component and person ability parameters. The higher the percentage of truth parameters that fall within the estimated cognitive components or person ability confidence intervals the greater the accuracy of the parameter estimates. The nominal coverage probability was set at .95 for this study. It is hoped that the nominal coverage probability and the actual coverage probability will be approximately equal.

CI width will inform the researcher about the precision of the interval estimate, and was calculated as the average difference between the upper and lower limits of the CI for the cognitive component and person ability estimates. The CI width is a relative index rather than an absolute index. For the purposes of this study 95% confidence intervals were constructed with the understanding researchers prefer to see narrow CI widths.

### **Cognitive Components Pairwise Phi Correlation Coefficients**

The Pearson product-moment coefficient,  $r$ , calculated on dichotomous data is called the *phi coefficient*,  $r_\phi$ . Pearson's phi coefficient is similar in interpretation to the Pearson product moment correlation coefficient and describes the strength and direction of the relationship between two dichotomous variables (Glass & Hopkins, 1996). Pairwise cognitive components in the Q-matrix with high positive phi correlation coefficients indicate a predominance of co-occurring concepts (i.e., 0,0 & 1,1) within items on an examination, strong negative phi correlation coefficients indicates a predominance of uniquely assessed concepts (i.e., 0,1 & 1,0) within items on the examination, and phi correlation coefficients approaching zero suggest a mixture of co-occurring (i.e., 0,0 & 1,1) and uniquely assessed (i.e., 0,1 & 1,0) concepts within items on the examination.

The following example may help illustrate this effect. Let CC1 represent addition and CC2 represent subtraction. Table 6 presents three examples of how addition and subtraction might be specified within items on an assessment. Examining the specification in the first pair of cognitive components reveals that when addition, CC1, is present in an item subtraction, CC2, is not specified. The pattern is either 0,1 or 1,0 across all 20 items. Addition and subtraction can be thought of as uniquely assessing addition and subtraction within items on the assessment. The cognitive component phi correlation coefficient in this case is -1.00. In this example the examinee would be required to either add or subtract to solve the item, but never both concepts together within items on the exam.

Examining the specification for the second pair of cognitive components reveals that when addition, CC1, is present in an item subtraction, CC2, is also present. The pattern is either 0,0, or 1,1 across all 20 items. Addition and subtraction can be thought of as predominantly co-occurring within items on the assessment. The cognitive components phi

correlation in this case is 1.00. In this instance, the examinee would be required to predominantly add and subtract to solve items on the exam.

Table 6

*Phi Pairwise Cognitive Component Correlation Coefficients Exemplifying Negative, Positive, and Neutral Phi Correlations*

Item	CC1	CC2	CC1	CC2	CC1	CC2
1	1	0	0	0	1	0
2	0	1	0	0	0	1
3	1	0	1	1	1	0
4	1	0	1	1	1	0
5	1	0	0	0	1	0
6	0	1	0	0	0	1
7	0	1	1	1	0	1
8	0	1	1	1	0	1
9	1	0	0	0	1	0
10	0	1	0	0	0	1
11	0	1	1	1	1	1
12	1	0	1	1	1	1
13	1	0	0	0	0	0
14	0	1	0	0	0	0
15	1	0	1	1	1	1
16	1	0	1	1	1	1
17	0	1	0	0	0	0
18	1	0	0	0	0	0
19	1	0	1	1	1	1
20	0	1	1	1	1	1
Phi Corr	-1.00		1.00		-0.01	

Examining the specification for the third pair of cognitive components reveals that in items 1-10 when addition, CC1, is present in an item subtraction, CC2, is not specified. The pattern is either 0,1 or 1,0 across all 10 items. . Addition and subtraction can be thought of as uniquely assessing addition and subtraction within items 1-10. In items 11-20 when addition, CC1, is present in an item subtraction, CC2, is also present. The pattern is either 0,0, or 1,1 across the final 10 items. Addition and subtraction can be thought of as predominantly co-occurring within items on items 11-20. The cognitive components phi correlation coefficient in

this case is -0.01. In this instance, the examinee would be required to add or subtract in items 1-10 on the exam but never the two together, and on items 11-20 of the exam the examinee would always be required to add and subtract to solve an item. In this instance there is a mixture of co-occurring or uniquely assessed concepts on items in the exam.

For the purposes of this study high positive phi coefficients will be referred to as cognitive component vectors with a predominance of co-occurring concepts, high negative phi coefficients will be referred to as cognitive components vectors with a predominance of uniquely assessed concepts, and phi confidents approaching zero will be referred to as cognitive components vectors with a mixture of co-occurring, and uniquely assessed concepts.

### **Data Analysis**

Research questions one through six were addressed by computing and comparing bias, CI Width, RMSE, and CI Coverage values in the cognitive components, item difficulty, and person ability estimates. These values were computed for 0% or truth, 1%, 5%, 10% and 15% misspecification by form of misspecification by sample size by density of the Q-matrix by number of items by skewness of person ability. The differences in bias, RMSE, CI Coverage, and CI Width for the cognitive components, item difficulty, and person ability were averaged over the 1085 replications in the study. Boxplots for the average overall bias, RMSE, CI Coverage, and CI Width were computed for the cognitive components, item difficulty, and person ability. When examination of the boxplots suggested further investigation was warranted an analysis of variance (ANOVA) was conducted for the six design factors in this study (i.e., percent of misspecification, form of misspecification, sample size, density of Q-matrix, test length, and skew of person ability distribution) for Bias, RMSE, CI coverage, and CI width. The eta squared values of the main effects and first level interaction effects (i.e., the six design factors and their interactions) were computed to determine if the effects size is greater than 0.0588, which is defined as a medium effect size (Cohen, 1988).

To explore the strength and direction of the relationships between the cognitive components in the Q-matrix the Pearson phi correlation coefficients were computed (N=2,050,650) for 0% or truth, 1%, 5%, 10% and 15% misspecification by form of misspecification by sample size by density of the Q-matrix by number of items and by skewness of person ability. The Fisher Z-transformed values were computed and the values compared across Q-matrices. The differences in the Pearson phi correlation coefficient for the cognitive components in the Q-matrix were averaged and analyzed over the 1085 replications in the study.

### **Major Steps in the SAS Simulation Code**

Table 7 outlines the major steps in the SAS simulation code.

Table 7

#### *Major Steps in SAS Code*

Step	Brief Description
Step 1	Set Replications to 1085
Step 2	Set estimation Technique to DBLDOG
Step 3	Construct Truth Q-matrix in PROC IML
Step 4	Set Parameters for the Cognitive Components
Step 5	Generate Person Abilities and Item Responses
Step 6	Set Percent and Form of Misspecification to be Imposed on the Q-matrix
Step 7	Set Skew
Step 8	Set Sample Size
Step 9	Set Test Length
Step 10	Set Density of Q-matrix
Step 11	Analyze simulated data in PROC NL MIXED

## CHAPTER FOUR: RESULTS

Green and Smith (1987), Baker (1993) and Cassuto (1996) represent the sum of the LLTM literature on Q-matrix simulation. Further, this study is the first to focus on the cognitive components vectors within the Q-matrix as opposed to examining the functioning of the Q-matrix as a matrix.

To examine the effect of the design factors (i.e., percent of misspecification, form of misspecification, sample size, density of Q-matrix, test length, and skew of person ability distribution) on item difficulty, cognitive components, and person ability the results from the simulation were analyzed by constructing and examining box-and-whisker plots for bias, CI coverage, RMSE, and CI width. Bias was selected because it measures how close the estimates of cognitive components, item difficulty, and theta are to the truth when they are impacted by the design features in this study. The researchers want to see box-and-whisker plots with means close to zero but less than  $\pm 0.20$  and tight boxes which are interpreted as the standard deviations being small. Root mean squared error quantifies the difference between values implied by an estimated parameter and the truth. The RMSE combines the variance of the estimated parameter about the mean and its bias. RMSE will help quantify the impact of the design factors on the cognitive components, item difficulty, and person ability. The researchers want to see RMSE box-and-whisker plots with means close to zero but less than 0.20 with tight boxes which are interpreted as the standard deviations being small. Confidence Interval coverage was selected because it measures how often the cognitive components, item difficulty, and person ability parameter estimates fall within the 95% confidence interval when they are impacted by the design features in this study. The researchers want to see box-and-whisker plots close to the nominal level and tight boxes which are interpreted as the standard deviations



being small. Confidence interval width was selected because it measures how precise the parameter estimates of cognitive components, item difficulty, and person ability are when the parameter estimates are impacted by the design factors in this study. The researchers want to see box-and-whisker plots with tight boxes which are interpreted as the standard deviations being small.

Next, an ANOVA was conducted to discover what impact, if any, the design factors (i.e., percent of misspecification, form of misspecification, sample size, density of Q-matrix, test length, and skew of person ability distribution) had on item difficulty, cognitive components, and person ability as measured by bias, CI coverage, RMSE, and CI width. An eta squared proportion of variance effect size, not to be confused with a Cohen's d effect size, was computed for the design factors and first order interaction effects. Cohen's (1988) categorization of eta squared effect sizes in which he defines 0.0099 as a small effect size, 0.0588 as a medium effect size, and 0.1379 as a large effect size was used in this study. Conclusions will be drawn employing medium effect sizes. Tabular and graphical materials were constructed to provide support for the presentation and discussion of results.

Examination of box and whisker plots for overall distribution of statistical bias, estimated root mean squared error, estimated CI coverage, and CI width across all simulation conditions demonstrated heterogeneity of variance across cognitive components, item difficulty, and person ability. A decision was made to analyze the results of the simulation study for the estimated parameters individually because (a) CC1-CC5 performed differentially and any aggregate analysis risked masking individual effects, (b) beta is computed as a linear composition of CC1-CC5 and performed differentially when compared to CC1-CC5, (c) theta is estimated employing empirical Bayes methodology while the cognitive components, and indirectly beta, are computed employing the double dog leg technique algorithm within the

marginal maximum likelihood technique, and (d) theta performed differentially when compared to beta and the cognitive component coefficients.

The decision to analyze cognitive components, item difficulty (beta), and person ability (theta) separately is consistent with the previous simulation reporting done by Cassuto (1996) in his doctoral dissertation. The decision to examine cognitive components separately is a departure from the reporting done for simulation work in the past (Baker, 1993; Cassuto, 1996; Green, 1987); however, it is warranted given the differential nature of the cognitive components functioning in this study.

To examine the model fit of the Q-matrices an ANOVA was conducted for the BIC, AIC, and AICC, for each of the design factors in the study (percent of misspecification, form of misspecification, sample size, density of the Q-matrix, number of items, and skewness of the person ability distribution). Finally, tabular and graphical materials were constructed to provide support for the presentation and discussion of results.

Correlation matrices were computed for truth Q-matrices and the misspecified Q-matrices across replications for all design factors in this simulation study. The Fisher Z-transformed values were computed and the values compared across Q-matrices. Next, the correlational matrices were averaged across replications for the truth Q-matrices and the misspecified Q-matrices for all design factors.

Recall that (a) item difficulty is a linear combination of the estimated cognitive components coefficients, (b) cognitive components coefficients are valid latent psychological units whose estimated parameters are capable of bridging cognitive processing and psychometric models, and (c) person ability in LLTM modeling representing a person's ability level estimated as a random effect over the population of persons computed using an empirical Bayes approach.

## Estimated Bias

The bias statistic will inform the researchers how close the average estimated parameters of cognitive components, item difficulty and person ability are to the truth. Smaller bias values closer to 0 suggest the parameter estimates are closer to the truth. In this study the researcher prefers bias values to be less than  $\pm 0.20$  which is less than a small effect as defined by Cohen (1988).

The average overall distribution of statistical bias across all simulation conditions for item difficulty, five cognitive components, and person ability are presented in Figure 3. The mean, standard deviation, minimum and maximum overall average estimated bias values by beta (item difficulty parameter), CC1-CC5 (cognitive component parameters 1-5), and theta (person ability parameter) are presented in Table 8.

Table 8

*Means, Standard deviations, Minimum, and Maximum Bias Values for Beta, CC1-CC5, and Theta (N=1890)*

	MEAN	SD	MIN	MAX
Beta	-0.2394	0.2167	-0.7884	0.0347
CC1	-0.0550	0.0946	-0.3529	0.1559
CC2	-0.1913	0.1494	-0.5071	0.0255
CC3	0.0789	0.0850	-0.0143	0.2972
CC4	-0.0739	0.1030	-0.4041	0.1144
CC5	-0.2545	0.2413	-0.8347	0.0401
Theta	-0.1743	0.1938	-0.7062	0.0458

Table 8 values and Figure 3 boxplots demonstrate that the means bias values for beta, theta, and CC2, and CC5 are not close to 0 and have boxes which are not tight. In other words, the mean bias values for item difficulty, person ability, and cognitive components 2 and 5 have large standard deviations.

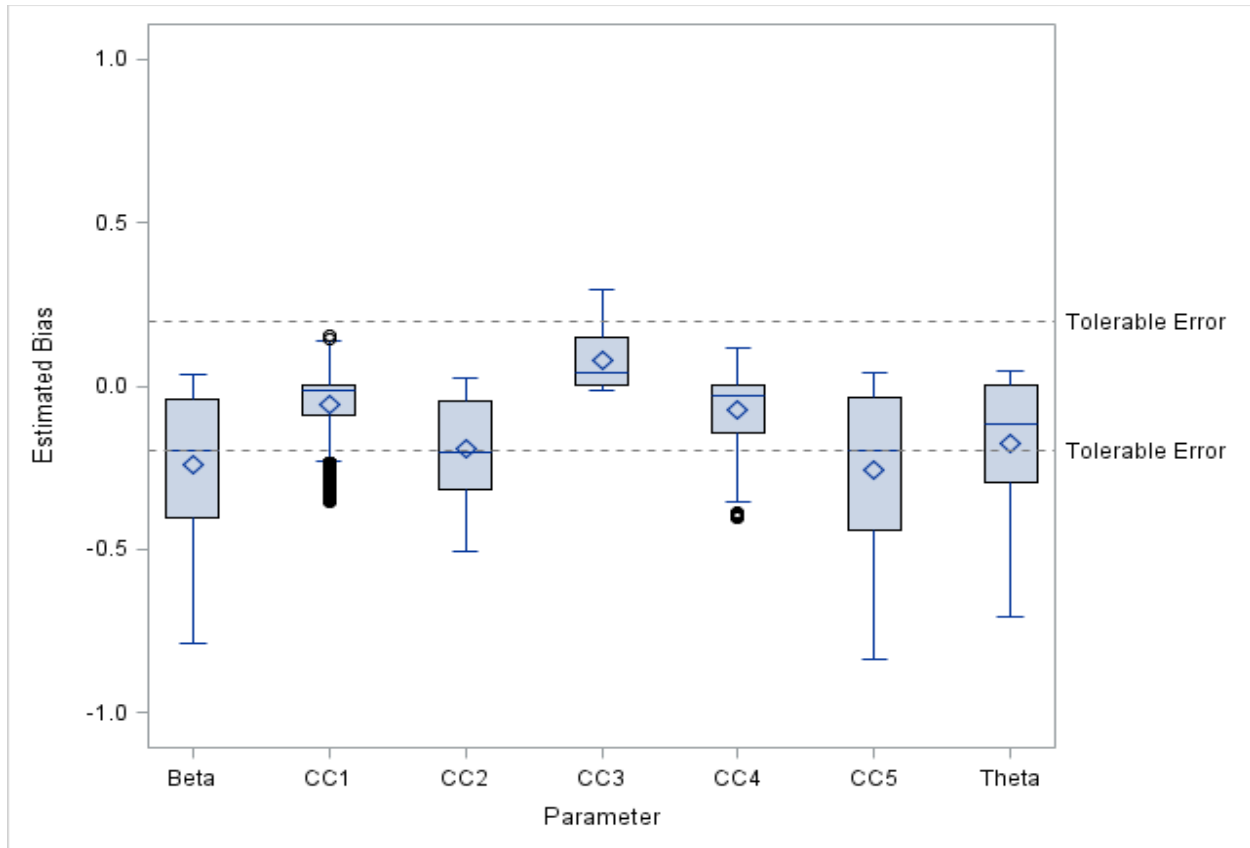


Figure 3. Distribution of average overall statistical bias estimates across beta, CC1-CC5, and theta.

An ANOVA for the design factors and their first order interaction effects was computed for item difficulty, cognitive components, and person ability as measured by the overall estimated bias values. Recall that bias is a measure of how far away the mean estimated values for cognitive components, item difficulty or person ability parameter estimates are from their true values. Table 9 presents eta-squared values, ( $\eta^2$ ), which quantifies the impact the design factors and first order interaction effects have on item difficulty, cognitive components, and person ability parameter estimates as measured by bias.

The design factors and their first order interaction effects total eta squared values are .9406 for CC1, .9944 for CC2, .9641 for CC3, .9771 for CC4, .9915 for CC5, .9883 for beta, and

.9875 for theta. In other words, the design factors and their first order interaction effects accounted for 94.06% of the variance in CC1, 99.44% of the variance in CC2, 96.41% of the variance in CC3, 97.71% of the variance in CC4, 99.15% of the variance in CC5, 98.83% of the variance in beta, and 98.75% of the variance in theta.

Table 9

*Eta-Squared Values for the Association of Design Factors and 1st Level Interaction Effects with the Average Estimated Overall Bias for CCs, Beta, and Theta*

	CC1	CC2	CC3	CC4	CC5	Beta	Theta
Percent	** .2190	** .8031	** .5013	** .3821	** .6822	** .7031	** .6200
Form	* .1319	.0425	.0174	** .2554	** .1559	.0561	.0514
Sample Size	.0001	.0001	.0001	.0008	.0004	.0005	.0000
Density	** .1988	.0362	** .2254	.0348	.0092	.0027	.0063
Items	* .1085	.0469	.0362	.0521	.0285	* .1245	** .1839
Skew	.0014	.0003	.0001	.0001	.0001	.0007	.0009
Percent*Form	* .0831	.0204	.0105	** .1879	* .0933	.0276	.0269
Percent*SS	.0000	.0000	.0001	.0000	.0000	.0000	.0000
Percent*Density	* .0977	.0222	* .0999	.0181	.0041	.0016	.0033
Percent*Items	.0575	.0196	.0324	.0295	.0122	* .0627	* .0847
Percent*Skew	.0005	.0001	.0001	.0001	.0001	.0004	.0005
Form*SS	.0000	.0000	.0000	.0001	.0000	.0000	.0000
Form*Density	.0319	.0017	.0289	.0049	.0049	.0048	.0048
Form*Items	.0097	.0009	.0043	.0015	.0004	.0026	.0040
Form*Skew	.0000	.0000	.0000	.0000	.0000	.0000	.0000
SS*Density	.0001	.0000	.0000	.0000	.0000	.0000	.0000
SS*Items	.0000	.0000	.0000	.0001	.0001	.0000	.0000
SS*Skew	.0000	.0000	.0000	.0000	.0000	.0000	.0000
Density*Items	.0000	.0001	.0073	.0092	.0002	.0007	.0007
Density*Skew	.0002	.0000	.0001	.0001	.0000	.0000	.0000
Items*Skew	.0002	.0001	.0000	.0000	.0000	.0001	.0001
Total Explained	.9406	.9944	.9641	.9771	.9915	.9883	.9875

Note 1. Percent=Percent of Misspecification, Form=Form of Misspecification, Density=Density of Q-Matrix, Items=Number of Items, Skew=Skewness of Person Ability Distribution, SS=Sample Size,

Note 2. \* indicates a medium effect size, \*\* indicates a large effect size

Examining Table 9 results reveals there are large eta squared effect sizes (Cohen, 1998) for main effects for (a) percent of misspecification: CC1 ( $\eta^2 = .2190$ ), CC2 ( $\eta^2 = .8031$ ), : CC3 ( $\eta^2 = .5013$ ), CC4 ( $\eta^2 = .3821$ ), CC5 ( $\eta^2 = .6822$ ), Beta ( $\eta^2 = .7031$ ), and Theta ( $\eta^2 = .6200$ ), (b) Form of Misspecification: CC4 ( $\eta^2 = .2554$ ), and CC5 ( $\eta^2 = .1559$ ), (c)

Density of Q-matrix: CC1 ( $\eta^2 = .1988$ ), and CC3 ( $\eta^2 = .2254$ ), and (d) number of items: Theta ( $\eta^2 = .1839$ ). There are medium eta squared effect sizes (Cohen, 1998) for main effects for: (a) form of misspecification: CC1 ( $\eta^2 = .1319$ ), (b) number of items: CC1 ( $\eta^2 = .1085$ ), and Beta ( $\eta^2 = .1245$ ). There are large eta squared effect sizes (Cohen, 1998) for first order interaction effects for: (a) percent of misspecification by form of misspecification: CC4 ( $\eta^2 = .1879$ ).

There are medium eta squared effect sizes (Cohen, 1998) for first order interaction effects for: (a) percent of misspecification by form of misspecification for CC1 ( $\eta^2 = .0831$ ), and CC5 ( $\eta^2 = .0933$ ), (b) percent of misspecification by density of the Q-matrix for CC1 ( $\eta^2 = .0977$ ), and CC3 ( $\eta^2 = .0999$ ), and (c) percent of misspecification by number of items for beta ( $\eta^2 = .0627$ ), and theta ( $\eta^2 = .0847$ ).

### **Percent of Misspecification**

Percent of misspecification has a large impact on the mean bias values in CC1 ( $\eta^2 = .2190$ ), CC2 ( $\eta^2 = .8031$ ), CC3 ( $\eta^2 = .5013$ ), CC4 ( $\eta^2 = .3821$ ), CC5 ( $\eta^2 = .6822$ ), Beta ( $\eta^2 = .7031$ ), and Theta ( $\eta^2 = .6200$ ). Percent of misspecification accounts for 21.90% of the variance of bias in CC1, 80.31% of the variance of bias in CC2, 50.13% of the variance of bias in CC3, 38.21% of the variance of bias in CC4, 68.22% of the variance of bias in CC5, 70.31% of the variance of bias in beta, and 62% of the variance of bias in theta.

The effect of misspecifying the Q-matrixes causes item difficulty, cognitive components, and person ability parameter estimates to be biased negatively or positively. This effect is graphically represented in Figure 4 which demonstrates that as the percent of misspecification increased mean bias estimates for item difficulty, person ability, and cognitive components increased.

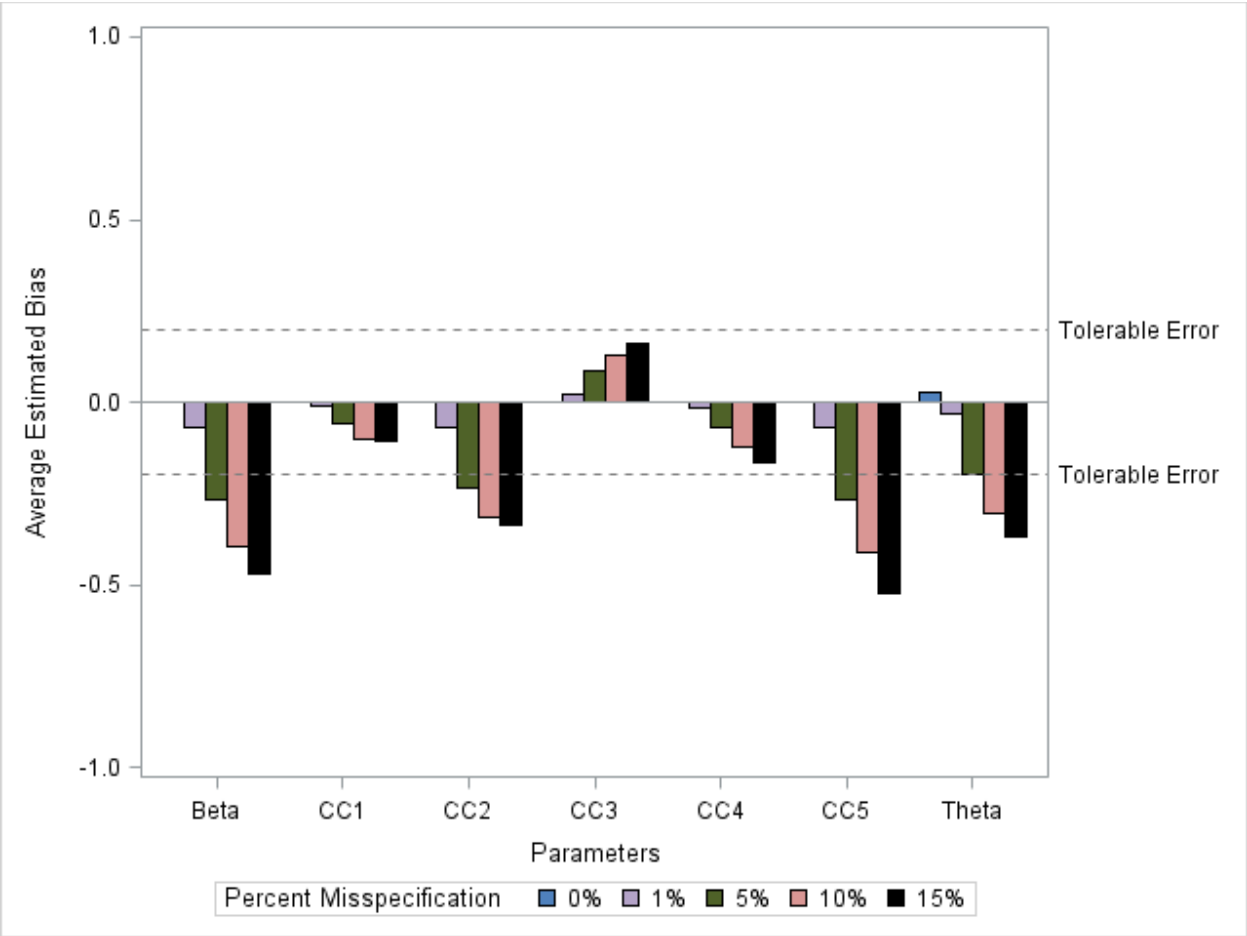


Figure 4. Average estimated bias for beta, CC1-CC5, and theta by percent of misspecification.

It is important to note that average estimated bias increased negatively for beta, CC1, CC2, CC4, CC5, and theta, whereas the average estimated overall bias values increased positively for CC3. Further, item difficulty, and person ability parameter estimates surpassed tolerable error when the Q-matrix was misspecified greater than 5%; however, the cognitive components parameter estimates demonstrated tolerable error for a majority of the cognitive components when the Q-matrix was misspecified 15%.

The parameter estimates for CC1-CC5 were arbitrarily set at .18, .42, .03, .65, and 1.2. The results of this study suggest that positive cognitive components parameter estimates (e.g., CC1, CC2, CC4, and CC5) are likely to be biased negatively when the Q-matrix is misspecified

and cognitive component parameter estimates close to zero in the logit unit (e.g., CC3) are likely to be biased positively when the Q-matrix is misspecified. At zero percent misspecification the average bias is close to zero for the cognitive components and item difficulty; however, for person ability at zero percent misspecification the mean bias values are slightly positive.

To illustrate the effect of misspecification of the Q-matrix, consider that Chen and MacDonald (2011) reported a cognitive component value for *Providing Interpretation* of 1.11, which has a value similar to CC5 in this study. If the Chen and MacDonald Q-matrix were misspecified by 10% as in CC5 above the biased value for *Providing Interpretation* might be 0.70 instead of 1.11. IRT item difficulty parameter estimates in standardized testing, especially in the K-12 environment, are often scaled with a mean of 300 and a standard deviation of 50. If this were the case *Providing Interpretation* would be assigned a biased score of 335.5 instead of their true score of 355.5. As a second example from the MacDonald and Chen (2011) study, consider *Using Illustrations* which was reported to have a cognitive component estimate of 0.07 similar to CC3 in this study. If the Chen and MacDonald Q-matrix were misspecified 15% as in CC3 above the biased value might be 0.25 instead of 0.07. In a scaled version the cognitive component *Using Illustrations* would be assigned a biased score of 325 instead of their true score of 307.

### **Form of Misspecification**

Form of misspecification has a large impact on the estimated mean bias values in CC4 ( $\eta^2 = .2554$ ) and CC5 ( $\eta^2 = .1559$ ). Form of misspecification has a medium effect in CC1 ( $\eta^2 = .1319$ ). Form of misspecification accounts for 25.54% of the variance of bias in CC4, 15.59% of the variance of bias in CC5, and 13.19% of the variance of bias in CC1. Figure 5 demonstrates that under-specification of the Q-matrix tended to yield less bias compared to balanced- and over-misspecification of the Q-matrix for cognitive components 1, 4, and 5. Form



of misspecification did not have not have at least a medium impact on the mean bias estimates in item difficulty, person ability, CC2 or CC3.

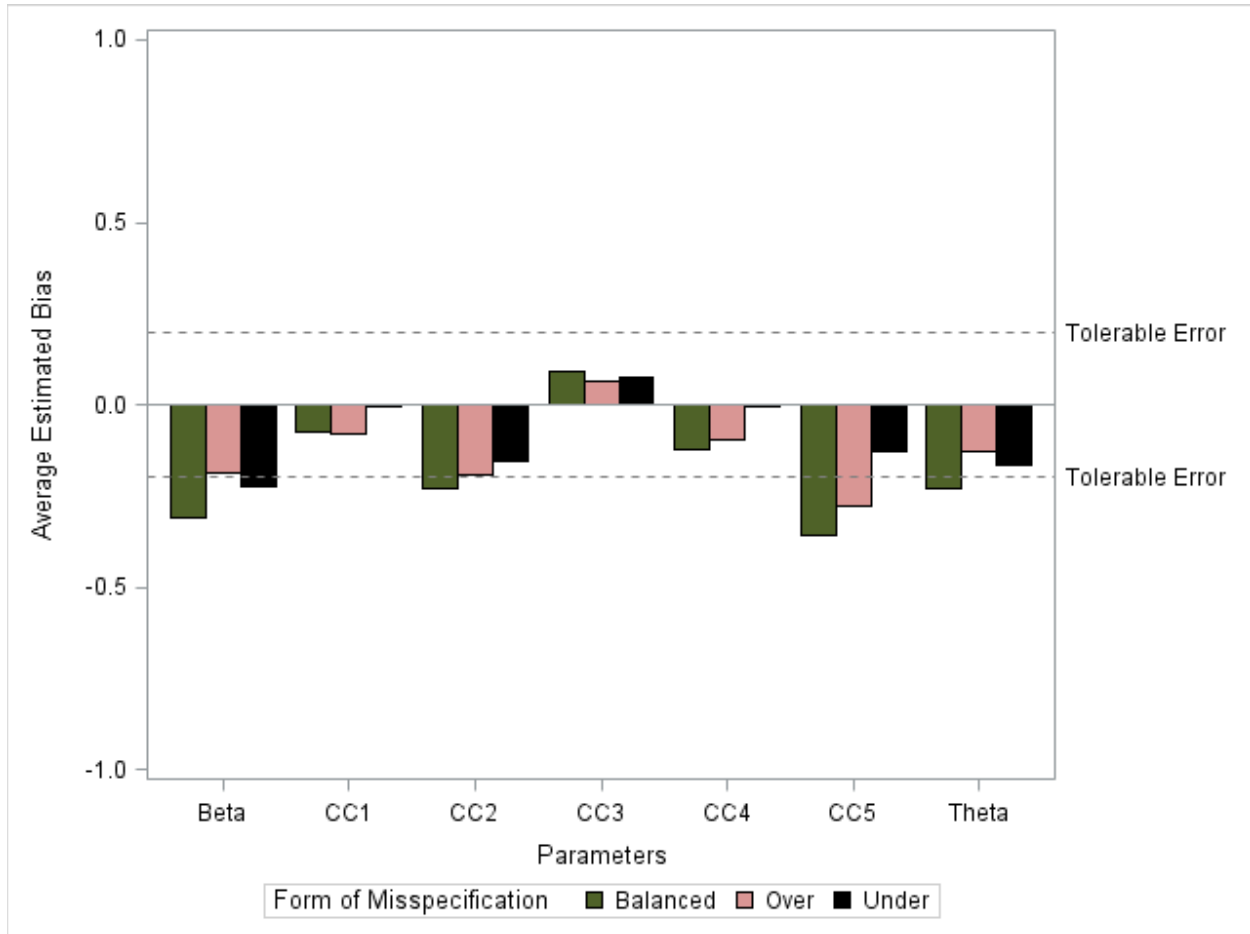


Figure 5. Average estimated bias for beta, CC1-CC5, and theta by form of misspecification.

Note that researchers specifying Q-matrices can indicate an item contains a cognitive component when in fact that item does not contain a cognitive component, or they can mark an item in a Q-matrix as not containing a cognitive component, when in fact that item does contain a cognitive component.

As figure 5 demonstrates the estimated mean bias values for the cognitive components and person ability parameter estimates fell within tolerable error when the Q-matrix was under-specified. When the Q-matrix was over-specified the bias values were generally higher than in the under-specified condition; however, the values for the cognitive components, item difficulty, and person ability fell within tolerable error except in the case of one cognitive component. The bias values were highest when the Q-matrix was balanced misspecified with a majority of the cognitive components bias values falling within tolerable error limits.

### **Percent of Misspecification by Form of Misspecification**

Percent of misspecification by form of misspecification has a large impact on the estimated mean bias values in CC4 ( $\eta^2 = .1879$ ). Percent of misspecification by form of misspecification has a medium impact on the estimated mean bias values in CC1 ( $\eta^2 = .0831$ ), and CC5 ( $\eta^2 = .0933$ ). Percent of misspecification by form of misspecification accounts for 18.79% of the variance of bias in CC4, 8.31% of the variance of bias in CC1, and 9.33% of the variance of bias in CC5.

As Figure 6 graphically demonstrates for CC1, CC4, and CC5, when the percent of misspecification increased, the balanced-misspecified and over-specified Q-matrix dramatically yielded more estimated mean bias, compared to the under-specified Q-matrix. In other words, the impact of percent of misspecification on mean bias values is greater for cognitive components when the Q-matrix is balanced misspecified and over-specified compared to under-specified Q-matrices.

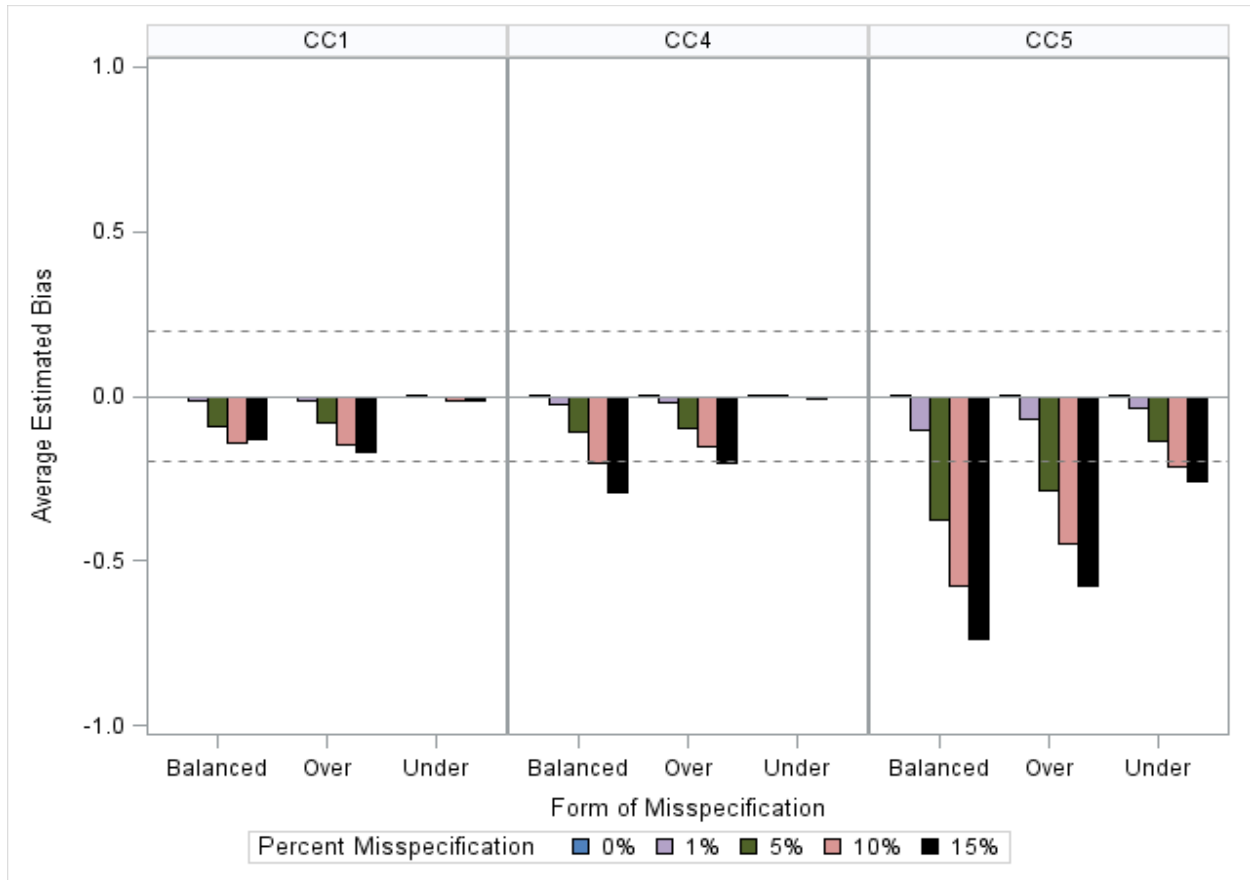


Figure 6. Average estimated bias for CC1, CC4, and CC5 for percent of misspecification by form of misspecification. The reference lines are placed at -0.20 and 0.20 to define the limits of tolerable error.

### Density of the Q-matrix

In this study the dense matrix (see table 4) was specified at 80%, 80%, 60%, 40% and 40% for CC1-CC5 respectively for an overall Q-matrix specification of 60%. The sparse Q-matrix (see table 5) was specified at 60%, 50%, 40%, 40%, and 40% for CC1-CC5 respectively for an overall Q-matrix specification of 46%. Densely or sparsely specifying the Q-matrix had a large impact on the mean bias values in CC1 ( $\eta^2 = .1988$ ), and CC3 ( $\eta^2 = .2254$ ). Densely or sparsely specifying the Q-matrix had no effect or a small effect on the mean bias values in CC5,

beta, theta, CC2, and CC4. Densely or sparsely specifying the Q-matrix accounted for 19.88% of the variance of bias in CC1, and 22.54% of the variance of bias in CC3.

As demonstrated in Figure 7 the sparse Q-matrix yielded less bias in CC1 and CC3 than the dense Q-matrix. For the remaining parameters, the sparse Q-matrix produced slightly larger estimated mean bias values than the dense Q-matrix, but the differences of the estimate mean bias values between the sparse and dense Q-matrix were negligible.

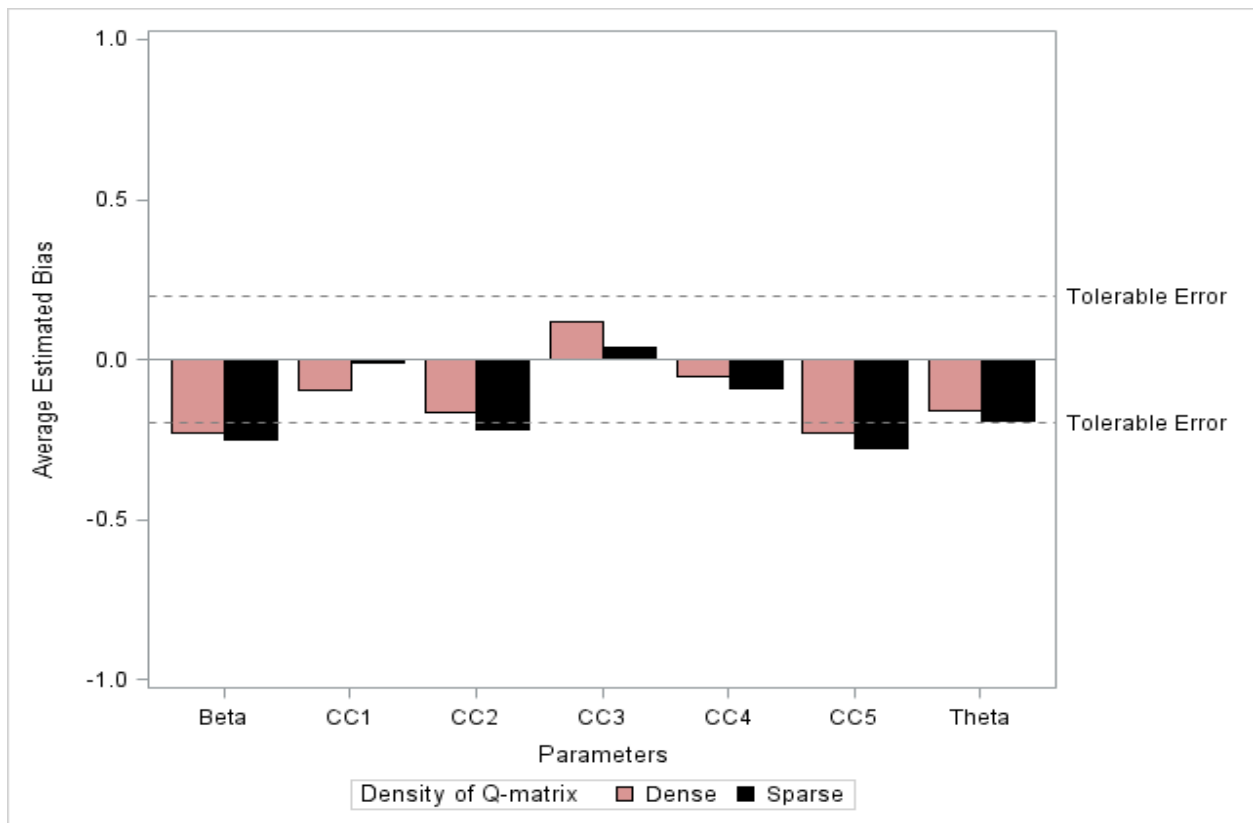


Figure 7. Average estimated bias for beta, CC1-CC5, and theta by density of the Q-matrix.

### Percent of Misspecification by Density of the Q-matrix

Percent of misspecification by density of the Q-matrix has a medium impact on the mean bias values in CC1 ( $\eta^2 = .0977$ ), and CC3 ( $\eta^2 = .0999$ ). Percent of misspecification by density

of the Q-matrix accounts for 9.77% of the variance of bias in CC1, and 9.99% of the variance of bias in CC3. As Figure 8 graphically demonstrates for CC1 and CC3, when the percent of misspecification increased the estimated mean bias values increased dramatically for the dense Q-matrix compared to the sparse Q-matrix, when the percent of misspecification increased. In other words for CC3, percent of misspecification had a greater impact on the estimated mean bias values in the dense Q-matrix than in the sparse Q-matrix.



Figure 8. Average estimated bias for CC1 & CC3 for percent of misspecification by density of the Q-matrix. The reference lines are placed at -0.20 and 0.20 to define the limits of tolerable error.

## Number of Items

The number of items was specified at 20, 40, and 60, respectively. The number of items had a large impact on the estimated mean bias values in theta ( $\eta^2 = .1839$ ), CC1 ( $\eta^2 = .1085$ ), and beta ( $\eta^2 = .1245$ ). The number of items accounted for 18.39% of the variance of bias in theta, 10.85% of the variance of bias in CC1, and 12.45% of the variance of bias in beta. As Figure 9 demonstrates, estimate mean bias values for 20 items for cognitive components, item difficulty, and person ability parameters fall within tolerable limits. As that number increases to

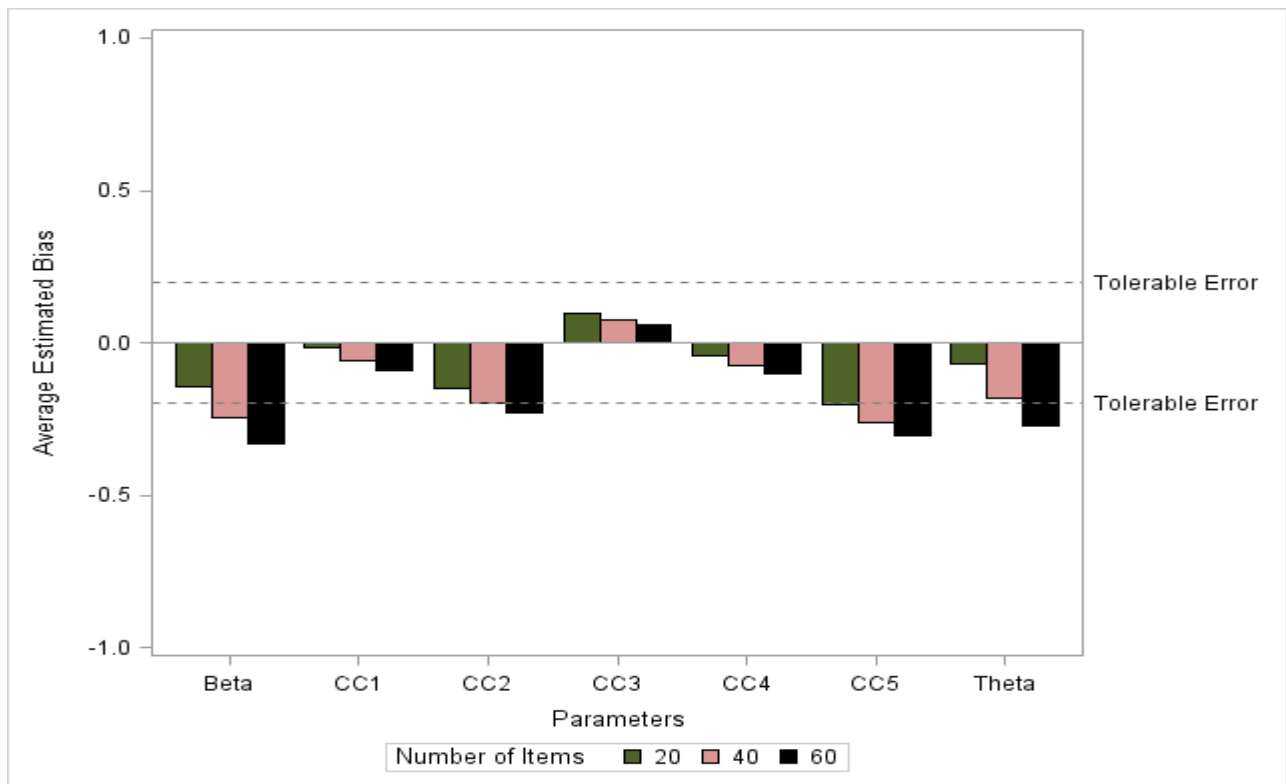


Figure 9. Average estimated bias for beta, CC1-CC5, and theta by number of items.

40 items bias values for a majority of the cognitive components fall within tolerable limits except for beta and CC5. When there are 60 items a minority of cognitive components fall within tolerable limits except for CC2, CC5, beta, and theta. In addition, for CC1, beta, and theta, the estimated mean bias values increased as the number of items increased. The mean bias for CC2, CC4, and CC5 seemed to follow the same pattern, but the difference in estimated mean bias values for these parameters was negligible.

### **Percent of Misspecification by Number of Items**

Percent of misspecification by number of items has a medium impact on the estimated mean bias values in beta ( $\eta^2 = .0627$ ), and theta ( $\eta^2 = .0847$ ). Percent of misspecification by number of items accounts for 6.27% of the variance of bias in beta, and 8.47% of the variance of bias in theta. As Figure 10 graphically demonstrates item difficulty and person ability have similar interaction patterns; namely, that as the percent of misspecification increased the estimated mean bias values for item difficulty (i.e., beta) and person ability (i.e., theta) increased as the number of items increased. In other words, the impact of percent of misspecification on mean bias in person ability and item difficulty is greater for longer tests.

Item difficulty bias estimates for 20 items remained within tolerable limits when the Q-matrix was misspecified up to 5%. In comparison, item difficulty bias estimates for 40 and 60 items remained within tolerable limits when the Q-matrix was misspecified up to 1%. Person ability bias estimates for 20 items remained within tolerable levels when the Q-matrix was misspecified up to 15%. Person ability bias estimates for 40 items remained with tolerable levels when the Q-matrix was misspecified up to 5%, and person ability bias estimates for 60 items remained within tolerable levels when the Q-matrix was misspecified up to 1%.

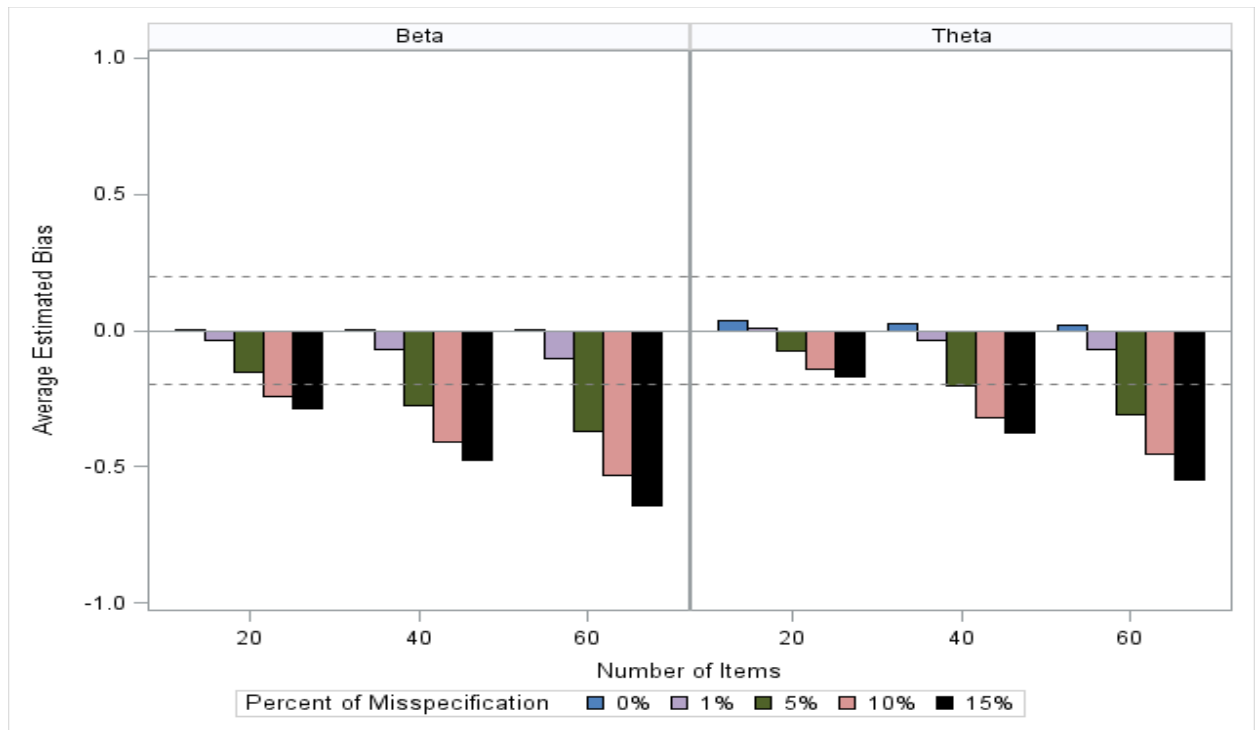


Figure 10. Average estimated bias for beta, and theta for percent of misspecification by number of items. The reference lines are placed greater than -0.20 and less than 0.20 to define the limits of tolerable error.

### Summary for Estimated Mean Bias Results

The summary result of the estimated mean bias can be found in Table 10. Based on the estimate mean bias values, percent of misspecification had the largest practically significant effect on the estimates of cognitive components, item difficulty and person ability. The form of misspecification had a practically significant effect on the estimates of a majority of the cognitive components, but not on parameter estimates of item difficulty or person ability. There was also a first-order interaction effect between percent of misspecification and form of misspecification on a majority of cognitive components.



Table 10

*Summary of Estimated Mean Bias Results*

Design Factor	Impact on mean Bias Estimates for CCs, Item Difficulty & Person Ability by Design Factor
Percent	As the percent of misspecification increased mean bias estimates for item difficulty, person ability, and all cognitive components increased
Form	Under-specification of the Q-matrix tended to yield less bias compared to balanced- and over-specification of the Q-matrix for cognitive components 1, 4, and 5
Percent*Form	As the percent of misspecification increased, the balanced-misspecified and over-specified Q-matrix dramatically yielded more estimated mean bias, compared to the under-specified Q-matrix for cognitive components 1, 4, and 5.
Percent*SS	No practical significant effect on all outcome variables
Form*SS	No practical significant effect on all outcome variables
Percent*Density	As the percent of misspecification increased the estimated mean bias values increased dramatically for the dense Q-matrix for cognitive components 1 and 3 compared to the sparse Q-matrix, when the percent of misspecification increased.
Form*Density	No practical significant effect on all outcome variables
Percent*Items	As the percent of misspecification increased the estimated mean bias values for item difficulty and person ability increased as the number of items increased.
Form*Items	No practical significant effect on all outcome variables
Percent *Shape	No practical significant effect on all outcome variables
Form *Shape*	No practical significant effect on all outcome variables

There were two other practically significant first-order interactions related to Q-matrix misspecification: percent of misspecification with Q-matrix density and percent of misspecification with number of items. Percent of misspecification by Q-matrix density had an interactive effect on only one parameter estimate of the five cognitive components. Percent of

misspecification by number of items had an interactive effect on the estimates of item difficulty and person ability, but not on the estimates of the cognitive components. All other first order interaction effects related to Q-matrix misspecification (i.e., percent\*SS, form\*SS, form\*density, form\*items, percent\*shape, and form\*shape) did not have practically significant effects on all the estimates of cognitive components, item difficulty and person ability.

### **Root Mean Squared Error**

Root mean squared error (RMSE) was used to quantify the typical difference between the true and estimated values of the cognitive components, item difficulty and person ability. The RMSE provides a measure of the variance of the estimated parameter about the mean and its bias. RMSE will help quantify the impact of the design factors on the cognitive components, item difficulty, and person ability. Smaller RMSE values close to 0 suggest the parameter estimates are closer to the truth, and values less than 0.20 of a standard deviation are less than a small effect as defined by Cohen (1988)

The overall distribution of the estimated root mean squared error (RMSE) for CC1 - CC5, beta, and theta values across all simulation conditions is illustrated in box plots in Figure 11. The mean, standard deviation, minimum, and maximum RMSE values for in CC1 - CC5, beta, and theta are presented in Table 11. Closely examining the box plots in Figure 11 and the descriptive statistics in Table 11 demonstrate that average RMSE values in beta, theta, and CC1-CC5 are not close to 0 and have boxes that are not tight. In other words, the average RMSE values for these statistics have large standard deviations. These results warrant further investigation.

Table 11

Mean, Standard Deviation, Minimum, and Maximum values for the Estimated RMSE for CC1 - CC5, Beta, and Theta

	MEAN	SD	MIN	MAX
Beta	0.4079	0.2286	0.0310	0.9483
CC1	0.1684	0.0907	0.0206	0.6363
CC2	0.2788	0.1310	0.0250	0.7101
CC3	0.1790	0.1009	0.0182	0.5279
CC4	0.1908	0.0982	0.0191	0.4648
CC5	0.3242	0.2129	0.0223	0.8508
Theta	0.5100	0.0972	0.3204	0.8073

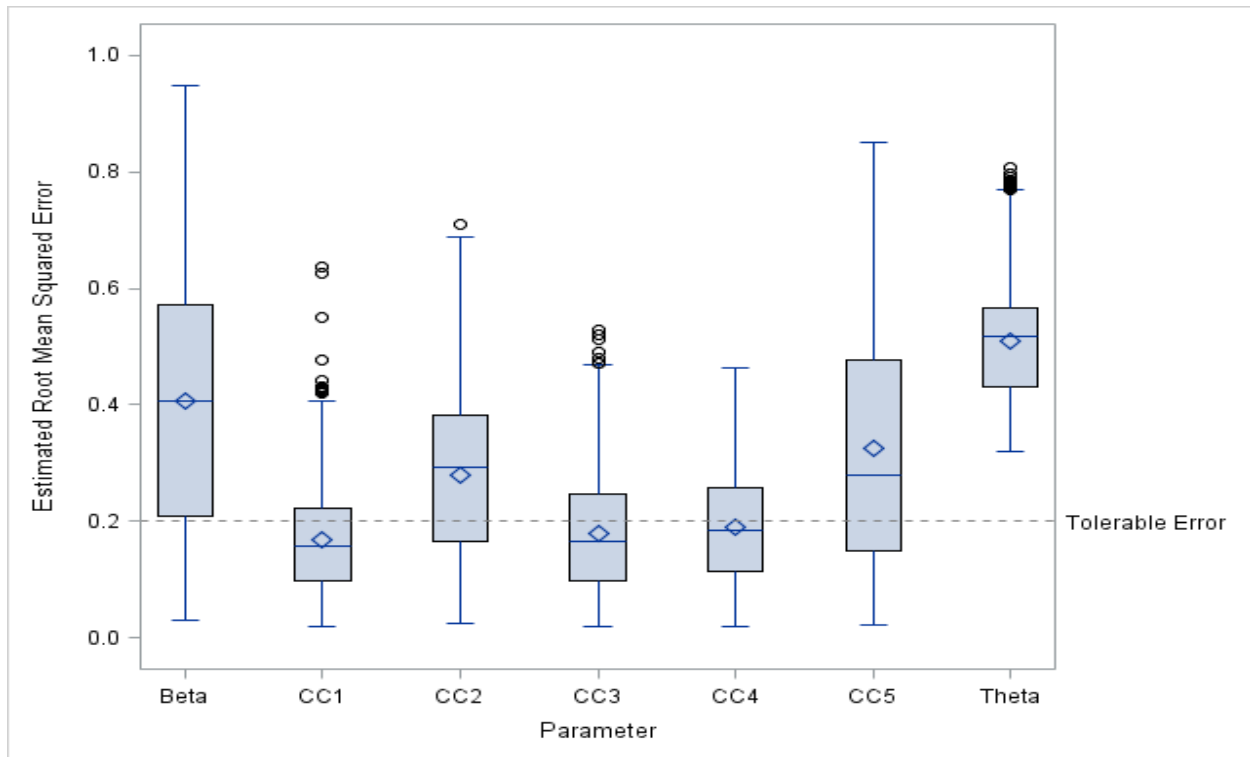


Figure 11. Distribution of estimated RMSE for CC1-CC5, beta, and theta.

An ANOVA for the design factors (i.e., percent of misspecification, form of misspecification, sample size, density of Q-matrix, number of items in Q-matrix, and the skew of the person ability distribution) and their first order interaction effects was computed for item

difficulty, cognitive components, and person ability as measured by the overall estimated RMSE values. Table 12 presents eta-squared values,  $(\eta^2)$ , that quantify the impact the design factors and first order interaction effects have on item difficulty, cognitive components, and theta parameter estimates as measured by RMSE.

Table 12

*Eta-Squared Values for the Association of Design Factors and 1st Level Interaction Effects with the Average Estimated Overall Root Mean Squared Error for CCs, Beta, and Theta*

	CC1	CC2	CC3	CC4	CC5	Beta	Theta
Percent	** .3072	** .6839	** .3028	** .4204	** .6454	** .7621	** .4862
Form	.0364	.0497	.0110	* .0782	** .1555	* .0593	* .0713
Sample Size	** .2672	* .1142	** .2050	** .2238	.0315	.0433	.0247
Density	* .0925	.0080	** .1531	.0231	.0080	.0013	.0001
Items	.0582	.0042	** .1823	.0573	.0051	.0319	* .0764
Skew	.0001	.0001	.0000	.0001	.0000	.0001	.0136
Percent*Form	.0337	.0271	.0069	.0491	* .0996	.0265	.0475
Percent*SS	.0259	.0358	.0220	.0276	.0119	.0135	.0002
Percent*Density	.0450	.0221	.0230	.0144	.0037	.0010	.0049
Percent*Items	.0057	.0177	.0273	.0047	.0167	.0381	** .2035
Percent*Skew	.0001	.0001	.0000	.0000	.0000	.0000	.0023
Form*SS	.0015	.0009	.0009	.0050	.0015	.0002	.0001
Form*Density	.0254	.0020	.0232	.0063	.0037	.0053	.0089
Form*Items	.0074	.0023	.0013	.0065	.0007	.0017	.0175
Form*Skew	.0001	.0000	.0000	.0000	.0000	.0000	.0002
SS*Density	.0002	.0023	.0007	.0050	.0001	.0001	.0003
SS*Items	.0195	.0138	.0119	.0196	.0039	.0031	.0000
SS*Skew	.0003	.0001	.0001	.0000	.0000	.0001	.0003
Density*Items	.0087	.0056	.0001	.0266	.0005	.0005	.0012
Density*Skew	.0000	.0000	.0000	.0000	.0000	.0000	.0001
Items*Skew	.0001	.0000	.0000	.0000	.0000	.0000	.0019
Total Explained	.9352	.9898	.9718	.9678	.9879	.9881	.9610

Note 1. Percent=Percent of Misspecification, Form=Form of Misspecification, Density=Density of Q-Matrix, Items=Number of Items, Skew=Skewness of Person Ability Distribution, SS=Sample Size,

Note 2. \*indicates a medium effect size, \*\* indicates a large effect size

The design factors and their first order interaction total eta squared values are .9352 for CC1, .9898 for CC2, .9718 for CC3, .9678 for CC4, .9879 for CC5, .9881 for beta, and .9610 for theta. In other words, the design factors and their first order interaction effects account for 93.52% of the variance of RMSE in CC1, 98.98% of the variance of RMSE in CC2, 97.18% of

the variance of RMSE in CC3, 96.78% of the variance of RMSE in CC4, 98.79 of the variance of RMSE in CC5, 98.81% of the variance of RMSE in beta, and 96.10% of the variance of RMSE in theta.

Examining Table 12 results reveals there are large eta squared effect sizes (Cohen, 1998) for main effects for: (a) percent of misspecification: CC1 ( $\eta^2 = .3072$ ), CC2 ( $\eta^2 = .6839$ ), : CC3 ( $\eta^2 = .3028$ ), CC4 ( $\eta^2 = .4204$ ), CC5 ( $\eta^2 = .6454$ ), Beta ( $\eta^2 = .7621$ ), and Theta ( $\eta^2 = .4862$ ), (b) Form of Misspecification: CC5 ( $\eta^2 = .1555$ ), (c) sample size: CC1 ( $\eta^2 = .2672$ ), CC3 ( $\eta^2 = .2050$ ), CC4 ( $\eta^2 = .2238$ ), (d) Density of Q-matrix: CC3 ( $\eta^2 = .1531$ ), and (d) number of items: CC3 ( $\eta^2 = .1823$ ). There are medium eta squared effect sizes (Cohen, 1998) for main effects for: (a) form of misspecification: CC4 ( $\eta^2 = .0782$ ), beta ( $\eta^2 = .0593$ ), and theta ( $\eta^2 = .0713$ ), (b) sample size: CC2 ( $\eta^2 = .1142$ ), (c) Density of Q-matrix: CC1 ( $\eta^2 = .0925$ ), (d) number of items: theta ( $\eta^2 = .0764$ ).

There are large eta squared effect sizes (Cohen, 1998) for first order interaction effects for: (a) percent of misspecification by number of items: theta ( $\eta^2 = .2035$ ). There are medium eta squared effect sizes (Cohen, 1998) for first order interaction effects for: (a) percent of misspecification by form of misspecification: CC5 ( $\eta^2 = .0996$ ).

### **Percentage of Misspecification**

Percent of misspecification has a large impact on the mean RMSE values in CC1 ( $\eta^2 = .3072$ ), CC2 ( $\eta^2 = .6839$ ), : CC3 ( $\eta^2 = .3028$ ), CC4 ( $\eta^2 = .4204$ ), CC5 ( $\eta^2 = .6454$ ), Beta ( $\eta^2 = .7621$ ), and Theta ( $\eta^2 = .4862$ ). Percent of misspecification accounts for 30.72% of the variance of RMSE in CC1, 68.39% of the variance of RMSE in CC2, 30.28% of the variance of

RMSE in CC3, 42.04% of the variance of RMSE in CC4, 64.54% of the variance of RMSE in CC5, 76.21% of the variance of RMSE in beta, and 48.62% of the variance of RMSE in theta.

The effect of misspecifying the Q-matrices causes item difficulty, cognitive components, and person ability parameter estimates to have higher mean RMSE values. This effect is graphically represented in Figure 12 which demonstrates that as the percent of misspecification increased mean RMSE values for item difficulty, person ability, and cognitive components increased.

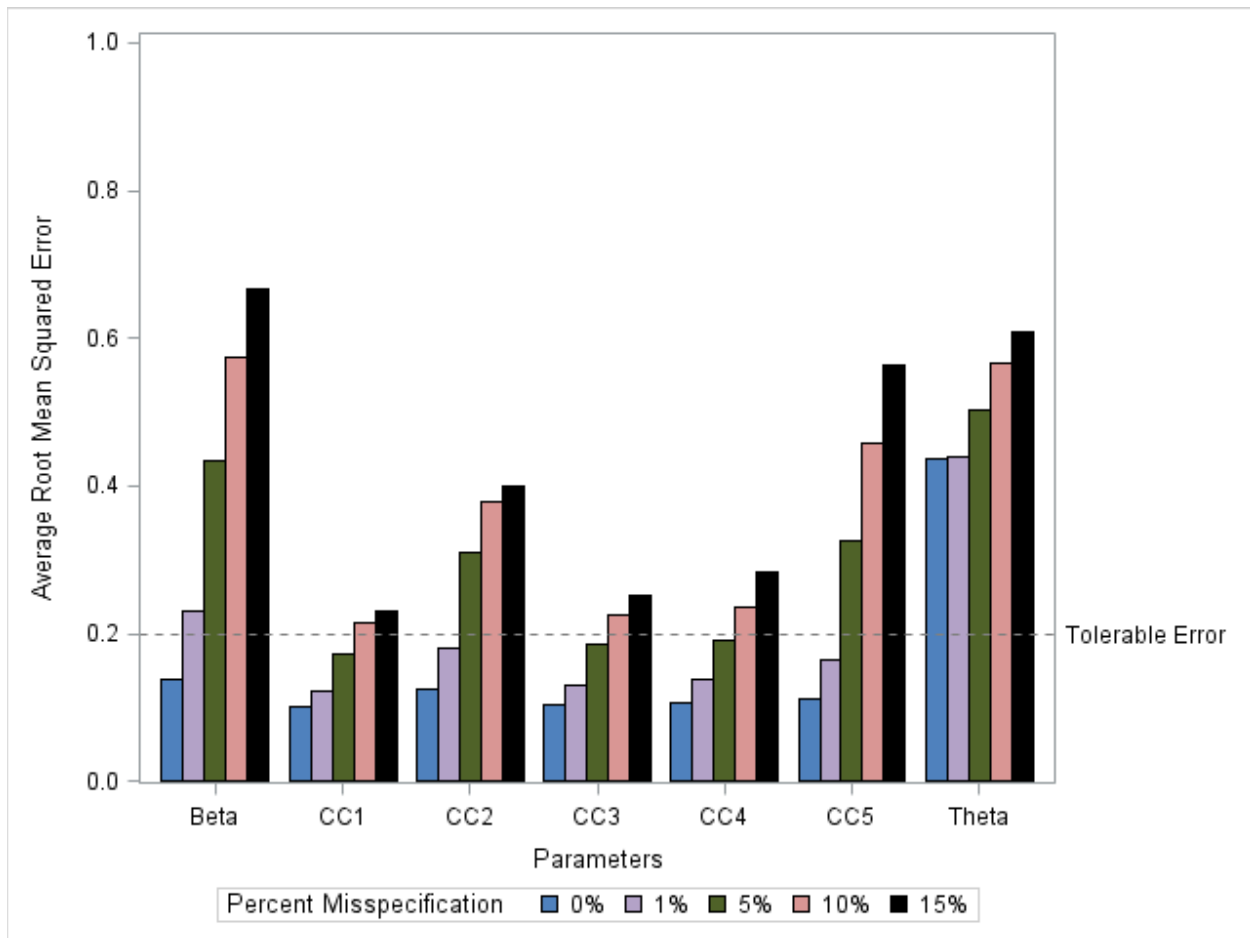


Figure 12. Average estimated root mean squared error for beta, CC1-CC5, and theta by percent of misspecification.

Person ability (i.e., theta) estimates demonstrate higher mean RMSE values. Item difficulty is sensitive to small percentages of misspecification demonstrating RMSE values well above 0.20 when the Q-matrix is misspecified at 5% or above. The cognitive components parameter estimates often demonstrated tolerable error but only when the Q-matrix was misspecified up to 5%.

### **Form of Misspecification**

Form of misspecification has a large effect on the on the estimated mean RMSE values in CC5 ( $\eta^2 = .1555$ ). Form of misspecification had a medium effect on the mean RMSE values in CC4 ( $\eta^2 = .0782$ ), beta ( $\eta^2 = .0593$ ), and theta ( $\eta^2 = .0713$ ). Form of misspecification accounts for 15.55% of the variance of RMSE in CC5, 15.59% of the variance of RMSE in CC5, 7.82% of the variance of RMSE in CC4, 5.93% of the variance of RMSE in beta, and 7.13% of the variance of RMSE in theta. Figure 13 demonstrates that under-specification of the Q-matrix tended to yield smaller RMSE values compared to over-specification of the Q-matrix, and over-specification of the Q-matrix tended to yield smaller RMSE values compared to balanced misspecification of the Q-matrix for cognitive components 4, and 5. Over-specification of the Q-matrix tended to yield smaller RMSE values compared to under-specification of the Q-matrix, and under-specification of the Q-matrix tended to yield smaller RMSE values compared to balanced misspecification for item difficulty and person ability. Form of misspecification did not have not have at least a medium impact on the mean RMSE estimates in CC1, CC2, or CC3.

As Figure 13 demonstrates the estimated mean RMSE values for a majority of the cognitive components parameter estimates fell within tolerable error when the Q-matrix was under-specified. When the Q-matrix was over-specified the mean RMSE values were generally higher than in the under-specified condition; however, a majority of the cognitive components fell within tolerable error. The mean RMSE values were highest when the Q-matrix was balanced misspecified with only a minority of the cognitive components mean RMSE values

falling within tolerable error limits. The estimated mean RMSE values for item difficulty and person ability did not fall within tolerable error for any of the forms of misspecification.

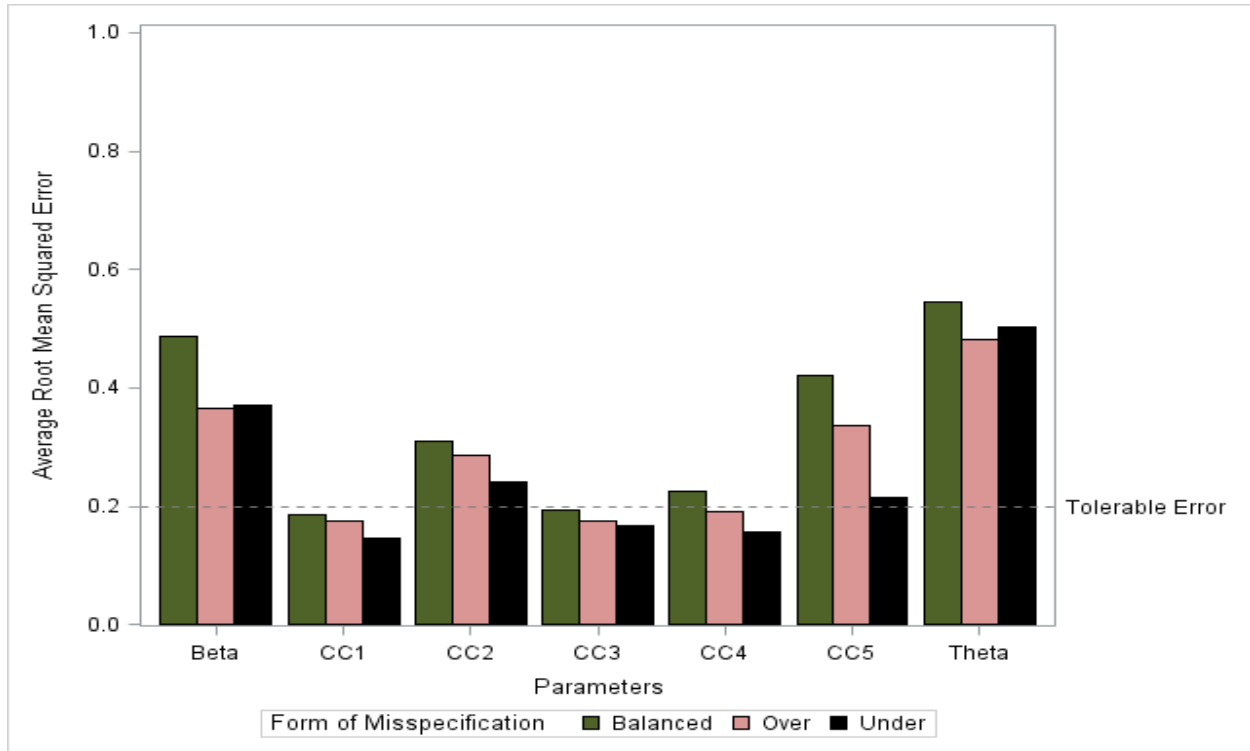


Figure 13. Average estimated root mean squared error for beta, CC1-CC5, and theta by form of misspecification.

### Percent of Misspecification by Form of Misspecification

Percent of misspecification by form of misspecification has a medium impact on the mean RMSE values in CC5 ( $\eta^2 = .0996$ ). Percent of misspecification by form of misspecification accounts for 9.96% of the variance of mean RMSE values in CC5.

As figure 14 graphically demonstrates, for CC5 as the percent of misspecification increased the balanced-misspecified Q-matrix yielded higher estimated mean RMSE values compared to the over-specified Q-matrix. When the percent of misspecification increased the



over-specified Q-matrix yielded higher estimated mean RMSE compared to the balanced-misspecified Q-matrix. In other words, the impact of percent of misspecification on estimated mean RMSE values is greater for cognitive component 5 when the Q-matrix is balanced-misspecified compared to when the Q-matrix is over-specified; further, the impact of percent of misspecification on estimated mean RMSE values is greater when the Q-matrix is over-specified compared to when the Q-matrix is under-specified. CC5 demonstrated tolerable levels of RMSE when the Q-matrix was balanced misspecified, over-specified, or under-specified up to 1%.

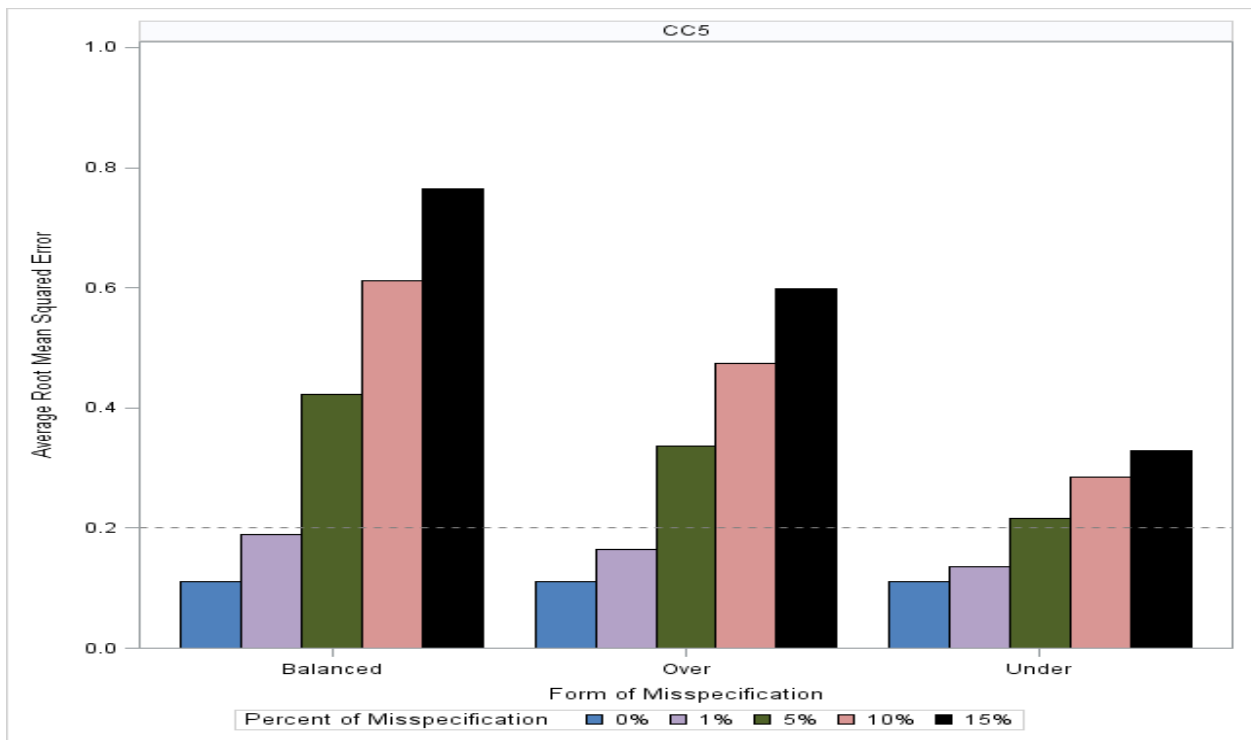


Figure 14. Average estimated root mean squared error for CC5 for percent of misspecification by form of misspecification. The reference line is placed at 0.20 to define the limits of tolerable error.

## Sample Size

Sample size has a large impact on the estimated mean RMSE values in CC1 ( $\eta^2 = .2672$ ), CC3 ( $\eta^2 = .2050$ ), and CC4 ( $\eta^2 = .2238$ ). Sample size has a medium impact on the mean RMSE values in CC2 ( $\eta^2 = .1142$ ). Sample size accounts for 26.72% of the variance of RMSE in CC1, 20.50% of the variance of RMSE in CC3, 22.38% of the variance of RMSE in CC4, and 11.42% of the variance of RMSE in CC2.

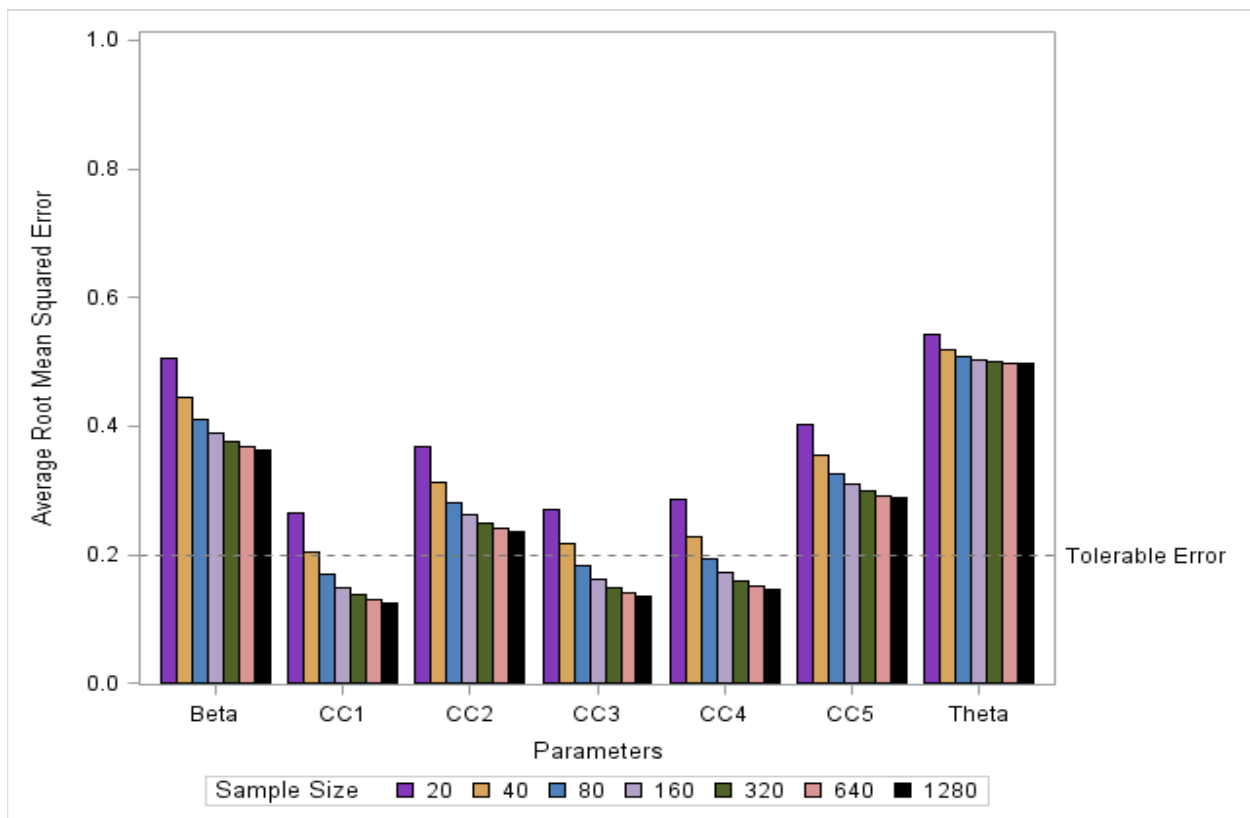


Figure 15. Average estimated root mean squared error for beta, CC1-CC5, and theta by sample size.

As the sample size increased item difficulty, person ability, and cognitive components estimated mean RMSE values decreased. This effect is graphically represented in Figure 15. In

other words, as the size of the sample increased the mean estimated RMSE values became more tolerable. While RMSE values decreased as the sample size increased, item difficulty, and person ability parameter estimates did not achieve tolerable RMSE values (i.e., 0.20) even when the sample size was 1,280. In a majority of the cases the cognitive components the mean estimated RMSE values fell within tolerable error limits with a sample size of approximately 80 or greater. Researchers are warned even with larger sample sizes that cognitive components RMSE values do not always attain tolerable error levels while item difficulty and person ability RMSE error values never fall within tolerable limits.

### **Density of the Q-matrix**

In this study the dense matrix (see table 4) was specified at 80%, 80%, 60%, 40% and 40% for CC1-CC5 respectively for an overall Q-matrix specification of 60%. The sparse Q-matrix (see table 5) was specified at 60%, 50%, 40%, 40%, and 40% for CC1-CC5 respectively for an overall Q-matrix specification of 46%. Densely or sparsely specifying the Q-matrix had a large impact on the mean RMSE values in CC3 ( $\eta^2 = .1531$ ). Densely or sparsely specifying the Q-matrix had a medium impact on the mean RMSE values in CC1 ( $\eta^2 = .0925$ ). Densely or sparsely specifying the Q-matrix had no effect on the mean RMSE values in CC2, CC4, CC5, beta or theta. Densely or sparsely specifying the Q-matrix accounted for 15.31% of the variance of RMSE in CC3, and 9.25% of the variance of RMSE in CC1.

As demonstrated in Figure 16 the sparse Q-matrix yielded smaller estimated mean RMSE values in CC1 and CC3 than the dense Q-matrix. For the remaining parameters (i.e., CC2, CC4, CC5, beta, and theta), the differences in the estimated mean RMSE values between the sparse and dense Q-matrix were not practically significant. The sparse Q-matrix produced slightly larger the RMSE values for sparsely specified Q-matrices fell within tolerable error limits for a majority of the cognitive components. The RMSE values for densely specified Q-matrices are at or slightly above tolerable error limits for a majority of cognitive components. When

considering RMSE values the sparsely specified Q-matrix is preferred over the densely specified Q-matrix for cognitive components parameter estimates. The RMSE values of item difficulty and person ability for densely and sparsely specified Q-matrices do not fall within tolerable error limits.

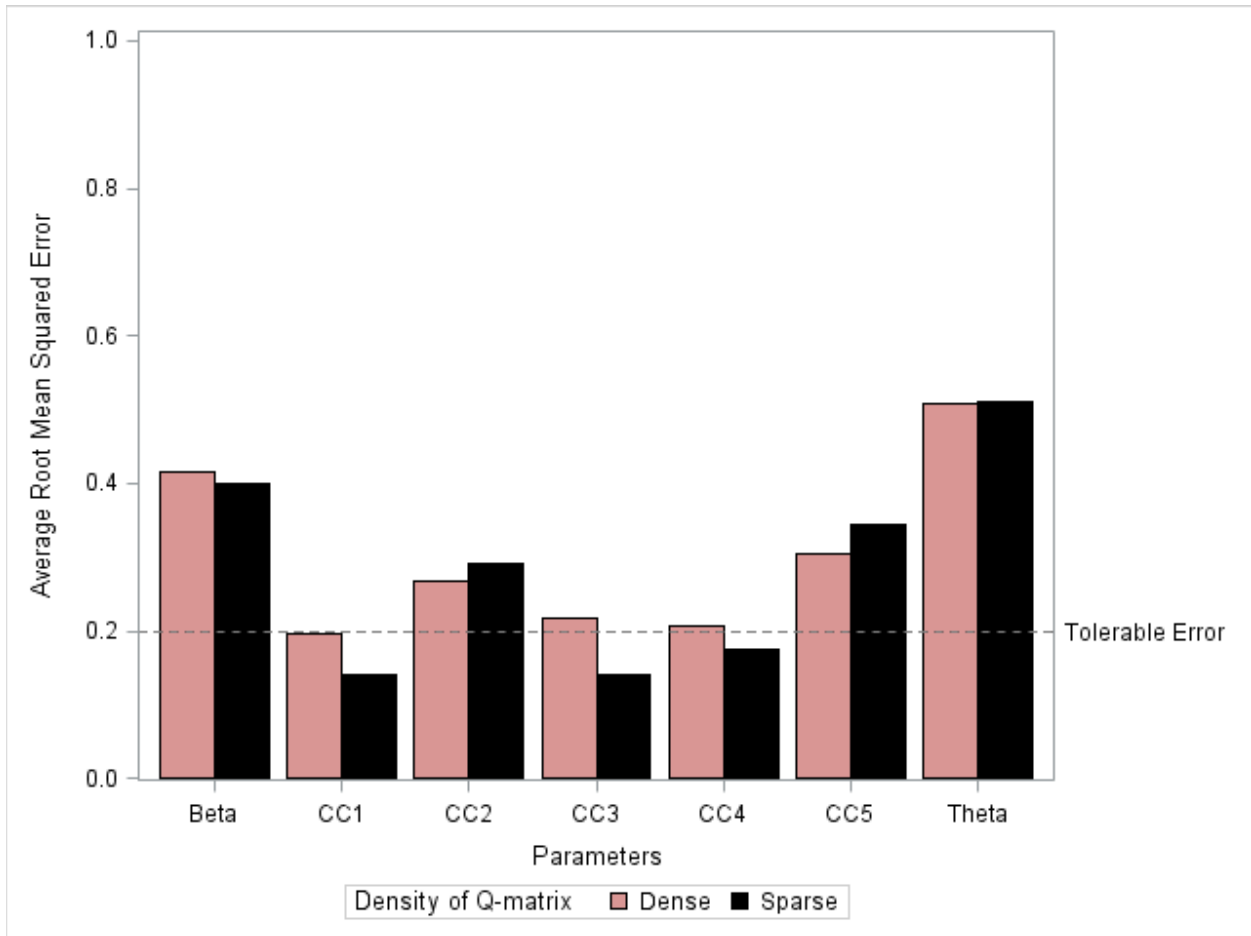


Figure 16. Average estimated root mean squared error for beta, CC1-CC5, and theta by density of the Q-matrix.

### Number of Items

The number of items was specified at 20, 40, and 60 respectively. The number of items had a large impact on the estimated mean RMSE values in CC3 ( $\eta^2 = .1823$ ). The number of

items had a medium impact on the mean RMSE values in theta ( $\eta^2 = .0764$ ). The number of items accounted for 18.23% of the variance of RMSE in CC3, and 7.64% of the variance of RMSE in theta.

As Figure 17 demonstrates when there are 40 or fewer items the mean estimated RMSE values for a majority of cognitive components parameter estimates fall within tolerable error limits; however item difficulty and person ability parameter estimates never fall within tolerable error limits. In other words, the mean estimated RMSE values for cognitive component fall within tolerable error limits when the number of items is 40 or fewer. Item difficulty and person ability parameter estimates never fell within tolerable error limits.

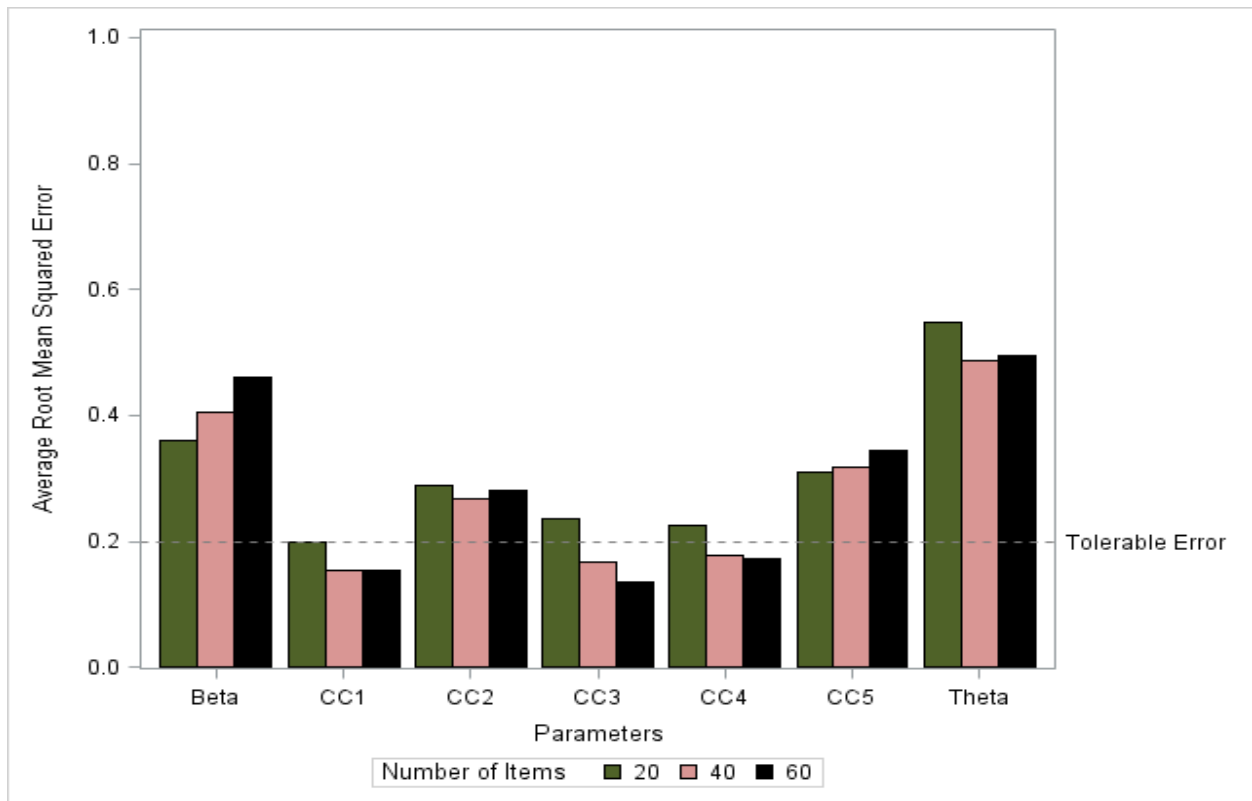


Figure 17. Average estimated root mean squared error beta, CC1-CC5, and theta by number of items.

### Percent of Misspecification by Number of Items

Percent of misspecification by number of items has a medium impact on the mean RMSE values in theta ( $\eta^2 = .2035$ ). Percent of misspecification by number of items accounts for 20.35% of the variance of RMSE in theta. As Figure 18 graphically demonstrates, as the percent of misspecification increased up to 1% the estimated mean RMSE values for person ability decreased as the number of items increased; however, as the percent of misspecification increased from 5% through 15% the estimated mean RMSE values for person ability increased

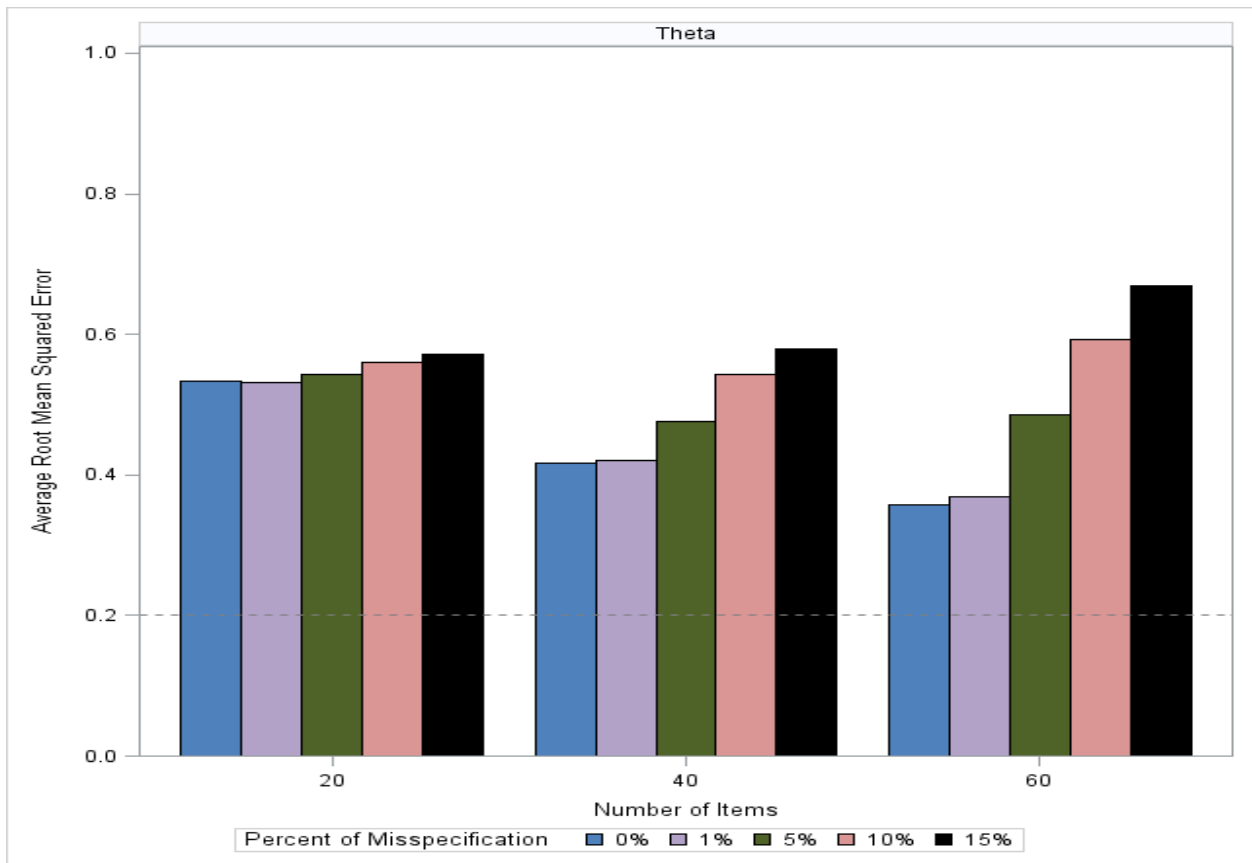


Figure 18. Average root mean squared error for theta for percent of misspecification by number of items. The reference line is placed at 0.20 to define the limits of tolerable error.

as the number of items increased. In other words, the impact of percent of misspecification on mean RMSE values for person ability decreases as the number of items increases when the percentage of misspecification is small; however, when the percent of misspecification is larger the impact on mean RMSE values for person ability increases as the number of items increases. It is worth noting that, person ability mean estimated RMSE values were very large and never came close to falling within tolerable error.

### **Summary for Mean Estimated Root Mean Squared Error**

The summary results of the estimated mean RMSE can be found in Table 13. Based on the estimate mean RMSE values percent of misspecification had the largest practically significant effect on the estimates of cognitive components, item difficulty and person ability. The form of misspecification had a practically significant effect on the estimates of a minority of the cognitive components but not on item difficulty or person ability. There was also a first-order interaction effect between percent of misspecification and form of misspecification on a cognitive component.

There was one other practically significant first-order interaction related to Q-matrix misspecification, namely, percent of misspecification and number of items. Percent of misspecification by number of items had an interactive effect on the estimates of person ability but not on the estimates of cognitive components or item difficulty. All other first order interaction effects related to the Q-matrix misspecification (i.e., percent\*SS, form\*SS, percent\*density, form\*density, form\*items, percent\*shape, and form\*shape) did not have practically significant effects on all the estimates of cognitive components, item difficulty or person ability.

Table 13

*Summary of Estimated Mean RMSE Results*

Design Factor	Impact on mean RMSE Estimates for CCs, Item Difficulty & Person Ability by Design Factor
Percent	As the percent of misspecification increased mean RMSE estimates for item difficulty, person ability, and all cognitive components increased.
Form	Under-specification yielded smaller RMSE values compared to over-specification, and over-specification yielded smaller RMSE values compared to balanced-misspecification of the Q-matrix for cognitive components 4, and 5. Over-specification and under-specification tended to yield smaller RMSE values compared to balanced misspecification of the Q-matrix for item difficulty and person ability; however these parameter estimates did not fall with tolerable error limits.
Percent*Form	As the percent of misspecification increased the balanced-misspecified Q-matrix yielded larger estimated mean RMSE values compared to the over-specified Q-matrix for cognitive component 5. As the percent of misspecification increased the over-specified Q-matrix yielded larger estimated mean RMSE values compared to the balanced-misspecified Q-matrix for cognitive component 5.
Percent*SS	No practically significant effect on any outcome variable.
Form*SS	No practically significant effect on any outcome variable.
Percent*Density	No practically significant effect on any outcome variable.
Form*Density	No practically significant effect on any outcome variable.
Percent*Items	As the percent of misspecification increased up to 1% the estimated mean RMSE values for person ability decreased as the number of items increased; however, as the percent of misspecification increased from 5% through 15% the estimated mean RMSE values for person ability increased as the number of items increased.
Form*Items	No practically significant effect on any outcome variable.
Percent *Shape	No practically significant effect on any outcome variable.
Form *Shape*	No practically significant effect on any outcome variable.



### Estimated Confidence Interval Coverage

CI coverage indicates the degree to which the parameter estimate is accurate. Note SAS employs the commonly used Wald-type confidence interval. SAS constructs CIs around cognitive components and person ability parameter estimates but does not construct CIs around item difficulty parameter estimates. CI coverage is calculated as the percentage of time the true parameter value falls within the confidence interval of the estimated parameter. When the truth parameter falls within the estimated parameters confidence interval the estimated parameter is understood to be accurate. For example CC1s value was set at .18, the lower CI for one estimated cognitive component was -0.3246 and the upper CI was 0.8540 which meant the truth parameter was located within the CI. In another case the lower CI was -0.9681 and the upper CI was 0.0464 which meant the truth parameter was not located within the estimated parameters CI. In the former case the estimated parameter is considered accurate but in the latter it is not. The nominal coverage probability was set at .95 for this study. Researchers want the actual coverage probability to be as close to .95 as possible.

The overall distribution of the estimated CI coverage for CC1 - CC5, and theta values across all simulation conditions is illustrated in box plots in Figure 19. The mean, standard deviation, minimum, and maximum values for the estimated CI coverage rates for CC1-CC5, and theta are presented in Table 14. Closely examining the box plots in Figure 19 and the descriptive statistics in Table 14 demonstrate that person ability parameter estimates fall within the confidence interval 87% of the time with a .12 standard deviation. The cognitive components (e.g., CC5) fall within the confidence interval approximately 47% of the time with a large standard deviation of .39.

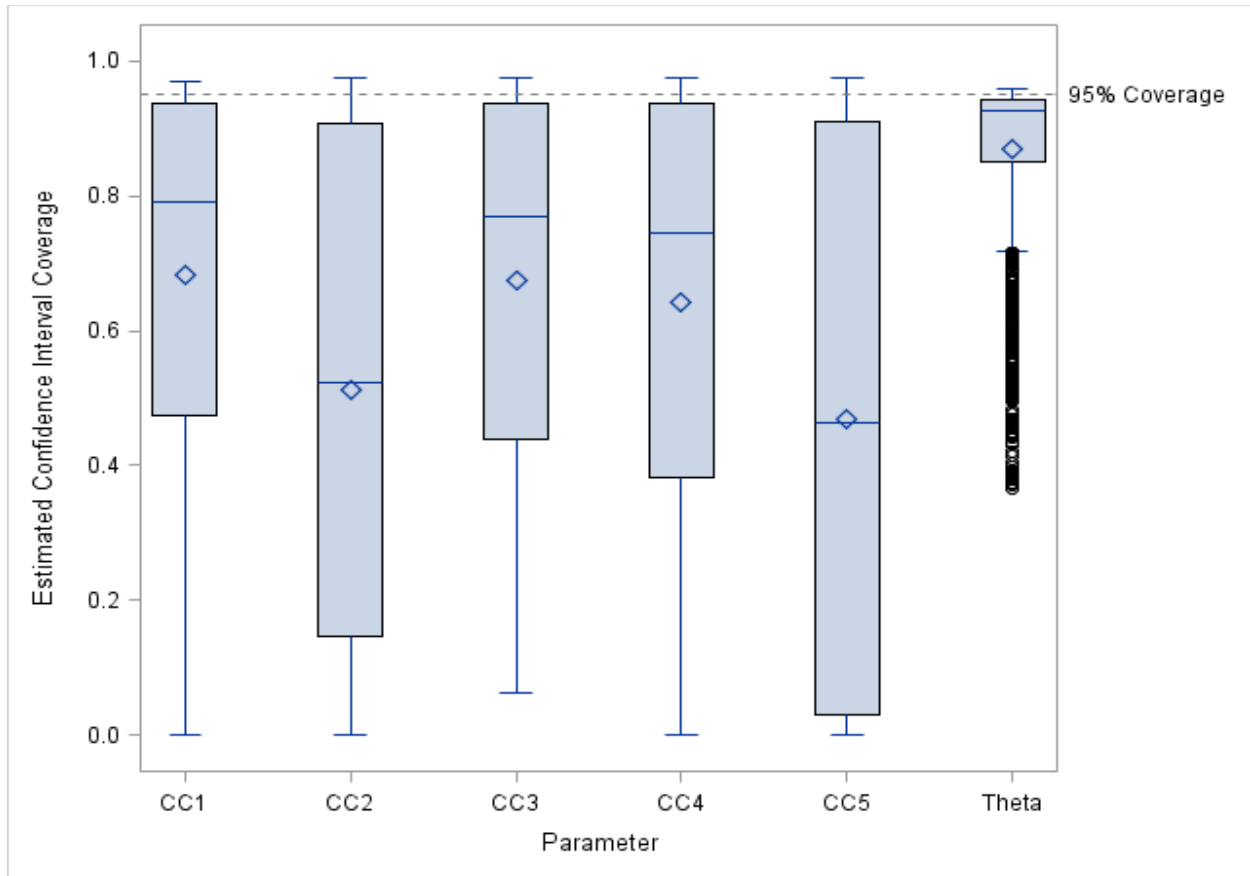


Figure 19. Distribution of estimated confidence interval coverage for CC1-CC5 and theta.

Table 14

Mean, Standard Deviation, Minimum, and Maximum Values for the Estimated Confidence Interval Coverage for CC1-CC5 and Theta

	MEAN	SD	MIN	MAX
CC1	0.6831	0.2853	0.0000	0.9714
CC2	0.5118	0.3606	0.0000	0.9742
CC3	0.6755	0.2785	0.0627	0.9760
CC4	0.6434	0.3095	0.0000	0.9760
CC5	0.4692	0.3892	0.0000	0.9751
Theta	0.8689	0.1202	0.3650	0.9584

An ANOVA for the design factors and their first order interaction effects was computed for cognitive components, and person ability as measured by the mean estimated CI coverage.

Table 15 presents eta-squared values ( $\eta^2$ ) for the association of design factors with estimated CI coverage rates for CC1-CC5, and theta.

Table 15

*Eta-Squared Values for the Association of Design Factors and 1st Level Interaction Effects with the Average Estimated Overall Confidence Interval Coverage for CCs and Theta*

	CC1	CC2	CC3	CC4	CC5	Theta
Percent	** .3963	** .6068	** .4297	** .4365	** .6434	** .3397
Form	.0244	.0153	.0139	* .0710	* .0713	.0511
Sample Size	** .3106	** .1866	** .3326	** .2739	* .1077	.0119
Density	.0350	.0149	.0458	.0038	.0048	.0053
Items	.0189	* .0588	.0002	.0304	.0414	** .2431
Skew	.0001	.0001	.0001	.0000	.0001	.0015
Percent*Form	.0132	.0061	.0062	.0263	.0240	.0353
Percent*SS	* .1011	.0548	* .1053	* .0827	.0353	.0055
Percent*Density	.0204	.0058	.0229	.0080	.0017	.0044
Percent*Items	.0097	.0205	.0008	.0144	.0129	.2151
Percent*Skew	.0001	.0000	.0000	.0000	.0000	.0002
Form*SS	.0019	.0007	.0008	.0043	.0060	.0006
Form*Density	.0114	.0008	.0061	.0003	.0055	.0081
Form*Items	.0069	.0002	.0007	.0061	.0022	.0223
Form*Skew	.0001	.0000	.0000	.0000	.0000	.0000
SS*Density	.0026	.0010	.0036	.0004	.0002	.0000
SS*Items	.0010	.0031	.0002	.0013	.0016	.0054
SS*Skew	.0000	.0000	.0000	.0000	.0000	.0000
Density*Items	.0065	.0003	.0099	.0066	.0001	.0016
Density*Skew	.0001	.0000	.0000	.0000	.0000	.0000
Items*Skew	.0000	.0000	.0000	.0000	.0000	.0001
Total Explained	.9604	.9759	.9787	.9661	.9582	.9511

Note 1. Percent=Percent of Misspecification, Form=Form of Misspecification, Density=Density of Q-Matrix, Items=Number of Items, Skew=Skewness of Person Ability Distribution, SS=Sample Size,

Note 2. \* indicates a medium effect size, \*\* indicates a large effect size

The design factors and their first order interaction total eta squared values are .9604 for CC1, .9759 for CC2, .9718 for CC3, .9787 for CC4, .9661 for CC5, and .9511 for theta. In other words, the design factors and their first order interaction effects account for 96.04% of the variance in the mean CI Coverage rates in CC1, 97.18% of the variance in the mean CI Coverage rates in CC2, 97.87% of the variance in the mean CI Coverage rates in CC3, 96.61%

of the variance in the mean CI Coverage rates in CC4, 95.82% of the variance in the mean CI Coverage rates in CC5, and 95.11% of the variance in the mean CI Coverage rates in theta.

Examining Table 15 results reveals there are large eta squared effect sizes (Cohen, 1998) for main effects for: (a) percent of misspecification: CC1 ( $\eta^2 = .3963$ ), CC2 ( $\eta^2 = .6082$ ), CC3 ( $\eta^2 = .4297$ ), CC4 ( $\eta^2 = .4365$ ), CC5 ( $\eta^2 = .6434$ ), Theta ( $\eta^2 = .3397$ ), (b) sample size: CC1 ( $\eta^2 = .3106$ ), CC2 ( $\eta^2 = .1866$ ), CC3 ( $\eta^2 = .3326$ ), CC4 ( $\eta^2 = .2739$ ), and (c) number of items: theta ( $\eta^2 = .2431$ ). There are medium eta squared effect sizes (Cohen, 1998) for main effects for: (a) form of misspecification: CC4 ( $\eta^2 = .0710$ ), CC5 ( $\eta^2 = .0713$ ), (b) sample size: CC5 ( $\eta^2 = .1077$ ), and (c) number of items: CC2 ( $\eta^2 = .0588$ ).

There are medium eta squared effect sizes (Cohen, 1998) for first order interaction effects for: (a) percent of misspecification by sample size: CC1 ( $\eta^2 = .1011$ ), CC3 ( $\eta^2 = .1053$ ), and CC4 ( $\eta^2 = .0827$ ).

### **Percentage of Misspecification**

Percent of misspecification had a large impact on the mean CI Coverage rates in CC1 ( $\eta^2 = .3963$ ), CC2 ( $\eta^2 = .6068$ ), : CC3 ( $\eta^2 = .4297$ ), CC4 ( $\eta^2 = .4365$ ), CC5 ( $\eta^2 = .6434$ ), and Theta ( $\eta^2 = .3397$ ). Percent of misspecification of the Q-matrix accounts for 39.63% of the variance of the mean CI Coverage rates in CC1, 60.68% of the variance of the mean CI Coverage rates in CC2, 42.97% of the variance of the mean CI Coverage rates in CC3, 43.65% of the variance of the mean CI Coverage rates in CC4, 64.34% of the variance of the mean CI Coverage rates in CC5, and 43.65% of the variance of the mean CI Coverage rates in theta. Figure 20 demonstrates that as the percent of misspecification of the Q-matrix increased the

percentage of cognitive components, and person ability parameter estimates falling within the confidence interval dramatically decreased.

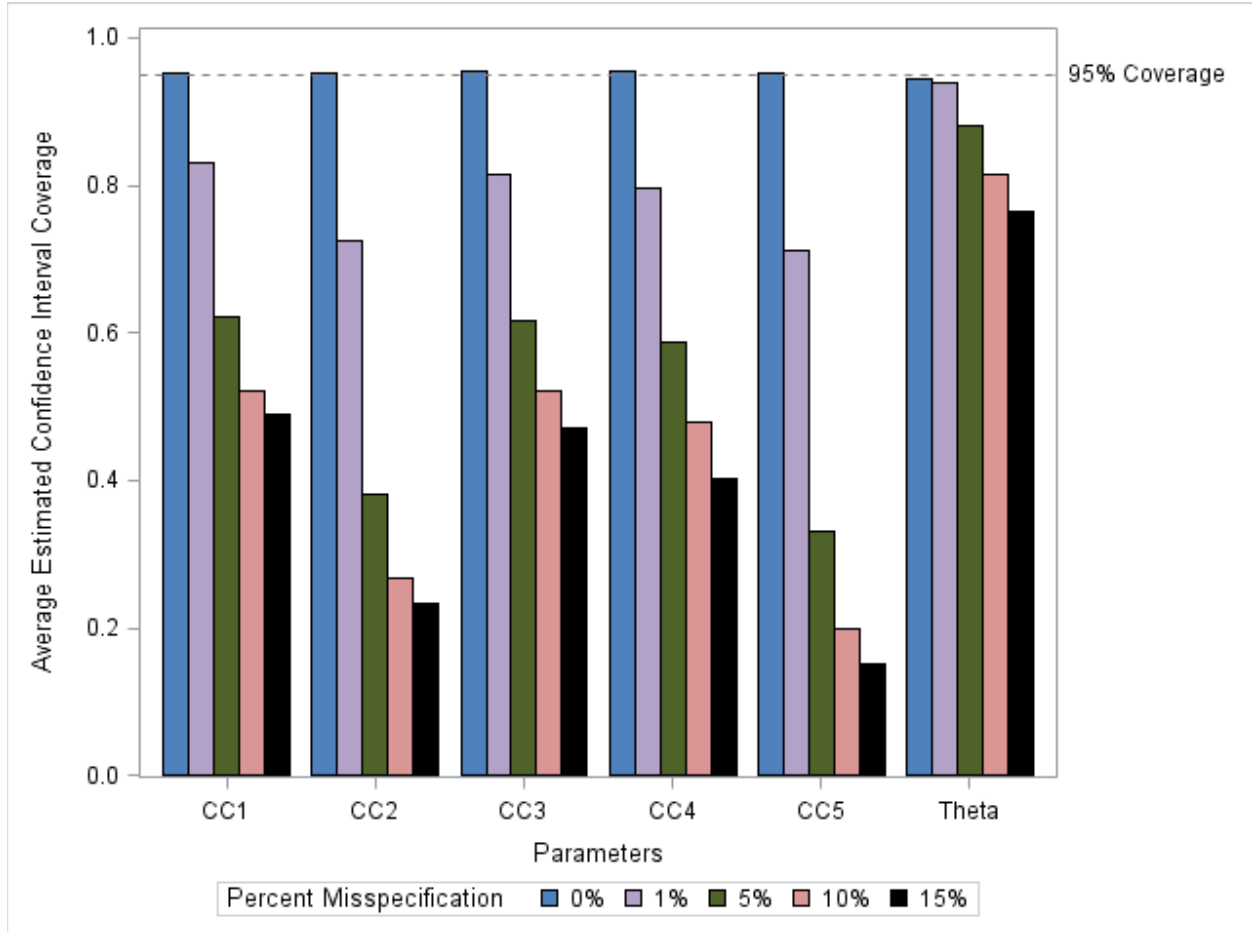


Figure 20. Average estimated confidence interval coverage for CC1-CC5 and theta by percent of misspecification.

Cognitive components and person ability parameter estimates can be thought of as accurate in the truth condition however as the percent of misspecification increased the permissiveness of the CI Coverage rates decreased dramatically. These CI Coverage rates are so permissive they can only be described as dreadful.

The truth parameter is behaving as expected, in that, the truth parameter falls within the confidence interval of the estimated parameters at approximately the 95% nominal rate; however, the CI coverage rate decreases dramatically as the percent of misspecification increases. To understand this result it is important to recall that the cognitive components truth parameters are fixed and that as percent of misspecification increases cognitive component parameter estimates have increasing mean bias. The Wald-type confidence intervals are constructed as a lower and upper bound around the estimated cognitive component but due to increasing mean bias values the lower and upper bounds will begin to increasingly move away from the fixed truth parameters as percent of misspecification increases. As the lower and upper bounds move away from the truth parameter that truth parameter will fall within the estimated confidence interval less and less.

To give a practical example, consider a sparse Q-matrix with 60 items and 1,280 people which is balanced misspecified at 15%. In this study the bias for CC5, whose fixed parameter estimate is 1.2, was -0.82 and the CI Width was 0.071. The lower bound on average would be  $0.309 (1.2_{\text{truth}} - 0.82_{\text{bias}} - 0.071_{\text{lower}})$  and the upper bound on average would be  $0.451 (1.2_{\text{truth}} + 0.82_{\text{bias}} + 0.071_{\text{upper}})$ . Given the large bias and tight CI Width it is unlikely the truth parameter will fall within the CI interval. This will have the effect of dramatically reducing the Wald-type confidence interval coverage rate even when it functions properly.

### **Form of Misspecification**

Form of misspecification had a medium impact of the mean CI Coverage rates in CC4 ( $\eta^2 = .0710$ ), and CC5 ( $\eta^2 = .0713$ ). Form of misspecification accounted for 7.10% of the variance of the mean CI Coverage rates in CC4, and 7.13% of the variance of the mean CI Coverage rates in CC5. Figure 24 demonstrates that under-specification of the Q-matrix tended to yield CI Coverage rates closer to the nominal coverage rate of .95 compared to over-specification for cognitive components 4, and 5. Over-specification yielded CI Coverage rates

closer to the nominal coverage rate of .95 compared to balanced misspecification for cognitive components 4, and 5. Form of misspecification did not have at least a medium impact on the mean CI Coverage rates in person ability, CC1, CC2, or CC3.

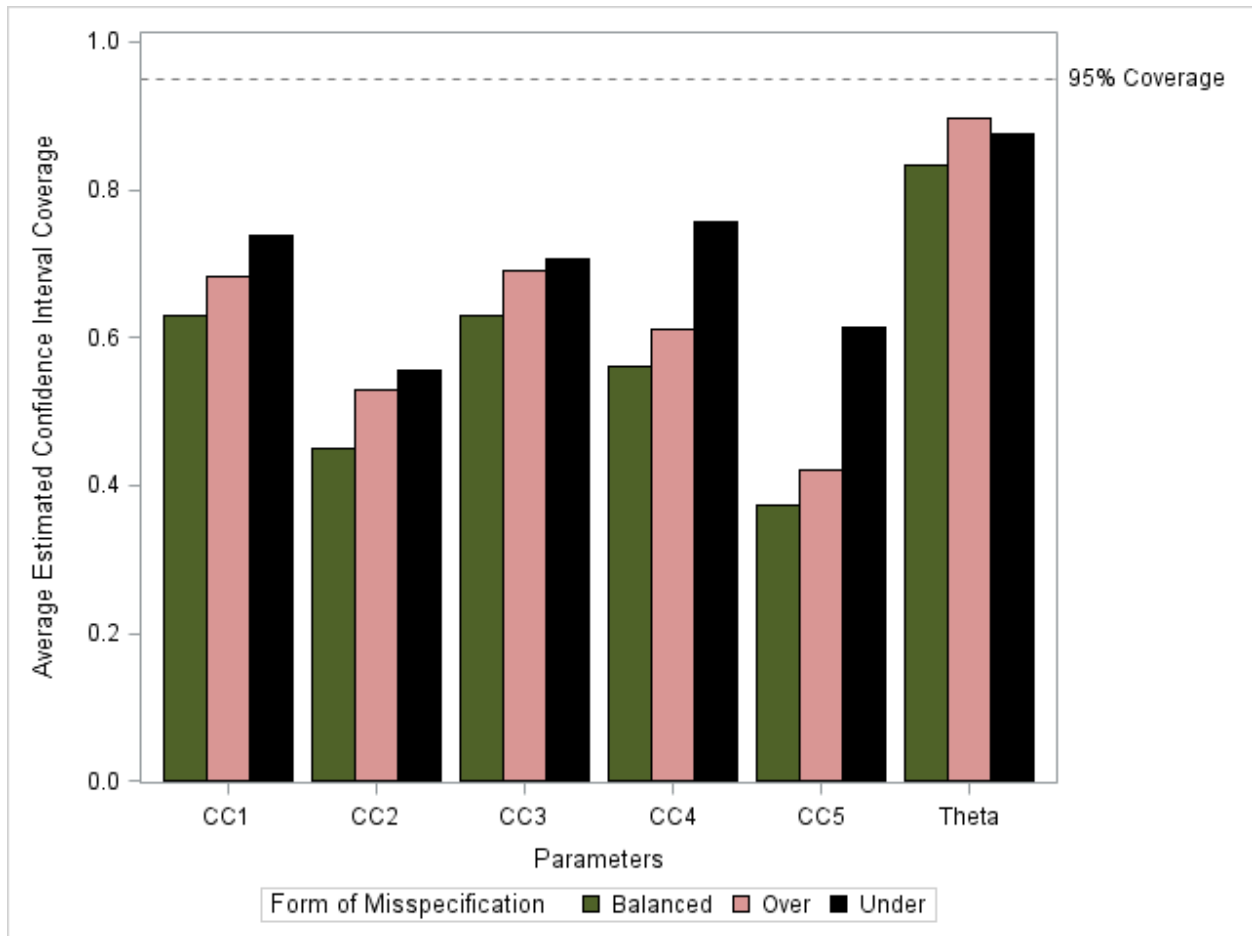


Figure 21. Average estimated confidence interval coverage for CC1-CC5 and theta by form of misspecification.

As Figure 21 demonstrates the cognitive components, and person ability parameter's confidence interval coverage rates were permissive; however the CI Coverage rates are impacted by percent of misspecification previously discussed. That said, when considering cognitive components under-specification is preferred to over-specification which in turn is

preferred to balanced misspecification. When considering person ability over-specification of the Q-matrix is preferred to underspecification of the Q-matrix, which in turn is preferred to balanced misspecification.

### **Sample Size**

Sample size (i.e., 20, 40, and 60) has a large impact of the mean CI Coverage rates in CC1 ( $\eta^2 = .3106$ ), CC2 ( $\eta^2 = .1866$ ), CC3 ( $\eta^2 = .3326$ ), and CC4 ( $\eta^2 = .2739$ ). Sample size has a medium impact of the mean CI Coverage rates in CC5 ( $\eta^2 = .1077$ ). Sample size accounts for 31.06% of the variance of the mean CI Coverage rates in CC1, 18.66% of the variance of the mean CI Coverage rates in CC2, 33.26% of the variance of the mean CI Coverage rates in CC3, 27.39% of the variance of the mean CI Coverage rates in CC4, and 10.77% of the variance of the mean CI Coverage rates in CC5. Figure 22 demonstrates that as sample size increased the fixed cognitive component truth parameter fell within the estimated cognitive component confidence interval less and less. In other words, as the size of the sample increased the accuracy of the cognitive components parameter estimates decreased dramatically.

These CI coverage results suggest that larger sample sizes produced less accurate parameter estimates which appear to contradict the RMSE and bias findings. This issue was explained under percentage of misspecification and will be discussed further in chapter 5. SAS employs a Wald-type algorithm to construct CIs around estimated parameters; however, these CIs may not function properly when estimating cognitive components. Future research may be conducted to determine if a bootstrapping approach to confidence interval estimation would be better choice of algorithm when constructing CIs around cognitive components, and person ability parameter estimates (DiCiccio & Efron, 1996; Efron & Tibshirani, 1986).



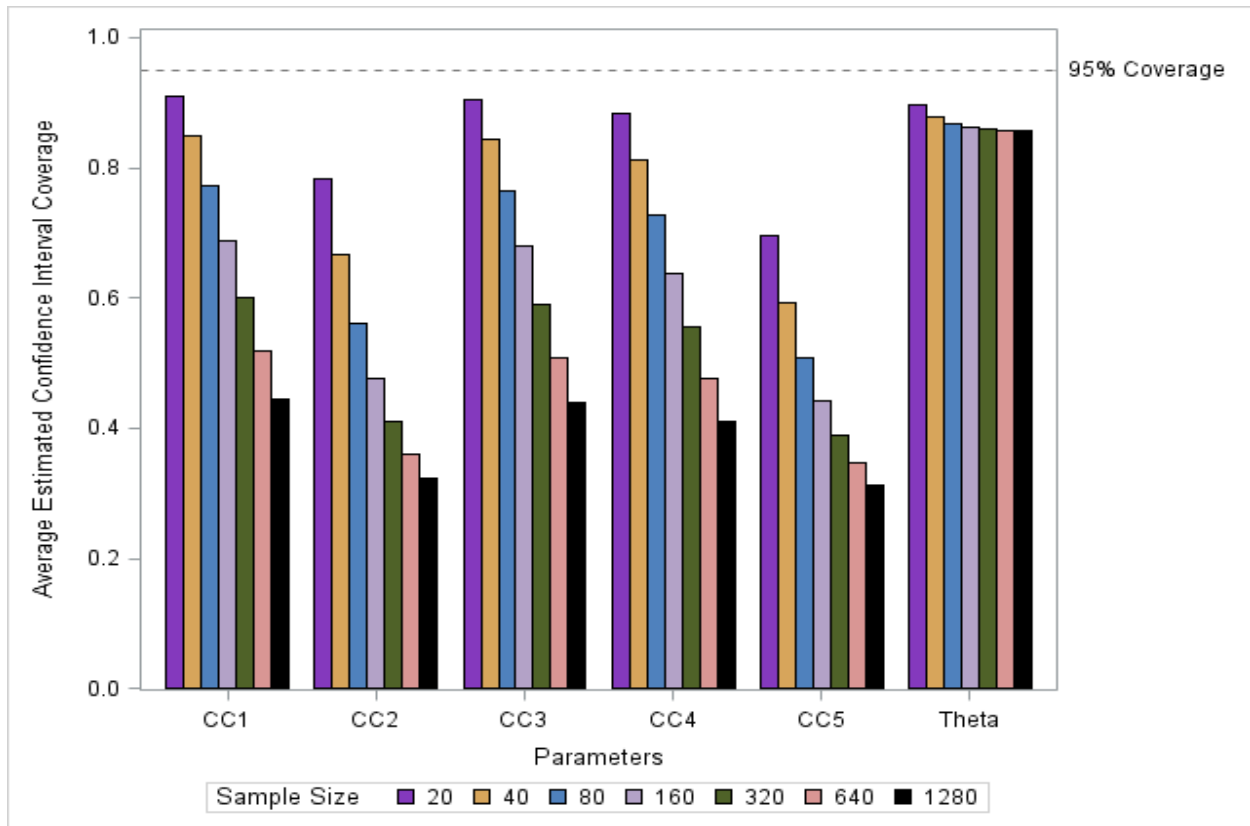


Figure 22. Average estimated confidence interval coverage for CC1-CC5 and theta by sample size.

Person ability parameter estimates do not fall within the confidence interval at the 95% nominal rate. Person ability parameter estimates fall within the CI interval at higher rates than the cognitive components parameter estimates; however, these rates are permissive.

#### Percent of Misspecification by Sample Size

Percent of misspecification by sample size has a medium impact on the mean CI Coverage rates in CC1 ( $\eta^2 = .1011$ ), CC3 ( $\eta^2 = .1053$ ), and CC4 ( $\eta^2 = .0827$ ). Percent of misspecification by sample size accounts for 10.11% of the variance of the mean CI Coverage rates in CC1, 10.53% of the variance of the mean CI Coverage rates in CC3, and 8.27% of the variance of the mean CI Coverage rates in CC4.

As Figure 23 graphically demonstrates as the percent of misspecification for cognitive components 1, 3, and 4 increased the estimated confidence interval coverage rates decreased dramatically as the sample size increased. In other words, for cognitive components 1, 3, and 4 percent of misspecification had a greater impact on the confidence interval coverage rates as the sample size increased.

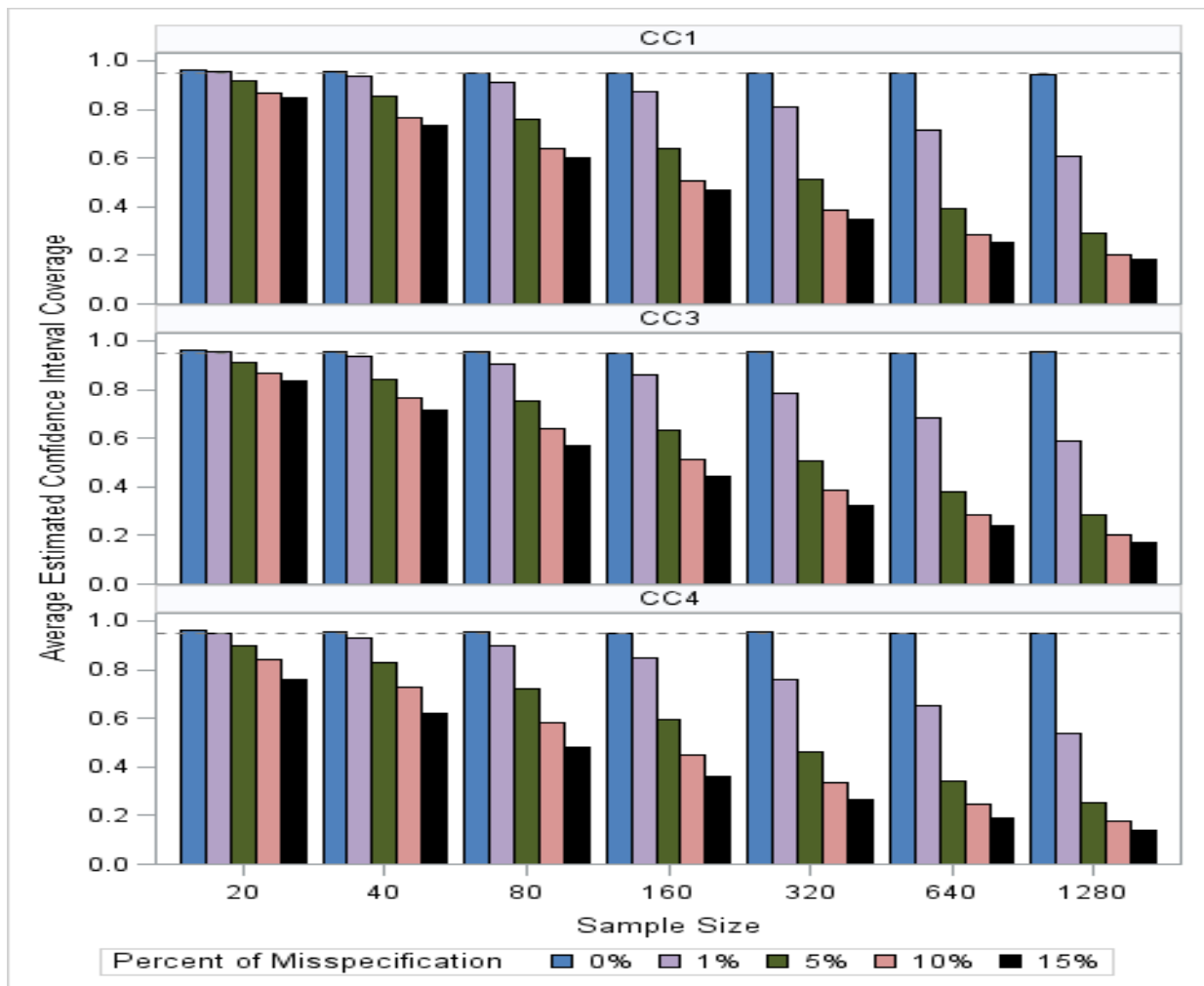


Figure 23. Average estimated confidence interval coverage CC1, CC3, and CC4 for percent of misspecification by sample size. The reference line marks 95% confidence interval coverage.

## Number of Items

The number of items (i.e., 20, 40, or 60) had a large impact had a large impact of the mean CI Coverage rates in theta ( $\eta^2 = .2431$ ). The number of items had a medium impact of the mean CI Coverage rates in CC2 ( $\eta^2 = .0588$ ). The number of items accounted for 24.31% of the variance of the mean CI Coverage rates in theta, and 5.88% of the variance of the mean CI Coverage rates in CC2.

As Figure 24 demonstrates as the number of items increased the cognitive component and person ability CI coverage rates decreased. The truth parameter is expected to fall within the estimated confidence interval approximately 95% of the time and there can be no doubt that these coverage rates are permissive.

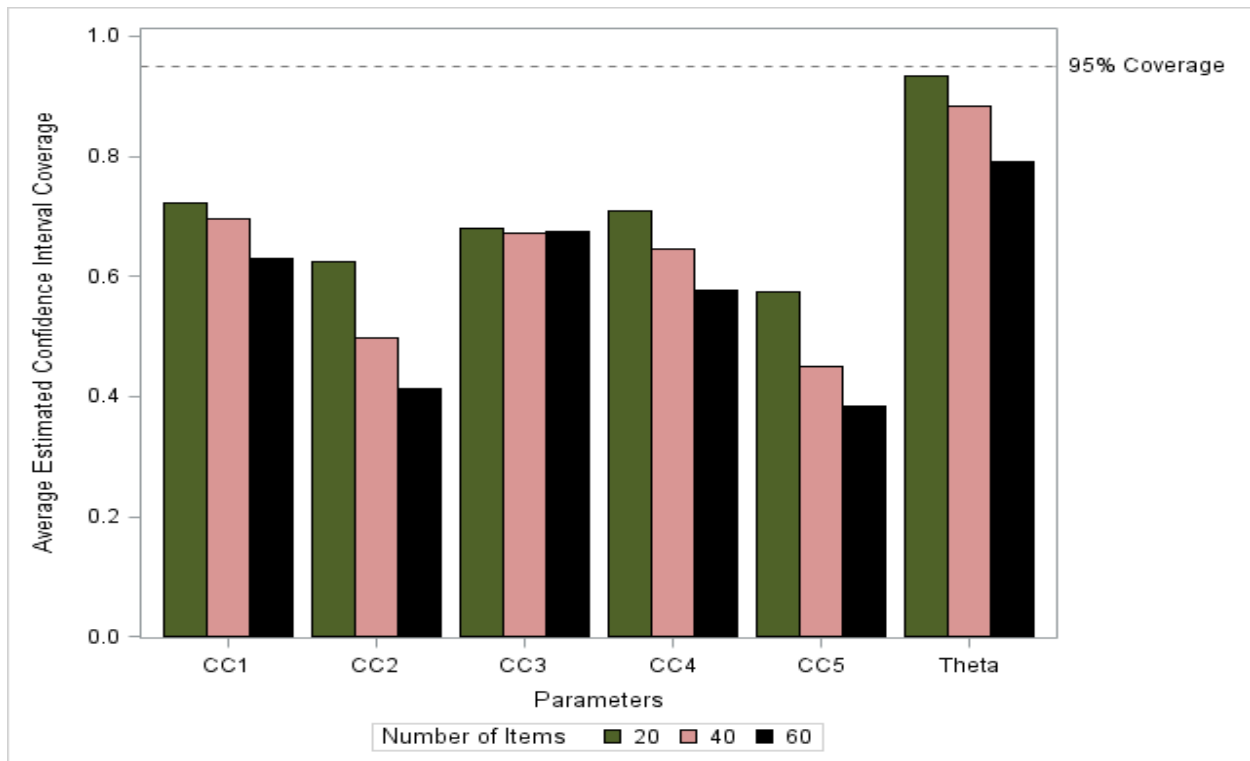


Figure 24. Average estimated confidence interval coverage for CC1-CC5 and theta.

## Summary of Estimated Confidence Interval Coverage

The summary results of the mean confidence interval coverage can be found in Table 16. Based on the confidence interval coverage percent of misspecification had the largest practically significant effect on the estimates of cognitive components, and person ability. The form of misspecification had a practically significant effect on the estimates of some cognitive components but not on parameter estimates of person ability.

Table 16

### Summary of Mean Confidence Interval Coverage

Design Factor	Impact on mean Confidence Interval Coverage Rates for CCs, & Person Ability by Design Factor
Percent	As the percent of misspecification increased the percentage of cognitive components, and person ability parameter estimates falling within the confidence interval dramatically decreased
Form	Under-speciation yielded CI Coverage rates closer to the nominal coverage rate of .95 compared to over-specification, and over-specification yielded CI Coverage rates closer to the nominal coverage rate compared to balanced misspecification for cognitive components 4, and 5.
Percent*Form	No practically significant effect on any outcome variable.
Percent*SS	As the percent of misspecification increased the estimated confidence interval coverage rates decreased dramatically as the sample size increased for cognitive components 1, 3, and 4.
Form*SS	No practically significant effect on any outcome variable.
Percent*Density	No practically significant effect on any outcome variable.
Form*Density	No practically significant effect on any outcome variable.
Percent*Items	No practically significant effect on any outcome variable.
Form*Items	No practically significant effect on any outcome variable.
Percent *Shape	No practically significant effect on any outcome variable.
Form *Shape*	No practically significant effect on any outcome variable.

There was one practically significant first-order interaction related to Q-matrix misspecification, namely, percent of misspecification with sample size. Percent of misspecification by sample size had an interactive effect on the estimates of some cognitive components but not on the estimates of person ability. All other first order interaction effects related to Q-matrix misspecification (i.e., percent\*form, form\*SS, percent\*density, form\*density, percent\*items, form\*items, percent\*shape, and form\*shape) did not have practically significant effects on all the estimates of cognitive components, or person ability.

### Confidence Interval Width

The confidence interval width (CI width) is a measure of precision of the parameter estimate and is calculated as the average difference between the upper and lower limits of the confidence interval. The overall distribution of the CI width for CC1-CC5 and theta values across all simulation conditions is illustrated in box plots in Figure 25.

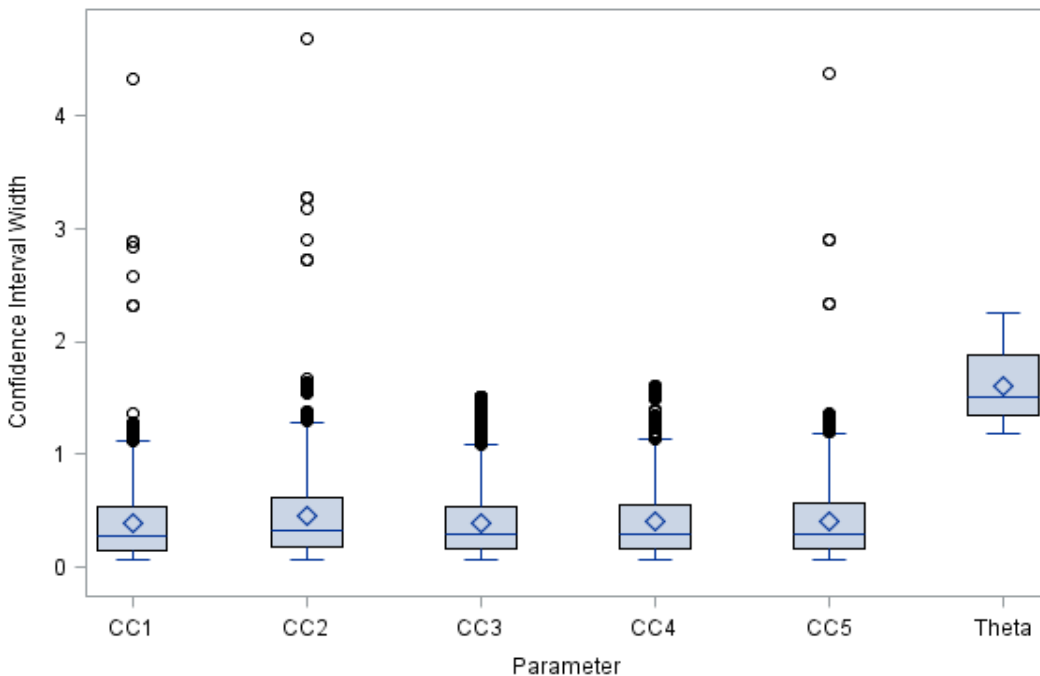


Figure 25. Distribution of the confidence interval width for CC1 - CC5 and theta.

The mean, standard deviation, minimum, and maximum values for the CI width for CC1-CC5, and theta are presented in Table 11. Closely examining the box plots in Figure 25 and the descriptive statistics in Table 17 reveals the cognitive components have CI widths with tight boxes but means ranging from .39 to .44. In comparison, person ability means are farther away from 0 (i.e., 1.6) and its boxes are not tight.

Table 17

*Mean, Standard Deviation, Minimum, and Maximum Values for Confidence Interval width for CC1-CC5 and theta*

	MEAN	SD	MIN	MAX
CC1	0.3917	0.3269	0.0696	4.3358
CC2	0.4492	0.3799	0.0696	4.6877
CC3	0.3957	0.3044	0.0681	1.5166
CC4	0.4052	0.3144	0.0672	1.6074
CC5	0.4102	0.3343	0.0698	4.3797
Theta	1.5994	0.2891	1.1901	2.2482

To explore variance patterns which might emerge from the CI width values an ANOVA for the design factors (i.e., percent of misspecification, form of misspecification, sample size, density of Q-matrix, number of items in Q-matrix, and the skew of the person ability distribution) in this study was computed. Table 18 presents eta-squared values ( $\eta^2$ ) for the association of design factors with estimated CI width values for CC1-CC5, and theta.

The design factors and their first order interaction total eta squared values are .8955 for CC1, .9134 for CC2, .9931 for CC3, .9920 for CC4, .9202 for CC5, and .9992 for theta. In other words, the design factors and their first order interaction effects account for 89.55% of the variance on the mean CI width values in CC1, 91.34% of the variance on the mean CI width values in CC2, 99.31% of the variance on the mean CI width values in CC3, 99.20% of the

variance on the mean CI width values in CC4, 92.02% of the variance on the mean CI width values in CC5, and 99.92% of the variance on the mean CI width values in theta.

Examining Table 18 results reveals there are large eta squared effect sizes (Cohen, 1998) for main effects for: (a) sample size: CC1 ( $\eta^2 = .7296$ ), CC2 ( $\eta^2 = .7033$ ), CC3 ( $\eta^2 = .7989$ ), CC4 ( $\eta^2 = .7872$ ), CC5 ( $\eta^2 = .7534$ ), Theta ( $\eta^2 = .1234$ ), and (b) number of items: theta ( $\eta^2 = .8637$ ). There are medium eta squared effect sizes (Cohen, 1998) for main

Table 18

*Eta-Squared Values for the Association of Design Factors and 1st Level Interaction Effects with the Average Overall Confidence Interval Width for CCs and Theta*

	CC1	CC2	CC3	CC4	CC5	Theta
Percent	.0039	.0169	.0029	.0039	.0090	.0002
Form	.0017	.0036	.0009	.0040	.0046	.0000
Sample Size	** .7296	** .7033	** .7989	** .7872	** .7534	** .1234
Density	.0051	.0116	.0217	.0281	.0000	.0063
Items	* .0750	* .0823	* .0992	* .0927	* .0777	** .8637
Skew	.0007	.0005	.0001	.0001	.0007	.0037
Percent*Form	.0018	.0023	.0006	.0026	.0020	.0001
Percent*SS	.0065	.0133	.0013	.0018	.0106	.0000
Percent*Density	.0008	.0017	.0052	.0067	.0003	.0002
Percent*Items	.0015	.0013	.0000	.0000	.0010	.0003
Percent*Skew	.0005	.0004	.0000	.0000	.0005	.0000
Form*SS	.0013	.0022	.0004	.0019	.0032	.0001
Form*Density	.0004	.0012	.0012	.0006	.0002	.0000
Form*Items	.0003	.0007	.0001	.0001	.0003	.0001
Form*Skew	.0001	.0001	.0000	.0000	.0001	.0000
SS*Density	.0073	.0107	.0102	.0134	.0012	.0003
SS*Items	.0552	.0559	.0477	.0447	.0523	.0002
SS*Skew	.0022	.0016	.0000	.0000	.0019	.0001
Density*Items	.0007	.0030	.0028	.0040	.0005	.0000
Density*Skew	.0003	.0002	.0000	.0000	.0002	.0000
Items*Skew	.0007	.0005	.0000	.0000	.0006	.0007
Total Explained	.8955	.9134	.9931	.9920	.9202	.9992

Note 1. Percent=Percent of Misspecification, Form=Form of Misspecification, Density=Density of Q-Matrix, Items=Number of Items, Skew=Skewness of Person Ability Distribution, SS=Sample Size,

Note 2. \*indicates a medium effect size, \*\* indicates a large effect size

## Sample Size

Sample size has a large impact on the mean CI width values in CC1 ( $\eta^2 = .7296$ ), CC2 ( $\eta^2 = .7033$ ), CC3 ( $\eta^2 = .7989$ ), CC4 ( $\eta^2 = .7872$ ), CC5 ( $\eta^2 = .7534$ ), and theta ( $\eta^2 = .1234$ ). Sample size accounts for 72.96% of the variance on the mean CI width values in CC1, 70.33% of the variance on the mean CI width values in CC2, 79.89% of the variance on the mean CI width values in CC3, 78.72% of the variance on the mean CI width values in CC4, 75.34% of the variance on the mean CI width values in CC5, and 12.34% of the variance on the mean CI width values in theta. The overall trend in Figure 12 demonstrates that as the sample size increased the cognitive components and person ability parameter estimates became more precise. It is worth noting, the CIs for theta are very wide.

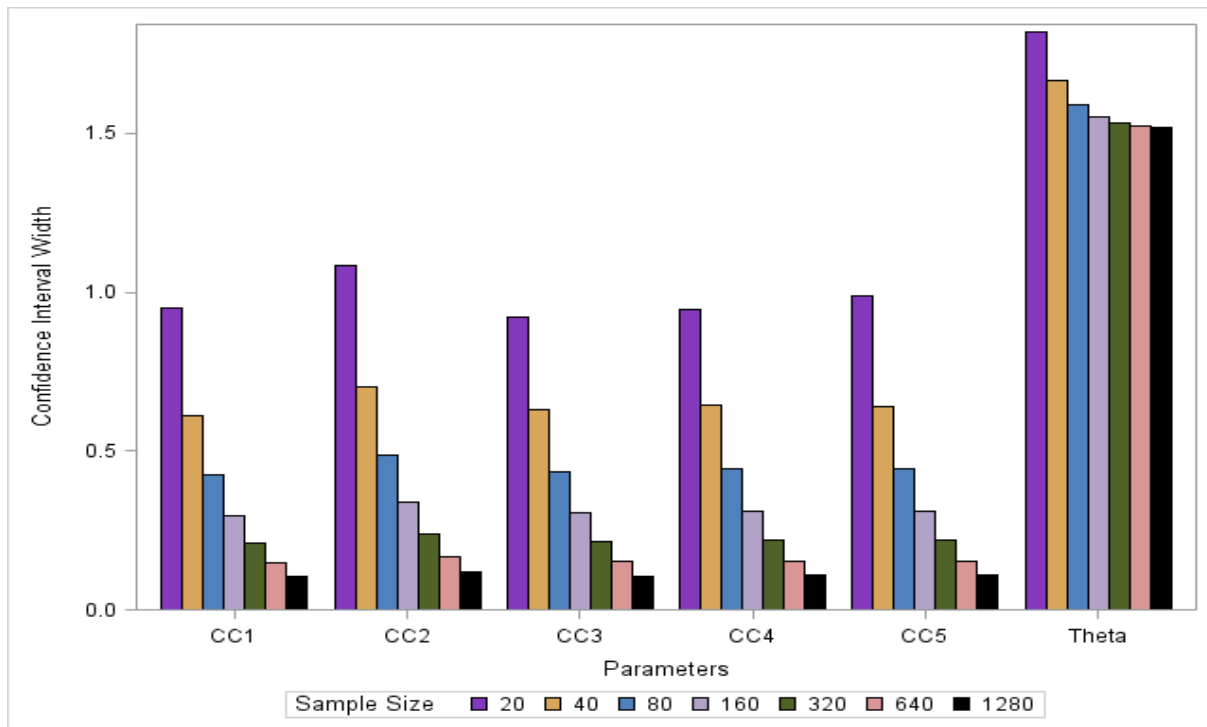


Figure 26. Average confidence interval width for CC1-CC5 and theta by sample size.



## Number of Items

Number of items had a large impact had a large impact on the mean CI width values in theta ( $\eta^2 = .8637$ ). Number of items had a medium impact on the mean CI width values in CC1 ( $\eta^2 = .0750$ ), CC2 ( $\eta^2 = .0823$ ), CC3 ( $\eta^2 = .0992$ ), CC4 ( $\eta^2 = .0927$ ), and CC5 ( $\eta^2 = .0777$ )

Increasing the number of items accounted for 86.37% of the variance on the mean CI width values in theta, 7.50% of the variance on the mean CI width values in CC1, 8.23% of the variance on the mean CI width values in CC2, 9.92% of the variance on the mean CI width values in CC3, 9.27% of the variance on the mean CI width values in CC4, and 7.77% of the variance on the mean CI width values in CC5.

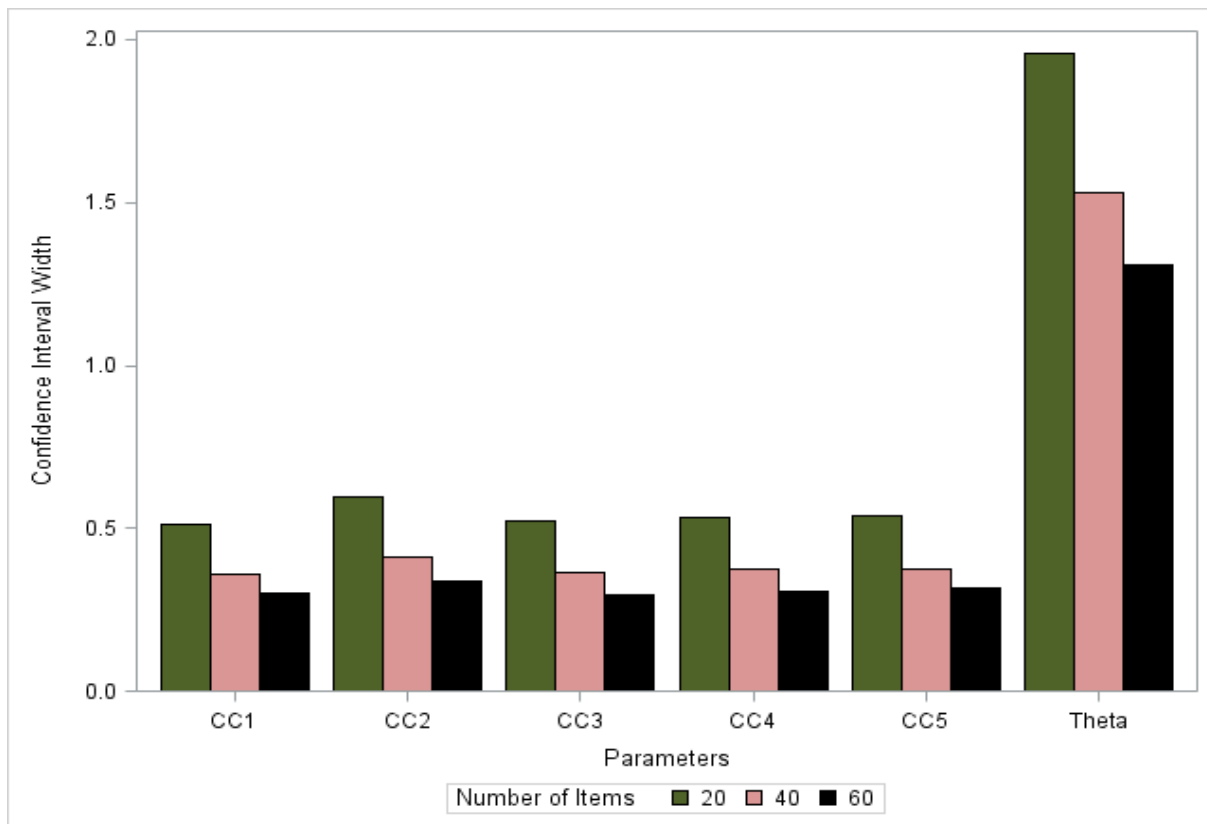


Figure 27. Average estimated confidence interval width for CC1-CC5 and theta.

Figure 27 demonstrates that as the number of items increased the cognitive components and person ability parameter estimates became more precise. The mean confidence interval width for the cognitive components fell to approximately .5 even when the number of items was 20; however, the confidence intervals width for person ability remained wide even when the number of items was at its largest.

### Summary of Confidence Interval Width

The summary results of the mean confidence interval width can be found in Table 19. Bases on the mean confidence interval width there were no practically significant effects related to Q-matrix misspecification on all the estimates of cognitive components or person ability.

*Table 19*

*Summary of Mean Confidence Interval Width*

Design Factor	Impact on mean Confidence Interval Width for CCs, & Person Ability by Design Factor
Percent	No practically significant effect on any outcome variable.
Form	No practically significant effect on any outcome variable.
Percent*Form	No practically significant effect on any outcome variable.
Percent*SS	No practically significant effect on any outcome variable.
Form*SS	No practically significant effect on any outcome variable.
Percent*Density	No practically significant effect on any outcome variable.
Form*Density	No practically significant effect on any outcome variable.
Percent*Items	No practically significant effect on any outcome variable.
Form*Items	No practically significant effect on any outcome variable.
Percent *Shape	No practically significant effect on any outcome variable.
Form *Shape*	No practically significant effect on any outcome variable.

### Phi Pairwise Cognitive Components Coefficients

To explore the impact of misspecification of the Q-matrix on the strength and direction of the pairwise relationships between cognitive components the phi pairwise correlation coefficients were computed (N=2,050,650) for all design factors in this simulation study. The Fisher Z-transformed values were computed and the values compared across Q-matrices. Table 20 presents the average phi pairwise correlation values for the cognitive components in columnar form for the Truth, under-specification, over-specification, and balanced misspecification by percent of misspecification for sparse Q-matrices. Table 21 presents the average phi pairwise correlation values for the cognitive components in columnar form for the Truth, under-specification, over-specification, and balanced misspecification by percent of misspecification for dense Q-matrices. Recall that the phi correlations provide a measure of whether a pair of cognitive component predominantly measure concepts within the same items, separately, or some mixture of the two.

Table 20

*Pairwise Cognitive Component Phi Correlation Coefficients for Truth, Under, Over, and Balanced Misspecification by Percent of Misspecification in the Sparse Q-matrix*

	Truth		Under			Over				Balanced			
	0%	1%	5%	10%	15%	1%	5%	10%	15%	1%	5%	10%	15%
r12	.41	.39	.31	.20	.08	.39	.34	.29	.24	.37	.25	.10	-.04
r13	-.38	-.36	-.30	-.26	-.23	-.36	-.31	-.26	-.21	-.34	-.24	-.15	-.08
r14	-.38	-.36	-.30	-.26	-.23	-.36	-.31	-.26	-.21	-.34	-.24	-.15	-.09
r15	-.38	-.36	-.30	-.26	-.24	-.36	-.31	-.26	-.21	-.34	-.24	-.15	-.10
r23	-.20	-.20	-.17	-.15	-.15	-.20	-.17	-.14	-.11	-.19	-.14	-.09	-.06
r24	-.20	-.20	-.17	-.15	-.16	-.20	-.17	-.14	-.11	-.19	-.14	-.09	-.08
r25	-.82	-.78	-.67	-.54	-.44	-.78	-.66	-.54	-.44	-.75	-.52	-.29	-.13
r34	-.25	-.24	-.21	-.18	-.17	-.24	-.20	-.16	-.13	-.23	-.16	-.10	-.06
r35	.17	.16	.13	.09	.03	.16	.13	.10	.08	.15	.10	.03	-.04
r45	-.04	-.04	-.04	-.05	-.08	-.04	-.03	-.03	-.02	-.04	-.03	-.04	-.05

The results of Green and Smith's (1987) analysis suggests that LLTM is sensitive to the presence of correlated coefficients, measurement disturbances, and misspecified cognitive components. In one part of the simulation, three cognitive components coefficients were highly correlated with  $r_{12} = .968$ ,  $r_{13} = .884$ , and  $r_{23} = .895$ . They found in the absence of collinearity of the cognitive component LLTM worked well. When the cognitive components were collinear accuracy of the estimation only increased once the sample size was greater than 200; however, in the case of collinearity, a unique decomposition of item parameters did not exist.

Table 21

*Pairwise Cognitive Component Phi Correlation Coefficients for Truth, Under, Over, and Balanced Misspecification by Percent of Misspecification in the Dense Q-matrix*

	Truth		Under			Over				Balanced			
	0%	1%	5%	10%	15%	1%	5%	10%	15%	1%	5%	10%	15%
r12	-.25	-.23	-.18	-.16	-.15	-.24	-.21	-.18	-.12	-.22	-.15	-.10	-.09
r13	-.41	-.38	-.31	-.25	-.21	-.39	-.34	-.28	-.17	-.37	-.24	-.14	-.07
r14	-.61	-.58	-.48	-.38	-.32	-.58	-.48	-.38	-.22	-.55	-.36	-.19	-.08
r15	-.61	-.58	-.48	-.39	-.33	-.58	-.48	-.38	-.23	-.55	-.36	-.20	-.11
r23	.61	.58	.45	.34	.25	.59	.51	.41	.23	.55	.35	.17	.05
r24	.41	.39	.31	.24	.18	.39	.32	.25	.14	.37	.23	.11	.03
r25	-.61	-.58	-.48	-.39	-.32	-.58	-.48	-.38	-.22	-.55	-.36	-.19	-.08
r34	.67	.64	.56	.46	.38	.63	.51	.39	.22	.61	.41	.21	.08
r35	-.17	-.16	-.14	-.12	-.11	-.16	-.13	-.10	-.06	-.15	-.11	-.06	-.04
r45	.17	.16	.14	.12	.09	.16	.12	.09	.04	.15	.10	.04	.01

Figure 28 graphically represents what is presented in tabular form in Tables 20 and 21. The phi coefficients have values closer to -1.0 or 1.0 in the Truth Q-matrix. As the percent of misspecification of the Q-matrix increases the cognitive components average phi coefficients move closer and closer to 0. When considering form of misspecification, the cognitive components phi coefficients are closest to 0 when the Q-matrix is balanced misspecified compared to when the Q-matrix is overspecified, and over-specification produces phi

coefficients closer to 0 when compared to under-specification as the percent of misspecification increases.

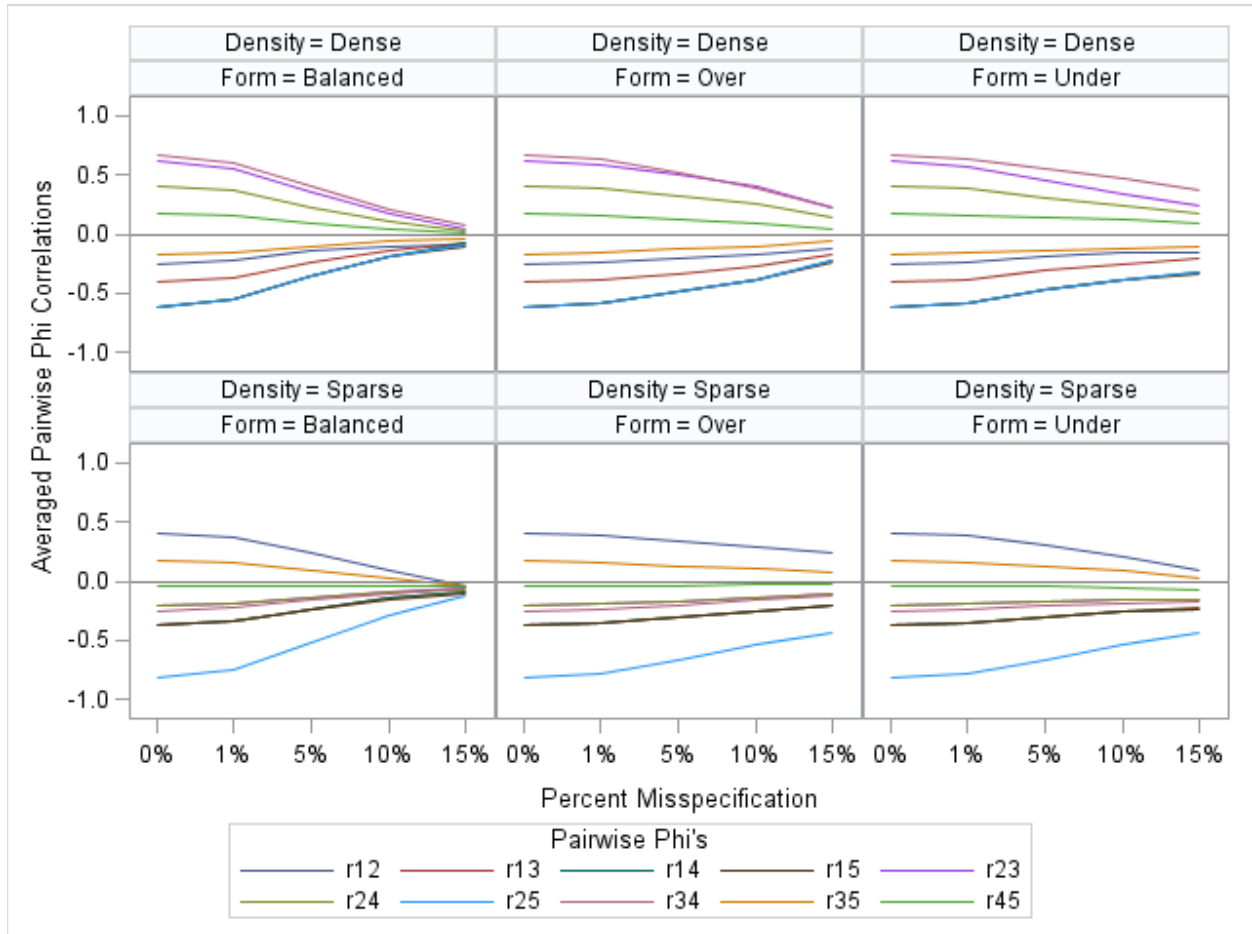


Figure 28. Average pairwise phi correlation coefficients by percent of misspecification for dense and sparse Q-matrices in all form of misspecification.

The phi coefficients in the sparse condition had a range of -.82 to .41 and in the dense condition the range was -.61 to .67. In effect the phi coefficients were never highly correlated in this study which allows for the accurate measurement of the parameters while providing a unique decomposition of cognitive components.

### **Performance of SAS PROC NLMIXED**

There are several results to report concerning the functioning of SAS 9.3. First, in all there were 2,050,650 replications conducted using SAS PROC NLMIXED. The models converged in every case without exception for every single replication for all combinations of the six design factors of this simulation. Second, SAS 9.3 offers PROC IML which is a very powerful environment for researchers who want to conduct simulation studies. Third, rather than compute beta as a sufficient statistic, SAS 9.3 employs a combination of marginal maximum likelihood modeling for estimating beta, and an empirical Bayes approach which is a random draw from a density defined over the person population for theta. This random draw over the person ability distribution fulfills the function of the polynomial equation in as much as theta is not constrained during the computation of beta. These advanced algorithms permit a complicated non-linear analysis which is not readily available in other software packages. In sum, SAS was a superb choice for conducting LLTM simulations.

## CHAPTER FIVE: DISCUSSION

This study focused on the effects of percent of misspecification and form of misspecification of the Q-matrix on cognitive components, item difficulty, and person ability parameter estimates as indexed by bias, RMSE, CI Coverage, and CI Width. Index is defined as the relative change occurring in bias, RMSE, CI Coverage, and CI Width for parameter estimates on the design factors in this study. To deepen the understanding of misspecification the impact of the first order interaction effects of percent of misspecification and form of misspecification on sample size, Q-matrix density, number of items, and skewness of the person distribution were examined.

The bias statistic informs the researchers how close the average estimated parameter is to the true parameter. For the purposes of this research anything less than 0.20, or less than a small effect size (Cohen, 1988), was considered tolerable error. Along with examining bias the precision of the parameter estimate was evaluated. SAS by default in its non-linear mixed procedure employs the Wald-type confidence intervals which are based on the asymptotic normality of the parameter estimators. The third statistic in this analysis, RMSE, is a combination of bias and variance allowing for a measure of overall variability. For the purposes of this research values less than 0.20, less than a small effect size (Cohen, 1988), were considered tolerable. Confidence interval coverage is a measure of accuracy which indicates the percentage of time the truth parameter falls within the lower and upper bounds of the estimated parameter. Typically the truth parameter will be found within the confidence interval 95% of the time.

The truth Q-matrix was created, the initial cognitive component values were set, the item-response matrix was generated, and the person ability matrix was randomly generated.

The non-linear mixed procedure fits the model by maximizing an approximation to the likelihood integrated over the random effects. The conditional maximum likelihood approach employed by Fischer was limited because no inferences are possible on the person effects, and the conditional maximum likelihood is maximized rather than the full likelihood (Tuerlinckx, et al, 2004). The approach taken in SASs non-linear mixed procedure is to form the marginal likelihood by integrating with respect to the random effects. In this way the undesirable limitation on person inferences is avoided.

Most model fit analysis does not address the correctness of the constructed Q-matrix (de la Torre, 2008). However, the amount of model misspecification that the Q-matrix can tolerate and still function adequately is unknown. Little simulation work examining the sensitivity of the LLTM to misspecification of the Q-matrix has been conducted (Baker, 1993). This was the issue of interest in this research study, to provide some evidence of the functioning of the LLTM when the Q-matrix or Weight Matrix is misspecified.

Questions 1 through 6 will examine:

- 1.) To what extent does the LLTM function well when the Q-matrix is progressively more misspecified?
- 2.) To what extent does the LLTM function well when the Q-matrix is properly specified, under specified, balanced misspecified, and over specified?
- 3.) To what extent does the LLTM function well under different conditions of model misspecification when the sample size varies?
- 4.) To what extent does the LLTM function well under different conditions of model misspecification when the Q-matrix is densely or sparsely populated?
- 5.) To what extent does the LLTM function well under different conditions of model misspecification when test length varies?



- 6.) To what extent does the LLTM function well under different conditions of model misspecification when the population distribution is normally distributed, negatively skewed, and positively skewed?

Green and Smith (1987), Baker (1993) and Cassuto (1996) represent the sum of the LLTM literature on Q-matrix simulation, but in fact the work of Green and Smith is focused on model comparison while Cassuto is focused on parameter recovery. So in some ways an argument could be made there is really only one true LLTM misspecification study. That said these studies are foundational and informed decision making about research questions, and methods in this present study.

### **The Impact on Parameters of Misspecifying the Q-matrix**

The percent of misspecification of the Q-matrix had a profound practically significant effect on cognitive components, item difficulty and person ability parameter estimates. As the percent of misspecification increased more and more error was introduced into the model. This effect, indexed by bias, RMSE, and CI Coverage values, was present for all cognitive components, item difficulty, and person ability with the proportion of variance explained ranging from 21% to 80%. Thus this can be thought of as the most important result of this study.

Interestingly, as the percent of misspecification increased the precision of the model parameters was not impacted, which was an unexpected result. In this non-linear mixed modeling environment the CI Width is a Wald-type confidence interval with lower and upper bounds computed as a standard deviation around the mean of the estimated parameter. In other words the CI Width statistic indexes the error variance of the parameter estimate. This suggests that percent misspecification of the Q-matrix causes the parameter estimates to move farther and farther away from the truth, but no matter how far those parameter estimates move away from the truth they still remain precise. In the simulation world researchers can establish

the truth and easily quantify bias, but in the observed world the truth is rarely known and bias is a theoretical construct. The danger is that researchers, who cannot quantify bias in their study, may take a degree of comfort in the precision of LLTM parameter estimates. The result could lead researchers to have confidence in precise parameter estimates; whereas, those estimates may contain a significant amount of bias causing them to over-estimate or under-estimate the truth.

It was found that as the percent of misspecification increased the cognitive components, item difficulty and person ability parameters were typically increasingly negatively biased, although one cognitive component was positively biased. Given that the cognitive component parameter estimates were arbitrarily set from .03 to 1.2. it is clear that misspecification causes positive cognitive components to be underestimated. Further, item difficulty and person ability parameter estimates were also underestimated when the Q-matrix was misspecified.

As in the case of Baker (1993) this simulation study is focused on misspecification of the Q-matrix. Other simulation studies, such as the Cassuto's (1996) dissertation, focused primarily on parameter estimation recovery. As has been discussed, the precision of the LLTM parameter estimates, indexed as CI Width, was not impacted by misspecification of the Q-matrix. Given that root mean squared error includes bias and variance error, indexed as CI Width, a clearer picture of the effects of misspecification of the Q-matrix will emerge when consulting the bias statistic rather than RMSE. In other words, it is preferable when thinking of the impact of misspecification of the Q-matrix to consider bias rather than RMSE to avoid clouding Q-matrix misspecification results with considerations of error variance that clearly have no impact in this study.

That said, at all levels of misspecification a majority of cognitive components had bias values within tolerable error limits when the Q-matrix was misspecified 15%. This suggests that cognitive components can be robust to the effects of misspecification of the Q-matrix. Item

difficulty and person ability parameter estimates only tolerated lower levels of misspecification, typically less than 5 % misspecification. This suggests that item difficulty and person ability are less robust to misspecification of the Q-matrix than the cognitive components. This is not surprising, in the case of item difficulty, because these parameters are computed as a linear combination of the cognitive components and therefore accumulate the error associated with each of the cognitive component. It is interesting that person ability parameter estimates only tolerate low levels of misspecification. These estimates are computed employing an empirical bayes approaches which begs the question of whether or not other empirical Bayes estimation algorithms might be more robust to percent of misspecification of the Q-matrix.

Misspecification of the Q-matrix has been a key issue in the literature since Fischer introduced LLTM in the 1970s. The Q-matrix has been found to be sensitive to misspecification (Baker, 1993; Cassuto, 1996; Green & Smith, 1987; Kunina-Habernicht, Rupp & Wilhelm, 2012; MacDonald & Kromrey, 2012; Rupp & Templin, 2008; Rupp et al., 2012). Baker (1993) concluded that a small degree of misspecification in the Q-matrix had a large impact on the parameter estimates. It may appear that Baker (1993) conclusions are contradicted by this studies results that a majority of cognitive components can tolerate up to 15% misspecification; however, a closer examination reveals than RMSE values in both studies are similar. The difference is this study weights the bias statistic more heavily in drawing conclusions about Q-matrix misspecification.

In the truth condition the cognitive component and person ability parameter estimates were accurate (i.e., the truth parameter fell within the confidence interval at the nominal rate). As the percent of misspecification in the Q-matrix increased the cognitive component, and person ability accuracy decreased dramatically. This suggests that misspecification of the Q-matrix in LLTM modeling can produces cognitive components and person ability parameter

estimate which are inaccurate. Thus the findings associated with the bias and RMSE indices are supported.

### **Under-specifying, Over-specifying or Balanced Misspecifying the Q-Matrix**

Whether the Q-matrix was under-specified, over-specified, or balanced misspecified had a practically significant effect on the estimates of a majority of the cognitive components but, interestingly, not on item difficulty or person ability. Under-specified Q-matrices led to a majority of cognitive components with less bias compared to when the Q-matrix was balanced-misspecified or over-specified Q-matrices. Further, as the percent of misspecification increased, the balanced-misspecified and over-specified Q-matrix led to a majority of cognitive components with dramatically more bias, compared to the under-specified Q-matrix. All cognitive components parameter estimates fell within tolerable error limits when the Q-matrix was under-specified compared to only a majority of cognitive components which fell within tolerable error limits for over-specification and balanced misspecification.

Often when researchers are specifying a Q-matrix they will have to decide if a cognitive component is present in an item. When faced with decisions like these researchers should keep in mind that under-specifying the Q-matrix will likely produce estimates closer to the truth. As well, researchers should also keep in mind that as the cognitive components are more and more misspecified under-specification of the Q-matrix produces cognitive components with dramatically closer to the truth than does over-specified or balanced misspecified Q-matrices. This suggests that under specified Q-matrices are to be preferred to over specified Q-matrices, and over specified Q-matrices are much better than balanced misspecified Q-matrices.

Interestingly, all three forms of misspecification of the Q-matrix had no practically significant effect on the precision (i.e., variance), indexed as CI Width, of cognitive components, and person ability estimates.

Under-speciation yielded CI Coverage rates closer to the nominal coverage rate of .95 compared to over-specification, and over-specification yielded CI Coverage rates closer to the nominal coverage rate compared to balanced misspecification for a majority of cognitive components; however, when the Q-matrix is under-specified, over-specified or balanced misspecified the permissiveness of the CI Coverage rates can only be described as dreadful.

Under-specification, over-specification and balanced misspecification has been a key design factor in the Q-matrix simulation literature; however, little there has been no discussion related to LLTM misspecification of the Q-matrix studies. The only discussion is found in Cassuto's doctoral dissertation (Cassuto, 1996) parameter recovery study in which he concludes that under-specification is preferred to other forms of misspecification.

### **Misspecification Interacting with Sample Size**

Surprisingly, LLTM parameter estimates do not demonstrate an increase or decrease in bias, RMSE, or precision when the Q-matrix is misspecified and the sample size increases. In fact there was only one significant interaction between misspecification and sample size. For a majority of the cognitive components as the percent of misspecification increased the accuracy of the cognitive components dramatically decreased once the sample size was 320 or greater. In other words, as the percent of misspecification increases larger sample sizes can be more impacted and much less accurate. This is important because it is well known that increasing sample size and number of items increases the precision of error variance. If researchers wish to improve the precision of their LLTM models they must do so carefully because they may cause the LLTM parameter estimates to become less accurate especially as the sample size increases beyond 320 persons. It is useful to know that a majority of the cognitive components parameter estimates fell within tolerable error once the sample size was between 40 to 80 participants.

Baker (1993) found that to obtain proper estimates of the cognitive components there must be a minimum number of examinees responding correctly to the items for each cognitive component. He further reported that once the estimation of cognitive components becomes stable, further increases in sample size have little additional effect on parameter estimates. Cassuto (1996) demonstrates a similar concern when he suggested if there are not enough examinees in the sample size to stabilize the estimation of the cognitive components in the correlated sparse Q-matrix the resulting sharing of variance across a correlated Q-matrix caused attenuation in the recovery of cognitive components coefficients. He suggests that estimation of cognitive component coefficients should settle down somewhere between 50 to 250 examinees.

The results of this study are consistent with the results of both Baker (1993), and Cassuto (1996) although his results are more aligned with parameter recovery rather than misspecification. The suggestion is that cognitive components settle down somewhere above 50 participants and that adding sample size does not necessarily improve cognitive component parameter estimates. That being said this research further cautions researchers that larger sample sizes pose a threat to accuracy not previously identified in the literature.

### **Misspecification Interacting with Dense or Sparse Q-matrices**

Interestingly, the only interaction effect was between percent of misspecification with dense and sparse matrices, as indexed by bias, for a minority of cognitive component. In this instance the dense Q-matrix produced larger bias values for a minority of cognitive components when compared to sparse Q-matrices. The suggestion is that under-specification of the Q-matrix can produce cognitive components that are closer to the truth, as indexed by bias. Care must be taken interpreting this finding given this interaction effect only applied to a minority of the cognitive component and not to item difficulty or person ability.

Baker (1993) suggests that the density of the Q-matrix is an important factor affecting the estimates of cognitive components. He notes that when the misspecification of the Q-matrix was from 1%-3%, the sparse Q-matrix yielded a larger root mean square (RMS) average than did the dense Q-matrix. This was also true for higher levels of misspecification between 5%-10%. Cassuto (1996) found that sparse Q matrices are more tolerable than dense Q matrices because each cognitive component will have more weight in determining the actual cognitive components coefficients and will yield items that are less complex in structure. Therefore he concluded, the density of the Q-matrix is an important factor affecting the estimates of cognitive components. He suggested further research needed to be conducted on sparse Q matrices especially when the sample size or number of respondents is small. He posits a dense Q-matrix with a low level of misspecification and a large number of items the effects of misspecification are masked.

The results of this study support the research findings of Baker (1993) and Cassuto (1996) that the density of the Q-matrix was an important factor affecting the parameter estimates. This suggests that researchers conducting LLTM analysis when choosing items for an assessment should be careful not to include too many items on the assessment that have multiple concepts in order to maintain a sparser specified Q-matrix. In other words, if there are 40 items on an assessment the researcher should be careful to select items that have one or two concepts per item rather than 4 or 5 concepts per item. In this way, the resulting Q-matrix will be more sparsely specified and the cognitive component parameter estimates should contain less bias.

### **Misspecification Interacting with Number of Items**

As the percent of misspecification increased the estimated mean bias values for item difficulty, and person ability increased as the number of items increased. In other words, the impact of percent of misspecification on mean bias in person ability and item difficulty is greater

for longer tests. When 40 items are present on an assessment person ability demonstrates tolerable error within acceptable limits when the Q-matrix is misspecified up to 5%, and for item difficulty tolerable error is within acceptable limits for 20 items when the Q-matrix is misspecified up to 5%. That said in all the models presented in the literature review the cognitive components were the parameters which were recovered not item difficulty or person ability; therefore, while this is an interesting result its application for real data sets is marginal.

Test length is an important design factor studied in the literature (Baker, 1993, Cassuto, 1996; Kunina-Habernicht, Rupp & Wilhelm, 2012; Rupp & Templin, 2009). The literature does not provide guidance on the number of items that should be used when employing LLTM modeling.

### **Misspecification Interacting with Skewness of the Person Distribution**

A slight skewness of -0.5 and 0.5 with a kurtosis of 2.0 had no practically significant effect on any design factors for cognitive components, item difficulty, and person ability. In other words, when specifying LLTM researchers can have confidence that the recovery of parameter estimates is robust to a slight skewness of the person distribution. It is well known that the normality of the distribution is a strong assumption of IRT. These results are encouraging given that -0.5 to 0.5 skewness of the person distribution represents a majority of distributions that educational researchers are likely to encounter; however, more research is required to determine what levels of skewness and kurtosis LLTM can accommodate and still accurately recover cognitive components, item difficulty, and person ability parameter estimates.

### **Conclusions**

Misspecification of the Q-matrix has a profound effect on the parameter estimates in LLTM modeling; that said, at all levels of misspecification a majority of cognitive components had mean bias values within tolerable error limits when the Q-matrix was misspecified up to 15%.



The statistical program, SAS 9.3, worked superbly and converged in all simulation conditions, for all key conditions without exception. Researchers are encouraged to take full advantage of the power of SAS as they contemplate conducting LLTM research.

Under-specified Q-matrices can lead to cognitive components with less bias compared to when the Q-matrix is balanced-misspecified or over-specified Q-matrices.

Sparsely specified Q-matrices can produce cognitive components with less bias.

Misspecification of the Q-matrix does not impact the precision, indexed as confidence interval width, of the LLTM parameter estimates.

Researchers are strongly advised to employ content expert groups and rigorous research design methodologies when specifying Q-matrices as the best control for the deleterious effects of estimated parameter bias.

### **Limitations of the Study and Future Research**

The LLTM was examined to determine if the model works well in the simulation world. It is not known how these models will function when subjected to the demands of wide scale state and district K-12 standardized testing. This is a rich area for future research.

The cognitive components parameter estimates ranged from .03 to 1.2 in this study. Analyzing the behavior of cognitive components ranging between  $\pm 2.00$  and possibly as much as  $\pm 4.0$  is a rich area of future research.

The skewness of the person ability distribution was simulated  $\pm 0.50$  with a kurtosis of  $\pm 2.0$ . It is not known how robust the LLTM parameter estimates are as the skewness of the person ability distribution and kurtosis increases. This is a rich area of future research

Confidence intervals are constructed in SAS using the Wald-type confidence interval; however, this statistic produced CI coverage results, especially when the sample size varied, which raised questions about the functioning of the statistic. Further research needs to be conducted on the Wald-type confidence interval when employed with nested models such as

the LLTM. It may be that a bootstrapping approach to the creation of confidence intervals is a better approach.

The person ability parameter estimates are computed employing an empirical Bayes approaches which begs the question of whether or not other empirical Bayes estimation algorithms might be more robust to percent of misspecification of the Q-matrix.

## REFERENCES

- Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automated Control*, 19, 716–723. doi:10.1109/TAC.1974.1100705
- Andrich, D. (1985). A latent-trait model for items with response dependencies: Implications for test construction and analysis. In S. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 245-275). New York, New York: Academic Press, Inc.
- Baker, F. (1993) Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement*, 17, 201-210.
- Cassuto, N. (1996). *The performance of the linear logistic test model under different testing conditions* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9711393)
- Cisse, D. (1995). *Modeling children's performance on arithmetic word problems with the linear logistic test model* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. NN95165)
- Chan, W.-H., Leu, Y.-C., & Chen, C.-M. (2007). Exploring group-wise conceptual deficiencies of fractions for fifth and sixth graders in Taiwan. *Journal of Experimental Education*, 76, 26-57.
- Chen, Y., MacDonald, G., & Leu, Y. (2011). Validating cognitive sources of mathematics item difficulty: Application of the LLTM to fraction conceptual items. *The International Journal of Education and Psychological Assessment*, 7, 74-93.
- Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, 37, 598-618. DOI: 10.1177/0146621613488436.

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533-559.
- De Boeck, P., & Wilson, M. (2004). *Statistics for social science and public policy: Explanatory Item response models*. New York, New York: Springer.
- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*, 1-28.
- de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*, 343-362.
- Dimitrov, D. (1996). *Cognitive item subordinations in linear logistic test modeling* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9634986)
- Dimitrov, D. (2007). Least squares distance method of cognitive validation and analysis for binary items using their item response theory parameters. *Applied Psychological Measurement*, *31*, 367-387.
- Draney, K., & Wilson, M. (2008). A LLTM approach to the examination of teachers' ratings of classroom assessment tasks. *Psychological Science Quarterly*, *50*, 417-432.
- Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika*, *80*, 3-26.
- Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*, 54-75.
- Embretson, S. (1992). Measuring and validating cognitive modifiability as an ability: A study in the spatial domain. *Journal of Educational Measurement*, *29*, 25-50.

- Embretson, S. (1995). A model for linking individual learning to processes and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement*, 32, 277-294.
- Embretson, S. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, 64, 407-433.
- Embretson, S. (2001). Measuring and validating cognitive modifiability as an ability: A study in the spatial domain. *Journal of Educational Measurement*, 29, 25-50.
- Embretson, S., & Daniels, R. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly*, 50, 328-344.
- Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, 38, 343-368.
- Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Embretson, S., & Schneider, L. (1981). Information structure for geometric analogies: A test theory approach. *Applied Psychological Measurement*, 5, 383-397.
- Embretson, S., & Wetzel, D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, 11, 175-193.
- Embretson, S., & Yang, X. (2006). Multicomponent latent trait models for complex tasks. *Journal of Applied Measurement*, 7, 335-350.
- Fleishman, A. (1978). A method for simulating non-normal distributions. *Psychometrika*, 4, 521-532.
- Fischer, G. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359-374.
- Gorin, J. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, 43, 351-373.

- Green, F., & Smith, R. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics, 12*, 369-381.
- Holling, H., Blank, H., Kuchenbacker, K., & Kuhn, J. (2008). Rule-based item design of statistical word problems: A review and first implementation. *Psychology Science Quarterly, 50*, 363-378.
- Hohensinn, C., Kubinger, K., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessments using the linear logistic test model. *Psychological Science Quarterly, 50*, 391-402.
- Im, S., & Corter, J. (2011). Statistical consequences of attribute misspecification in the rule space method. *Educational and Psychological Measurement, 71*, 712-731.
- Indiana Department of Administration on behalf of the Indiana Department of Education (2012, June). *Solicitation for: PARCC item tryout, field testing, operational form construction, and embedded research*. Request for Proposal retrieved from internet on June 30, 2012 from <http://www.parcconline.org/about-parcc> (RFP has been taken down and is available from author on request).
- Kubinger, K. (2008). On the revival of the Rasch-model based LLTM: From constructing tests using item generating rules to measuring item administration effects. *Psychological Science Quarterly, 50*, 311-327.
- Kubinger, K., Hohensinn, C., Holocher-Ertl, S., & Heuberger, N. (2011). Applying the LLTM for the determination of children's cognitive age-acceleration function. *Psychological Test and Assessment Modeling, 53*, 183-191.
- Kunina-Habenicht, O., Rupp, R., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement, 49*, 59-81.

- Leighton, J., & Gierl, M. (eds.). (2007). *Cognitive diagnostic assessment in education: Theory and applications*. New York, New York: Cambridge University Press.
- Leighton, J., Gierl, M., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement, 41*, 205-236.
- Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measures, 36*, 548-564.
- Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores*. Oxford, England: Addison-Wesley.
- MacDonald, G., & Kromrey, J. (2011). Linear logistic test model: Using SAS® to simulate the decomposition of item difficulty by algorithm, sample size, cognitive component and time to convergence. *Proceedings of the American Statistical Association's Joint Statistical Meeting, Social Statistics Section [CD-ROM]*, Miami, FL.
- MacDonald, G., & Kromrey, J. (2012, October). *The effects of Q-matrix misspecification when employing Proc NL MIXED: A simulation study*. Paper presented at the annual SESUG conference, Durham, NC.
- Medina-Diaz, M. (2009). Analysis of cognitive structure using the linear logistic test model and quadratic assignment. *Applied Psychological Measurement, 17*, 117-130.
- Mispelkamp, H. (1985). *Theoriegeleitete Sprachtestkonstruktion [Theory-based construction of a reading comprehension test] (Unpublished doctoral dissertation)*. University of Düsseldorf, Düsseldorf, Germany.
- National Research Council. (2001). *Adding it up: Helping children learn mathematics*. J. Kilpatrick, J. Swafford, and B. Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

- Neath, I., & Surprenant, A. (2003). *Human Memory* (2nd Ed.). Belmont, CA: Thomson Wadsworth.
- Pedhazur, E. (1997). *Multiple regression in behavioral research* (3<sup>rd</sup> Ed.). Orlando, Florida: Harcourt Brace College Publishers.
- Robey, R., & Barcikowski, R. (1992). Type 1 error and the number of iterations in monte carol studies of robustness. *British Journal of Mathematical and Statistical Psychology*, *45*, 283-288.
- Rupp, A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*, 78-96.
- Scheiblechner, H. (1985). Psychometric models for speed-test construction: The linear exponential model. In S. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 219-244). New York, New York: Academic Press, Inc.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.  
doi:10.1214/aos/1176344136
- Sheehan, K., & Mislevy, R. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement*, *27*, 255-272.
- Solso, R., Maclin, O., & Maclin, K. (2008). *Cognitive psychology* (5th Ed.). Boston, Massachusetts: Pearson.
- Sonnleitner, P. (2008). Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychology Science Quarterly*, *50*, 345-362.
- Spada, H., & Kluwe, R. (1980). Two models of intellectual development and their reference to the theory of Piaget. In R. Kluwe & H. Spada (Eds.), *Developmental model of thinking* (pp. 1-32). New York, New York: Academic Press.



- Stevens, J.(1986). *Applied multivariate statistics for the social sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Tabachnick, B., & Fidell, L. (2001). *Using multivariate statistics* (5<sup>th</sup> ed.). Upper Saddle River, NJ: Pearson Allyn & Bacon.
- Tanzer, N., Gittler, G., & Ellis, B. (1995). Cross-cultural validation of item complexity in a LLTM-calibrated spatial ability test. *European Journal of Psychological Assessment, 11*, 170-183.
- Tatsuoka, K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.
- Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York, New York: Guilford.
- Tuerlinckx, F., Rijmen, F., Molenberghs, G., Verbeke, G., Briggs, D., Van den Noortgate, W., Meulders, M., & De Boeck, P. (2004). Estimation and software. In P. De Boeck, & M. Wilson.(Eds.), *Statistics for social science and public policy: Explanatory Item response models* (pp.343-373). New York, New York: Springer.
- Whitley, S. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479-494.
- Xie, Y., & Wilson, M. (2008). Investigating DIF and extensions using a LLTM approach and also an individual differences approach: An international testing context. *Psychological Science Quarterly, 50*, 403-416.

## APPENDIX A: PROGRAMMING CODE

```
%let directory=%sysfunc(dcreate(George_dissertation,C:\));
%let directory=%sysfunc(dcreate(U_20_D_0_1,C:\George_dissertation\));
libname LLTM 'C:\George_dissertation\U_20_D_0_1';
```

```
%let form = "under";
%let test_Length= 20;
%let density = "dense";
%let skew =0;
%let percent = 1;
```

```
proc printto log=junk;
```

### The Item Matrix, Cognitive Components, Beta LLTM and Theta

```
proc iml;
```

```
N_reps = 1085; * N of samples to generate;
```

```
N_items = 20;
```

```
* Extent of misspecification required (enter zeros to leave model correctly
specified);
```

```
N_zeros = 1; * N of zeros to impose;
```

```
N_ones = 0; * N of ones to impose;
```

```
    * initialize seed for random number generator;
```

```
    seed1=round(1000000*ranuni(0)); * Randomly generate a seed;
```

```
    *seed1 = 12345;
```

```
    call randseed(seed1);
```

```
*20 Dense Item Matrix with 5 components;
```

```
    A = J(N_items,5,0);
```

```
    c1= {1 2 3 4 6 7 8 9 11 12 13 14 16 17 18 19};
```

```
    c2 = {2 3 4 5 7 8 9 10 12 13 14 15 17 18 19 20};
```

```
    c3 = {3 4 5 8 9 10 13 14 15 18 19 20};
```

```
    c4 = {4 5 9 10 14 15 19 20};
```

```
    c5 = {1 5 6 10 11 15 16 20};
```

```
* 40 Dense item matrix with 5 components;
```

```
    /*A = J(N_items,5,0);
```

```
32 33 34 36 37 38 39};
```

```
32 33 34 35 37 38 39 40};
```

```
40};
```

```
    c4 = {4 5 9 10 14 15 19 20 24 25 29 30 34 35 39 40};
```

```
    c5 = {1 5 6 10 11 15 16 20 21 25 26 30 31 35 36 40};*/
```

```
*60 Dense Item matrix with 5 components;
```

```

/*A = J(N_items,5,0);
*A[,1] = 1;
c1= {1 2 3 4 6 7 8 9 11 12 13 14 16 17 18 19 21 22 23 24 26 27 28 29 31
32 33 34 36 37 38 39 41 42 43 44 46 47 48 49 51 52 53 54 56 57 58 59};
c2 = {2 3 4 5 7 8 9 10 12 13 14 15 17 18 19 20 22 23 24 25 27 28 29 30
32 33 34 35 37 38 39 40 42 43 44 45 47 48 49 50 52 53 54 55 57 58 59 60};
c3 = {3 4 5 8 9 10 13 14 15 18 19 20 23 24 25 28 29 30 33 34 35 38 39
40 43 44 45 48 49 50 53 54 55 58 59 60};
c4 = {4 5 9 10 14 15 19 20 24 25 29 30 34 35 39 40 44 45 49 50 54 55 59
60};
c5 = {1 5 6 10 11 15 16 20 21 25 26 30 31 35 36 40 41 45 46 50 51 55 56
60};*/

*20 Sparse item matrix with 5 components;
/*A = J(N_items,5,0);
*A[,1] = 1;
c1= {1 2 3 6 7 8 11 12 13 16 17 18 };
c2 = {2 3 7 8 10 12 13 15 17 18};
c3 = {3 5 8 9 13 14 19 20};
c4 = {1 4 5 10 12 15 16 20};
c5 = {1 4 6 9 11 14 19 20};*/

* 40 Sparse item matrix with 5 components;
/*A = J(N_items,5,0);
*A[,1] = 1;
c1= {1 2 3 6 7 8 11 12 13 16 17 18 21 22 23 26 27 28 31 32 33 36 37
38};
c2 = {2 3 7 8 10 12 13 15 17 18 22 23 27 28 30 32 33 35 37 38};
c3 = {3 5 8 9 13 14 19 20 23 25 28 29 33 34 39 40};
c4 = {1 4 5 10 12 15 16 20 21 24 25 30 32 35 36 40};
c5 = {1 4 6 9 11 14 19 20 21 24 26 29 31 34 39 40};*/

*60 Sparse item matrix with 5 components;
/*A = J(N_items,5,0);
*A[,1] = 1;
c1= {1 2 3 6 7 8 11 12 13 16 17 18 21 22 23 26 27 28 31 32 33 36 37 38
41 42 43 46 47 48 51 52 53 56 57 58};
c2 = {2 3 7 8 10 12 13 15 17 18 22 23 27 28 30 32 33 35 37 38 42 43 47
48 50 52 53 55 57 58};
c3 = {3 5 8 9 13 14 19 20 23 25 28 29 33 34 39 40 43 45 48 49 53 54 59
60};
c4 = {1 4 5 10 12 15 16 20 21 24 25 30 32 35 36 40 41 44 45 50 52 55 56
60};
c5 = {1 4 6 9 11 14 19 20 21 24 26 29 31 34 39 40 41 44 46 49 51 54 59
60};*/

do i = 1 to ncol(c1);
    A[c1[1,i],1] = 1;
end;
do i = 1 to ncol(c2);
    A[c2[1,i],2] = 1;
end;
do i = 1 to ncol(c3);
    A[c3[1,i],3] = 1;
end;
do i = 1 to ncol(c4);

```

```

        A[c4[1,i],4] = 1;
    end;
do i = 1 to ncol(c5);
    A[c5[1,i],5] = 1;
end;
*print 'Guttman Matrix:' A;

corrA = corr (A);
varnames = {"CC1" "CC2" "CC3" "CC4" "CC5"};
mattrib corrA      rowname=varnames colname=varnames;
*print corrA;

ccname = {"CC1" "CC2" "CC3" "CC4" "CC5"};
create Corr_CCs from corrA [colname=ccname];
append from corrA;

B = {.18, .42, .03, .65, 1.2};
*print 'Parameters for Cognitive Components:' B;
* Send parameters to regular SAS for calculation of bias & RMSE;
ccname = {"TrueB"};
create TrueP from B [colname = ccname];
append from B;

Beta = A*B;
ccname = {"Rasch_D"};
create Rasch from Beta [colname = ccname];
append from Beta;
*print 'Item Difficulties:' Beta;

do sampsize = 1 to 7;
    if sampsize = 1 then N_persons = 20;
    if sampsize = 2 then N_persons = 40;
    if sampsize = 3 then N_persons = 80;
    if sampsize = 4 then N_persons = 160;
    if sampsize = 5 then N_persons = 320;
    if sampsize = 6 then N_persons = 640;
    if sampsize = 7 then N_persons = 1280;

do replication = 1 to N_reps;
* Generate person abilities N(0,1) and item responses;
    theta = J(N_persons,1);

    call randgen(theta, 'NORMAL');
        *theta=(-1*.06416925946524)+(.85011102914029*theta)+
        (.06416925946524*theta##2)+(.04641702467833*theta##3);
        *sk=.5 kr=2.0 for Fleishman power transformation parameters b, c, & d
where theta=(-1*cc)+(bb*theta)+(cc*theta^2)+(dd*theta^3)
        *print 'True Sample Abilities:' theta;

        *theta=(-1*-.06416925946524)+(.85011102914029*theta)+
        (-.06416925946524*theta##2)+(.04641702467833*theta##3);
        *sk= -.5 kr=2.0 for Fleishman power transformation parameters a, b, &
c where theta=(-1*cc)+(bb*theta)+(cc*theta^2)+(dd*theta^3)
        to obtain negative values of skew reverse the signs for c and a
parameters
        *print 'True Sample Abilities:' theta;

```

```

item_matrix = J(N_persons, N_items, 0);
  ranvec = J(1, N_items, 0);
  do row = 1 to N_persons;
    call randgen(ranvec, 'UNIFORM'); * Uniform random number to
compare with prob of correct response;
    do col = 1 to N_items;
      prob = (exp(theta[row,1] - beta[col,1])) / (1 +
(exp(theta[row,1] - beta[col,1])));
      if (prob > ranvec[1,col]) then item_matrix[row,col] = 1;
    end;
  end;

*print item_matrix;

```

### Impose Misspecification on the A Matrix

```

* Impose misspecification on the A matrix;
A_model = A;
col_totals = J(1, N_items, 1) * A_model;
row_totals = A_model * J(5, 1, 1);
*print 'Starting Matrix and Totals';
*print A_model col_totals row_totals;
* +-----+
  Change required number of ones to zeros
* +-----+;
if n_zeros > 0 then do; * Change ones into zeros;
  got_zeros = 0; * Counter for number of changes made;
  zeros_imposed = J(n_zeros, 2, 0); * matrix to track which elements were
changed to zeros;
  do until (got_zeros = n_zeros);
    column = ceil (5*ranuni(0));
    row = ceil (20*ranuni(0));
    if A_model[row, column] = 1 then do;
      if (col_totals[1, column] > 1 & row_totals[row, 1] > 1) then
do; * Only impose zero if the row and column will not subsequently sum to
zero;

      A_model[row, column] = 0;
      got_zeros = got_zeros + 1;
      zeros_imposed[got_zeros, 1] = row;
      zeros_imposed[got_zeros, 2] = column;
      col_totals[1, column] = col_totals[1, column] - 1;
      row_totals[row, 1] = row_totals[row, 1] - 1;
      *print replication row column got_zeros zeros_imposed
col_totals row_totals;
      end;
    end;
  end;
end;
* +-----+
  Change required number of zeros to ones
* +-----+;
if n_ones > 0 then do; * Change zeros into ones;
  got_ones = 0; * Counter for number of changes made;

```

```

ones_imposed = J(n_ones,2,0); * matrix to track which elements were
changed to ones;
do until (got_ones = n_ones);
  column = ceil (5*ranuni(0));
  row = ceil (20*ranuni(0));
  if A_model[row,column] = 0 then do;
    change_back = 0; * Check to be sure zeros are not
changed back into a ones;
    do check = 1 to NROW(zeros_imposed);
      if row = zeros_imposed[check,1] & column =
zeros_imposed[check,2] then change_back = 1;
    end;
    if change_back = 0 then do; * Only impose one if
element has not been previously changed;
      A_model[row,column] = 1;
      got_ones = got_ones + 1;
      ones_imposed[got_ones,1] = row;
      ones_imposed[got_ones,2] = column;
      col_totals[1,column] = col_totals[1,column] + 1;
      row_totals[row,1] = row_totals[row,1] + 1;
      *print replication row column got_ones ones_imposed
col_totals row_totals;
    end;
  end;
end;
end;
end;

```

### Send Generated Data to SAS for Analysis

```

* Send generated data to regular SAS for analysis;
do person = 1 to N_persons;
  score1 = item_matrix[person,];
  sizelabel = repeat(N_persons,N_items);          person_label =
repeat(person,N_items);
  rep_label = repeat(replication,N_items);
  to_SAS =
sizelabel||rep_label||person_label||score1||A||A_model||repeat(theta[person,],
,N_items);
  if (sampsiz = 1 & replication = 1 & person = 1) then do;
    cname = {"N_Persons" "Replication" "Person" "Score1" "a1"
"a2" "a3" "a4" "a5" "Ma1" "Ma2" "Ma3" "Ma4" "Ma5" "true_theta"};
    create combine from to_SAS [colname = cname];
    append from to_SAS;
    free to_SAS;
  end;
  if (sampsiz > 1 | replication > 1 | person > 1) then do;
    setout combine;
    append from to_SAS;
    free to_SAS;
  end;
end; * end the person loop;
end; * end the replication loop;
end; * end the N_Persons loop;
quit;
*proc print data = combine;
*run;

```

```

* These two statements will turn off standard printed output from proc
NLmixed;
filename junk = dummy;
proc printto print =junk;

data combine;
  set combine;
  skew = &skew;
  form = &form;
  percentage = &percent;
  Test_L = &Test_Length;
  Density = &Density;

data Truetheta;
  set combine;
  by N_Persons replication person;
  if first.person;

*proc contents data = truetheta;
  *title 'Contents of truetheta';
*run;

```

### LLTM Specified using Proc NLMIXED

```

Proc NLmixed data=combine tech=dbldog;
by N_Persons replication skew form percentage Test_L Density;
Parms b1-b5=0 sd0=1;
beta= b1*Ma1+b2*Ma2+b3*Ma3+b4*Ma4+b5*Ma5;
ex=exp(theta-beta);
p=ex/(1+ex);
model score1 ~ binary(p);
Random theta ~ normal(0,sd0**2)subject=person OUT = theta_hat;
Estimate 'sd0**2' sd0**2;
ods output Parameters = Parms ParameterEstimates = ParmEst
AdditionalEstimates = AddEst FitStatistics = FitStats ConvergenceStatus =
Converg;

data fit_LogLikelihood;
set fitstats;
if Descr = '-2 Log Likelihood';
Log_Likelihood = Value;

data fit_AIC;
set fitstats;
if Descr = 'AIC (smaller is better)';
AIC = Value;

data fit_BIC;
set fitstats;
if Descr = 'BIC (smaller is better)';
BIC = value;

data fitstats_AICC;
set fitstats;
if Descr = 'AICC (smaller is better)';

```

```

AICC = value;

data theta_hat_fit;
  merge theta_hat fitstats_aicc fit_aic fit_bic fit_loglikelihood;
  by N_Persons Replication;
  drop value descr;

data theta_hat;
  set theta_hat_fit;

data parms_FIT;
  merge parms fitstats_aicc fit_aic fit_bic fit_loglikelihood;
  by N_Persons Replication;
  drop value descr;

data parms;
  set parms_fit;

data addest_fit;
  merge addest fitstats_aicc fit_aic fit_bic fit_loglikelihood;
  by N_Persons Replication;
  drop value descr;

data addest;
  set addest_fit;

data parmest_fit;
  merge parmest fitstats_aicc fit_aic fit_bic fit_loglikelihood;
  by N_Persons Replication;
  drop value descr;

data parmest;
  set parmest_fit;

```

### Cognitive Components Analysis

```

*proc printto print = print; * Turn printing back on;

*proc print data = Parm;
  *title 'Contents of ODS Parameters File';
*proc print data = ParmEst;
  *title 'Contents of ODS ParameterEstimates File';
*proc print data = AddEst;
  *title 'Contents of ODS AdditionalEstimates File';

data TrueP;
  set TrueP;
  if _N_ = 1 then parameter = 'b1';
  if _N_ = 2 then parameter = 'b2';
  if _N_ = 3 then parameter = 'b3';
  if _N_ = 4 then parameter = 'b4';
  if _N_ = 5 then parameter = 'b5';

*proc print data = TrueP;

proc sort data = ParmEst;
  by parameter;
run;

```



```

data ParmEst2;
merge ParmEst TrueP;
by parameter;
if parameter ne 'sd0';
deviation = estimate - TrueB;
dev_square = deviation**2;
CICoverage = 0;
if trueB < upper and trueB > lower then CICoverage = 1;
CIWidth = upper - lower;
proc sort data = ParmEst2;
by N_Persons parameter;
proc means noprint data = ParmEst2;
var deviation dev_square CICoverage CIWidth aicc aic bic
log_likelihood;
by N_Persons skew form percentage Test_L Density;
output out = final mean = bias RMSE CICoverage CIWidth aicc aic bic
log_likelihood n = n_samples ;

data final;
set final;
RMSE = SQRT(RMSE);

*proc print data = final;
*title 'Estimates of Bias and RMSE for Parameter Estimates, CI Coverage and
Width';

data from_IML;
set combine;
if person = 1;
by N_Persons replication person;
if first.N_Persons | first.person then item = 0;
item + 1;
keep N_Persons replication Ma1 - Ma5 item skew form percentage Test_L
Density;

*proc print data = from_IML;
* title 'This is the dataset from_IML';

proc sort data = ParmEst;
by N_Persons replication parameter;

data from_NLMixed;
set ParmEst;
by N_Persons replication;
if parameter = 'b1' then b1 = estimate;
if parameter = 'b2' then b2 = estimate;
if parameter = 'b3' then b3 = estimate;
if parameter = 'b4' then b4 = estimate;
if parameter = 'b5' then b5 = estimate;
retain b1 - b5;
if last.replication;
keep N_Persons replication b1 - b5 skew form percentage Test_L Density aicc
aic bic log_likelihood;

*proc print data = from_NLMixed;
* title 'This is the data set from_NLMixed';

```

```

data beta_hat;
  merge from_IML from_NLMixed;
  by N_Persons replication;
  beta_LLTM = b1*Ma1 + b2*Ma2 + b3*Ma3 + b4*Ma4 + b5*Ma5;

*proc print data = beta_hat;
*   title 'This is the data set beta_hat';

```

### Beta LLTM Analysis

```

data rasch_R;
  set rasch;
  item = _N_;

proc sort data=beta_hat;
  by item;

data BetaEstimated;
  merge beta_hat rasch_R;
  by item;
  deviation = beta_LLTM - Rasch_D;
  dev_square = deviation**2;

proc sort data = BetaEstimated;
  by N_Persons item percentage;

proc means noprint data = BetaEstimated;
  var beta_lltm rasch_D deviation dev_square aicc aic bic log_likelihood;
  by N_Persons percentage test_L form density skew;
  output out = final_beta mean = Beta_lltm Rasch_D bias RMSE aicc aic bic
log_likelihood n = n_samples ;

data final_beta;
  set final_beta;
  RMSE = SQRT(RMSE);

*proc print data = final_beta;
*title 'Estimates of Bias and RMSE for Parameter Estimates, CI Coverage and
Width';

```

### Theta Analysis

```

proc sort data = truetheta;
  by N_Persons person ;

proc sort data=theta_hat;
  by N_Persons person percentage test_L form density skew ;run;

data ThetaEstimated;
  merge truetheta theta_hat;
  by N_Persons person;
  deviation = estimate - true_theta;
  dev_square = deviation**2;
  CICoverage = 0;
  if true_theta < upper and true_theta > lower then CICoverage = 1;

```

```

CIWidth = upper - lower;

proc means noprint data = ThetaEstimated;
    var deviation dev_square CICoverage CIWidth aicc aic bic
log_likelihood;
    by N_Persons percentage test_L form density skew;
    output out = final_theta mean = bias RMSE CICoverage CIWidth aicc aic
bic log_likelihood;

data final_theta;
    set final_theta;
    RMSE = SQRT(RMSE);
*proc print data = final_theta;
    *title 'Estimates of Bias and RMSE for Parameter Estimates, CI Coverage and
Width';

data LLTM.addest; set work.addest;
data LLTM.corr_ccs; set work.corr_ccs;
data LLTM.final; set work.final;
data LLTM.final_beta; set work.final_beta;
data LLTM.final_theta; set work.final_theta;
data LLTM.parmest2; set work.parmest2;
data LLTM.converg; set work.converg;
proc sort data= combine;
    by replication;

proc sort data=work.combine;
    by N_Persons;

proc corr data=combine outp=corr_out;
    var Ma1 Ma2 Ma3 Ma4 Ma5;
    by N_Persons;
    data LLTM.corr_out; set work.corr_out;

run;

```

## ABOUT THE AUTHOR

George MacDonald is a native of Sydney, Nova Scotia, Canada and is the son of Rev. Donald J. MacDonald and Mrs. Marjorie A. MacDonald (nee Lane). He completed a B.A. in philosophy and English literature at Mount Alison University in 1979, a M.Div. from the Atlantic School of Theology in 1982, and a B.A. *Summa Cum Laude* in psychology from St. Leo University in 2009, where he was named student of the year in the college of Arts and Sciences. MacDonald was ordained a member of the Order of Ministry in the Maritime Conference of the United Church in Canada in 1982 and served for twenty-five years. MacDonald became a graduate research assistant in the David C. Anchin Center in 2008, was jointly named the inaugural Tampa Bay Educational Partnership (TBEP) fellow in 2009 through the Anchin Center and Hillsborough County Public Schools (HCPS), and accepted the position of Assistant Director for Research and Grant Development in the Anchin Center in 2010.