THOMAS, CHERYL, Ph.D. A Comparison of Traditional Test Blueprinting to Assessment Engineering in a Large Scale Assessment Context. (2016)
Directed by Dr. Richard M. Luecht & Dr. Terry Ackerman. 170 pp.

This dissertation investigates the plausibility of computing Assessment Engineering cognitive task model derived difficulty parameters through careful engineering design, and to compare the task model derived difficulty with empirical Rasch model 'b' parameter estimates. In addition, this research seeks to examine whether cognitive task model derived difficulty can replace the Rasch Model 'b' parameter estimates for scoring examinees. The study uses real data constituting four assessments from a large-scale testing company. The results of the analysis indicated strong correlations between the task model and the empirical difficulty parameter estimates. While most of the empirical items satisfied the standard requirements of fit, there were several misfitting task model items, however, the task model was able to provide adequate fit for most of the items. Furthermore the proficiency scores for the empirical and the task model matched each other quite well for all of the assessments, showing no differences among the empirical and task model scores. An examination of the standard error statistics showed no differences between the empirical Rasch model and the cognitive task models. Assessment engineering is a new field, therefore very little research exists on comparing assessment engineering cognitive task model derived difficulties to empirical Rasch model parameter estimates. Moreover, the effects of cognitive task model estimate on proficiency scores has not been investigated.

This study showed that through assessment engineering cognitive task modelling design process, it is possible to generate the item difficulty parameters a priori, without the use of any complex data hungry statistical models. For large scale testing companies, this will significantly reduce cost for pilot testing and make available hundreds of items that operate in a psychometrically similar manner. This design process produces difficulty parameters that operate in a similar manner to the statistical difficulty parameters computed in traditional ways using the Rasch model.

# A COMPARISON OF TRADITONAL TEST BLUEPRINTING TO ASSESSMENT ENGINEERING IN A LARGE SCALE ASSESSMENT CONTEXT

by

Cheryl Thomas

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

Greensboro
2016

Approved by

_____
Committee Co-Chair

_____
Committee Co-Chair

To Dr. Richard Luecht

To my sons Duwane and Jon-Anthony

APPROVAL PAGE

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

Committee Co-Chair _____

Committee Co-Chair _____

Committee Members _____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

# ACKNOWLEDGMENTS

I am greatly indebted to Dr. Richard Luecht for his unflagging, invaluable guidance and support throughout this entire process. He continually challenged me to expand my horizon and grow as a researcher. Without his expertise, this dissertation would have been a far-fetched reality.

Thanks to Dr. Terry Ackerman's for his unwavering support and assistance given, particularly in navigating the process of completing this research. I also owe sincere appreciation to Dr. Burke and Dr. Henson whose comprehensive and timely feedback provided the impetus needed to successfully complete my dissertation. To Dr. Sunnassee, I express my gratitude for his personal assistance/support and encouragement.

To the personnel at the College Board for the providing the support needed and your willingness to explore new frontiers that contributes to the growth and development of research in the field of assessment engineering.

I owe sincere gratitude to my family members, Reginald, Duwane and Jon-Anthony. The sacrifices made, support and encouragement contributed greatly to my success. I could not have done it without you.

Sincere appreciation to Mrs. Valeria Caviness for her expertise in formatting my dissertation. Your assistance was invaluable.

Last but not least to Almighty God. Everything I have done or accomplished centers on his presence in my life.

TABLE OF CONTENTS

Page

LIST OF TABLES ..... viii

LIST OF FIGURES ..... ix

CHAPTER

I. INTRODUCTION ..... 1

Background of the Problem ..... 1
Purposes of the Study ..... 6
Research Questions ..... 6
Significance of the Study ..... 7
Definition of Terms ..... 11

II. REVIEW OF LITERATURE ..... 13

Traditional Table of Specification and Test Development ..... 13
Cognitive Task Modeling and Test Development ..... 15
The Development of Cognitive Task Models ..... 18
Rules for Building Cognitive Task Models ..... 23
Task Model Maps ..... 23
Building Task Models That Control Difficulty and Dimensionality ..... 24
Methods of Comparing Cognitive Task Model Derived Difficulty Statistics ..... 26
Studies Comparing Cognitive Task Model Difficulty and Dimensionality ..... 28
Psychometric Analysis of Cognitive Task Models ..... 29
The Role of Cognitive Tasks Models in Large Scale Testing ..... 34
Contemporary Approaches to Test Design and Development ..... 36
Evidence Centered Design ..... 36
Automatic Item Generation ..... 38
Evaluating Sources of Item Difficulty Scales ..... 41
Cognitive Task Model Item Difficulty ..... 41
Item Difficulty Modeling ..... 42
Item Mapping ..... 43
Scale Anchoring ..... 45
Cognitive Task Analysis ..... 46
The Role of Validity in Assessment Engineering ..... 47

III. METHODOLOGY .....49

Reversed Assessment Engineering Cognitive Task Model Grammar Development.....49
Computation of Cognitive Task Model Difficulty Score .....52
Psychometric Analysis of Empirical Data .....55
Analysis of Cognitive Task Models.....56
Statistical Analysis.....57

IV. RESULTS .....59

Computation of Difficulty Parameter Estimates.....60
Relationship between Cognitive Task Model and Empirical Difficulty Parameter.....61
Correlation and R-squared Evaluation.....61
Rasch Model Fit of Task Model and Empirical Difficulties for English Language 2012.....64
Description of Misfitting Item for English Language 2012.....70
Rasch Model Fit of Task Model and Empirical Items for English Language 2013.....79
Description of Misfitting Items for English Language 2013 .....84
Rasch Model Fit of Task Model and Empirical Difficulties for Calculus 2012.....93
Description of Misfitting Items for Calculus 2012 .....98
Rasch Model Fit of Task Model and Empirical Difficulties for Calculus 2013.....104
Description of Misfitting Items for Calculus 2013 .....109
The Effects of Cognitive Task Model on Proficiency Scores.....118

V. CONCLUSIONS AND DISCUSSION .....130

Overview and Summary of Findings .....130
Practical Implications of the Results.....133
Limitations and Future Research .....135

REFERENCES .....137

APPENDIX A. TASK MODEL GRAMMARS FOR ENGLISH LANGUAGE AND CALCULUS.....151

APPENDIX B. SUBJECT MATTER EXPERTS TASK DESCRIPTIVE STATEMENTS .....154

APPENDIX C. TASK MODEL CODING SCHEMA AND SCORING INDICES....158

APPENDIX D. COGNITIVE COMPLEXITY SCORES....163

APPENDIX E. RUBRICS....168

APPENDIX F. WINSTEPS CODE....172

LIST OF TABLES

Page

Table 3.1 Subject Matter Experts Task Descriptive Statements for Calculus .....52

Table 4.1 Correlation and R-Squared between Task Model and Empirical Difficulty.....61

Table 4.2 Summary Statistics across Items for Empirical and Cognitive Task Model for English Language 2012.....65

Table 4.3 Summary Statistics across Persons for Empirical and Cognitive Task Model for English Language 2012.....66

Table 4.4 Summary Statistics of Misfitting Items for English Language 2012.....69

Table 4.5 Summary Statistics across Items for Empirical and Cognitive Task Model for English Language 2013.....79

Table 4.6 Summary Statistics across Persons for Empirical and Cognitive Task Model for English Language 2013.....80

Table 4.7 Summary Statistics of Misfitting Items for English Language 2013.....83

Table 4.8 Summary Statistics across Items for Empirical and Cognitive Task Model for Calculus 2012.....93

Table 4.9 Summary Statistics across Persons for Empirical and Cognitive Task Model for Calculus 2012.....94

Table 4.10 Summary Statistics of Misfitting Items for Calculus 2012.....97

Table 4.11 Summary Statistics across Items for Empirical and Cognitive Task Model for Calculus 2013.....104

Table 4.12 Summary Statistics across Persons for Empirical and Cognitive Task Model for Calculus 2013.....105

Table 4.13 Summary Statistics of Misfitting Items for Calculus 2013.....108

Table 4.14 Average Proficiency Estimates for Empirical and Task Model .....118

LIST OF FIGURES

Page

Figure 1.1 A Representation of the Assessment Engineering System....2

Figure 2.1 Task Model Map of Scale Construction....24

Figure 4.1 Scatterplot of Item Difficulty Estimates for Empirical and Cognitive Task Model for English Language....62

Figure 4.2 Scatterplot of Difficulty Estimates for Empirical and Cognitive Task Model for Calculus....63

Figure 4.3. Scatterplot of Misfitting MS Outfit Items for Empirical and Cognitive Task Model for English Language 2012....67

Figure 4.4. Scatterplot of Misfitting MS Infit Items for Empirical and Cognitive Task Model for English Language 2012....68

Figure 4.5 Empirical Conditional Mean Scores and Expected Response Function Item 10....70

Figure 4.6 Empirical Conditional Mean Scores and Expected Response Function Item 20....71

Figure 4.7 Empirical Conditional Mean Scores and Expected Response Function Item 22....72

Figure 4.8 Empirical Conditional Mean Scores and Expected Response Function Item 24....73

Figure 4.9 Empirical Conditional Mean Scores and Expected Response Function Item 30....74

Figure 4.10 Empirical Conditional Mean Scores and Expected Response Function Item 40....75

Figure 4.11 Empirical Conditional Mean Scores and Expected Response Function Item 41....76

Figure 4.12 Empirical Conditional Mean Scores and Expected Response Function Item 44....77

Figure 4.13 Empirical Conditional Mean Scores and Expected Response Function Item 45 ..... 78

Figure 4.14 Scatterplot of Misfitting MS Outfit Items for Empirical and Cognitive Task Model for English Language 2013 ..... 81

Figure 4.15 Scatterplot of Misfitting MS infit Items for Empirical and Cognitive Task Model for English Language 2013 ..... 82

Figure 4.16 Empirical Conditional Mean Scores and Expected Response Function Item 17 ..... 84

Figure 4.17 Empirical Conditional Mean Scores and Expected Response Function Item 1 ..... 85

Figure 4.18 Empirical Conditional Mean Scores and Expected Response Function Item 17 ..... 86

Figure 4.19 Empirical Conditional Mean Scores and Expected Response Function Item 35 ..... 87

Figure 4.20 Empirical Conditional Mean Scores and Expected Response Function Item 36 ..... 88

Figure 4.21 Empirical Conditional Mean Scores and Expected Response Function Item 41 ..... 89

Figure 4.22 Empirical Conditional Mean Scores and Expected Response Function Item 45 ..... 90

Figure 4.23 Empirical Conditional Mean Scores and Expected Response Function Item 46 ..... 91

Figure 4.24 Empirical Conditional Mean Scores and Expected Response Function Item 54 ..... 92

Figure 4.25 Scatterplot of Misfitting MS Outfit Items for Empirical and Cognitive Task Model for Calculus 2012 ..... 95

Figure 4.26 Scatterplot of Misfitting MS Infit Items for Empirical and Cognitive Task Model for Calculus 2012 ..... 96

Figure 4.27 Empirical Conditional Mean Scores and Expected Response Function Item 5 ..... 98

Figure 4.28 Empirical Conditional Mean Scores and Expected Response Function Item 6 ..... 99

Figure 4.29 Empirical Conditional Mean Scores and Expected Response Function Item 8 ..... 100

Figure 4.30 Empirical Conditional Mean Scores and Expected Response Function Item 11 ..... 101

Figure 4.31 Empirical Conditional Mean Scores and Expected Response Function Item 18 ..... 102

Figure 4.32 Empirical Conditional Mean Scores and Expected Response Function Item 21 ..... 103

Figure 4.33 Scatterplot of Misfitting MS Outfit Items for Empirical and Cognitive Task Model for Calculus 2013 ..... 106

Figure 4.34 Scatterplot of Misfitting MS infit Items for Empirical and Cognitive Task Model for Calculus 2013 ..... 107

Figure 4.35 Empirical Conditional Mean Scores and Expected Response Function Item 3 ..... 109

Figure 4.36 Empirical Conditional Mean Scores and Expected Response Function Item 4 ..... 110

Figure 4.37 Empirical Conditional Mean Scores and Expected Response Function Item 8 ..... 111

Figure 4.38 Empirical Conditional Mean Scores and Expected Response Function Item 11 ..... 112

Figure 4.39 Empirical Conditional Mean Scores and Expected Response Function Item 14 ..... 113

Figure 4.40. Empirical Conditional Mean Scores and Expected Response Function Item 23 ..... 114

Figure 4.41 Empirical Conditional Mean Scores and Expected Response Function Item 24 ..... 115

Figure 4.42 Empirical Conditional Mean Scores and Expected Response Function Item 28 ..... 116

Figure 4.43 Empirical Conditional Mean Scores and Expected Response Function Item 36 ..... 117

Figure 4.44 Empirical Item Person Map for English Language 2012 ..... 120

Figure 4.45 Cognitive Task Model Item Person Map for English Language 2012 ..... 121

Figure 4.46 Empirical Item Person Map for English Language 2013 ..... 122

Figure 4.47 Cognitive Task Model Item Person Map for English Language 2013 ..... 123

Figure 4.48 Empirical Item Person Map for Calculus 2012 ..... 124

Figure 4.49 Cognitive Task Model Item Person Map for Calculus 2012 ..... 125

Figure 4.50 Empirical Item Person Map for Calculus 2013 ..... 126

Figure 4.51 Cognitive Task Model Item Person Map for Calculus 2013 ..... 127

Figure 4.52 Scatterplot of Proficiency Scores for Calculus and English Language 2012 and 2013 ..... 128

# CHAPTER I

# INTRODUCTION

## Background of the Problem

The educational landscape across the United States continues to go through radical, unprecedented changes, geared towards improving the machinery and methodologies through which students learn. A paradigm shift in assessment practices must out of necessity be fostered, adapted and implemented, to accommodate modern advances in cognition, measurement and computer technology, as new integrated modern approaches that have the potential of creating high quality assessments, that are more useful and valid indicators of what students have learned are replacing isolated outmoded methods (Chudowsky & Pelligrino, 2010).

Most notably among these developments is assessment engineering Luecht, (2006b, 2007a, 2008b). It is an approach that can potentially guide and greatly enhance assessment practices and capable of meeting the growing demand for high quality test items, with sufficient psychometric properties. As Bennett (2013) noted, that in order for meaningful changes to occur in education, there must also be comparative accelerated changes in educational assessment, otherwise they will increasingly work against each other.

Gierl & Leighton, (2010) define assessment engineering as: “An innovative approach to measurement where engineering- based principles is used to direct the design

and development as well as the analysis, scoring, and reporting of assessment results". (p. 3). There are four fundamental processes that undergird assessment engineering. They are: (a) construct mapping and evidence modeling; (b) task modeling; (c) template design and validation; and (d) psychometric calibration and scaling (Luecht, 2012; Luecht, Dallas & Steed, 2010).

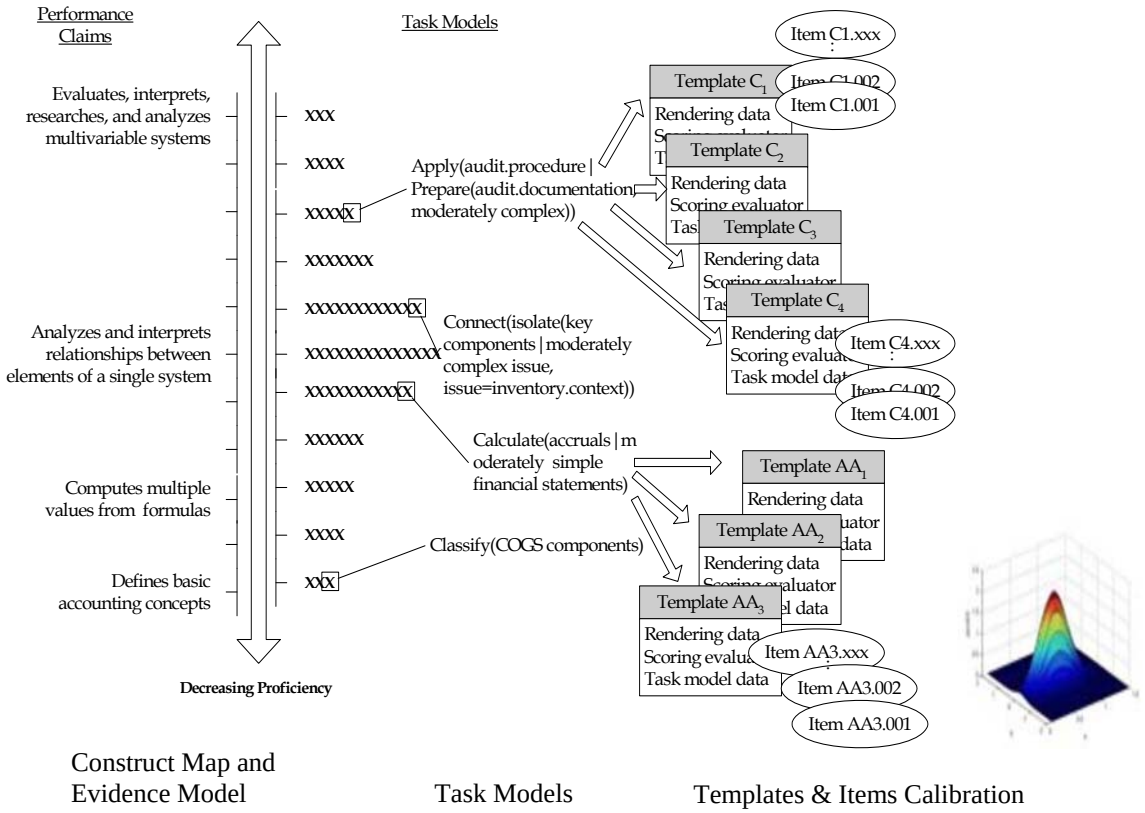



Figure 1.1 A Representation of the Assessment Engineering System

As depicted on the left of Figure 1.1, a construct map represents an ordered list of proficiency claims. It specifies the types of performance-based interpretations that are to be made at different levels of the construct, with higher-level proficiency claims

presuming that the lower-level proficiencies have been mastered (Luecht, 2007c, 2008, 2013). Located in the middle of Figure 1 are the empirically driven evidence models and cognitive task models. The cognitive task models represent skills and/or knowledge-based performance tasks that provide direct evidence about the claims. They are developed at specific levels of each construct, and replace traditional test blueprints and related specifications (Luecht, 2009; Luecht, Burke & Devore, 2009). Each "X" on Figure 1 is representative of a separate task model. The density of task models at different locations of the construct is directly proportional to the psychometric test information needs across the various levels of the construct (Luecht, 2013). Task templates are located on the far right of Figure 1.1. Each template is capable of producing large number of items, which according to Luecht, (2013), 'Share exactly the same cognitive task complexity specifications and perform as isomorphs from a statistical or psychometric perspective' (p. 7). Once validated, templates can be stored and used multiple times to generate new items (Luecht, Burke, & Devore, 2009; Lai, Gierl & Alves, 2010). Finally, to the extreme right of Figure 1.1 is the calibration of the assessment tasks and quality control mechanisms (Luecht, 2013).

Item difficulty index is an important statistic that is used to evaluate the effectiveness of an item or a test. Traditionally, difficulty is treated as an empirical issue, using a post hoc approach, in which the analysis is carried out using data hungry psychometric models. An understanding of the features that influence difficulty, and that valid estimates of item difficulty can be successfully achieved a priori, through reversed assessment engineering is imperative.

Through the use of assessment engineering cognitive task model grammar, cognitive task models can be created with unique design features which captures and controls task difficulty and other psychometric operating characteristics for the family of items, which operate similarly psychometrically and provide interpretations of knowledge and skills, exchangeable at specific location along the scale (Luecht, 2008, 2009, 2012, 2013). Zhou, (2009) defines a cognitive task model as a:

> Generic profile of an assessment task which contains descriptions of knowledge and skills, descriptions of key features (e.g., objects and their properties, variables for difficulty variation) of the task, specifications of task representation material and any required condition, and classifications of response actions returned for scoring. Task models are created at different locations along a construct map and, in turn, each model provides measurement information in a particular region of the construct map. (p. 8)

Each task model therefore represents a group of items in a cognitively meaningful way, incorporating aspects of declarative and procedural knowledge, relevant content and ancillary features, all which go together to directly impact the cognitive complexity of the task (Luecht, Dallas & Steed, 2010). In describing the specific nature of an item, task models incorporate the conditions under which the task is to be completed, the materials presented, and the nature of the work product that will be generated by the examinee (Gorin, 2007).

The Rasch statistical IRT model, a one-parameter logistic model, is often used for analyzing data from assessments to measure the level of difficulty of an item(s) based on a large number of examinees responses. The Rasch analysis is appropriate to make inferences about an examinees ability and the characteristics of an item. The level of

difficulty of the item ranges on a scale of -3 to + 3. Traditionally, the level of difficulty of a particular task is usually not known until after the specific items have been administered to a large number of examinees and estimates of difficulty analyzed. There is also no way of comparing the cognitive complexity across items or to establish the level of complexity that was intended to challenge the examinees (Mislevy, 2006). In addition, the vague content specifications make difficulty targets hard to reach, and most item-level content specifications pay no attention to item difficulty and task complexity attributes that contribute to difficulty. Therefore, it become difficult to understand exactly what is being measured and how best to measure it. Thus construct validity can be supported when the correspondence between the processes measured by items and tests are in alignment with those that were intended by the researcher (Gorin, 2006).

There is also the problem of ascertaining the uncertainty associated with whether items drawn from the same content and have different levels of difficulty are measuring the same complexity of content in relation to knowledge, skills, resource utilization, and context (Luecht, 2009; Lueong, 2006). Hence there is a state of imbalance between content and statistical test specifications.

Assessments engineering cognitive task models are designed to adequately test difficulty levels and also to maintain difficulty or their statistical location along the scale in addition to other psychometric characteristics of the tasks (Luecht, Burke & Devore, 2009). This will allow for test items on an assessment to be designed to test a wide range of complexity and difficulty levels for the entire range of candidates' cognitive skill levels.

The purpose of this research is to use reverse assessment engineering principles to develop cognitive task models, through the use of task model grammars and careful design principles, to determine the difficulty and complexity of the items on two Advanced Placements large scale high stakes assessments. Numerous colleges across the United States use Advanced Placement Tests to check the academic skill levels of entering students, so that they can be appropriately placed in classes at the right level, or may enable them to skip some introductory courses, or highlight areas where more preparatory work is needed.

## Purposes of the Study

The purposes of this research dissertation are:

- To demonstrate the plausibility of computing reverse assessment engineering cognitive task model derived difficulty parameters estimates through careful engineering design.
- Examine the impact of cognitive task models derived estimates and empirical Rasch model estimates on examinee's proficiency scores.

## Research Questions

This dissertation will address the following research questions.

1. To what extent can cognitively task model derived difficulty estimates be compared to statistical empirical Rasch model 'b' parameter estimates?
2. Can assessment engineering cognitive task model derived difficulty estimates replace the Rasch Model 'b' parameter estimates in scoring examinees?

## Significance of the Study

There are tremendous advantages to be gained from using the assessment engineering cognitive task modeling system. The cognitive task modeling process is capable of generating numerous field test items that could mimic existing operational test forms and create tests with established item formats that allow testing to be consistent from year to year (Perea, 2011). This can be very beneficial for large scale testing companies, where considerable numbers of items, are needed to support the development of large item banks. Thus items can be mass produced with known parameter estimates, without the need for pretesting. The result is cost effectiveness and increased potential flexibility for developers (Clauser & Margolis, 2006). In addition, developers are able to meet timely commitment for supplying the demands for summative assessments. At the classroom level, task modeling can produce an assessment tool that greatly reduces the time in which teachers gather and report critical information back to their learners, thus delivering timely formative feedback, identify student's areas of weakness and provide appropriate interventions.

Assessment engineering uniquely designed cognitive task models support proficiency claims, and have a relative ordering in terms of complexity and difficulty along the scale. In addition, changes in the content is also reflected as progress is made up the proficiency scale as the content gets increasingly more complex (Luecht, 2013). This confirmatory, iterative approach to test development is in stark contrast to traditional approaches, where content validity is only established through relevance and representativeness of the distribution of items for content relative to a larger domain

(Messick, 1989). Thus the central issue for assessment engineering is construct validity. This overcome the limitations of the traditional approaches as construct validity provides a strong foundation for test development and score interpretation when descriptions of the cognitive processes and hypothesized relationships among these processes are outlined (Embretson, 1998; Mislevy, 1994; Messick, 1995).

In traditional approaches to testing, item development is primarily a manual process, in which items are individually crafted, and reviewed. Drasgow, Luecht & Bennet (2006) corroborated this when they noted that:

> The demand for large numbers of items is challenging to satisfy because the traditional approach to test development uses the item as the fundamental unit of currency. That is, each item is individually hand-crafted—written, reviewed, revised, edited, entered into a computer, and calibrated—as if no other like it had ever been created before. A second issue with traditional approaches is that it is notoriously hard to hit difficulty targets, which results in having too many items at some levels and not enough at other levels. Finally, the pretesting needed for calibration in adaptive testing programs entails significant cost and effort. (p. 473).

The subject matter experts (SME's) and test developers, the principal decision makers, independently determine what content are to be tested, the design, test specification, quantitative constraints and statistical targets of the tests (Zhou, 2009; Luecht, 2013; Shu, Burke & Luecht, 2010). Content validity is often compromised, as the areas to be tested are prioritized, with most of the content blueprints representing a compromise of priorities within the domain of interest. In addition content validity does not provide any direct evidence that aids in the interpretation of scores or inferences drawn from observable performances on a particular form of the test (Luecht, 2008;

Messick, 1989). Moreover, assessment design and analysis, is one of the areas where many teachers lack formal training. Research conducted by Stiggins, (1999) showed, for example, that fewer than half of the states require competence in assessment for licensure as a teacher. Race, (2009), noted that many assessors struggle in many subject disciplines to make exams valid, reliable and transparent. The continued reliance on traditional blueprinting to generate items lacks a strong system of rules or concrete indicators, necessary for consistently writing or coding items to the content categories (Luecht, 2009).

Assessment engineering cognitive task modeling process provides a detailed design plan for test development in which psychometricians, subject matter experts and cognitive specialists all work together to design the assessment. The items within a family are designed based on well-developed and empirically verified cognitive task models of complexity and specifies the skills needed to perform a particular tasks within that family. In addition, declarative knowledge, auxiliary tools, and overall contextual complexity of the task setting are stipulated (Luecht, 2013). Through its well-designed approach, test developers understand what is being measured and how it is being measured. The processes which contribute to difficulty and complexity are clearly and explicitly laid out. Leighton, (2004) also points out that researchers should be trained in both cognitive psychology and educational measurement as this may be the most valuable resource to the test development industry. This will allow test developers and practitioners to become more educated about theories and methods of cognitive psychology so that these new tools can be utilized to tackle questions of task difficulty

and complexity in addition to a variety of psychological data collection methods such as verbal protocols and interview tools to gather information from these sources (Gorin, 2006).

Finally, assessment engineering can potentially improve the quality of assessments to which examinees are exposed on a regular basis and provide a most efficient and effective method in assessing student's skills and knowledge. Through its goals of supplying an extensive supply of low cost items and to generate one or more well-designed scales that do not require individual item pretesting or data-hungry psychometric models (Luecht, 2006, 2008, 2009, 2012, 2013), it is hoped that the purposes of testing which Fulcher & Davidson (2007), describes as building better tests and gaining a better understanding of what is actually being tested will be accomplished.

Olsen, Olsen & Smith, (2010) noted that assessment engineering task modeling and analysis processes can potentially propel the educational measurement profession forward in very significant and meaningful ways, ultimately influencing what and how Students' learn, drive improvements in education and impact the overall characteristic of assessments. Through its goals of supplying an extensive supply of low cost items and to generate one or more well-designed scales that do not require individual item pretesting or data-hungry psychometric models (Luecht, 2006, 2008, 2009, 2012, 2013).

Leighton, (2004) points out that researchers should be trained in both cognitive psychology and educational measurement as this may be the most valuable resource to the test development industry. Therefore, it is imperative that test developers and practitioners become more educated about theories and methods of cognitive psychology

so that these new tools can be utilized to tackle  questions of task difficulty and complexity in addition to a variety of psychological data collection methods such as verbal protocols and interview tools to gather information from these sources (Gorin, 2006).

## Definition of Terms

The following definitions of terms are provided to ensure clarity and aid understanding of the text:

*Assessment Engineering*: Is an innovative approach to measurement where engineering principles are used to direct the design, scoring, analysis, reporting and implementation of tests (Luecht, 2012).

*Reverse Assessment Engineering*: Describes an analytical process of test creation that begins with the actual test questions and make inferences about the language that drives it, such that equivalent items can be generated that closely relates text complexity that inextricably connects to reading comprehension and supports the fidelity of the reflection process **(**Fulcher & Davidson, 2007**).**

*Task models:* Cognitively oriented specification for a class or family of items that integrates declarative and procedural knowledge components as well as relevant content and auxiliary features that affect the cognitive complexity of the task (Luecht, 2013).

*Test blueprint:* Is a specification of the different quantitative constraints used to assembly a test.

*Task Model Map*: Is a descriptive representation of the precision of key decisions in the corresponding region of the construct map, or richer interpretations of performance in those regions of the scale (Luecht, Dallas & Steed. 2010).

*Isomorphs:* Items generated with the constraint that they all be of the same psychometric attributes (Bejar, 2002).

*Complexity:* Complexity describes the mental operations in which the brain deals with information through different levels of thought processes such as to demonstrate thinking at the levels of Bloom's Taxonomy and Webb's Depth of Knowledge. It relates to the kind of thinking, action, and knowledge needed to answer a question, solve a problem, or complete a task.

*Difficulty*: Item difficulty index refers to the amount of effort that the learner must expend, within a level of complexity in order to accurately respond to an item. It is a characteristic of both the item and the examinee taking the test. Thus it describes the interaction between the learner's mental processes and the item.

# CHAPTER II

# REVIEW OF LITERATURE

## Traditional Table of Specification and Test Development

Luecht, (2013) describes a traditional table of specification (TOS) or content blueprint as 'Consisting of a list of all the relevant topics or standards that adequately represents a particular domain to be assessed. Each topic or standard creates a specification for how many items are to be included on each test form' (p. 3).

Traditionally, a Table of Specification has been commonly used by test developers as the basis for constructing assessments, in order to adequately evaluate examinees on the subject matter that was taught, as well as the appropriate level of cognitive processing, or level of difficulty, to match how the subject matter was taught (Natar, Zuelke, Wilson & Yunker, 2004). Carey, (1988) enumerated six major elements to be considered when developing a Table of Specification: (1) weight or balance of the goals/objectives (2) balance among the levels of learning or levels of taxonomy; (3) the test format; (4) length of the test; (5) the number of test items for each goal and level of learning; and (6) skills selected from each goal framework. Linn and Gronlund (2000) further suggested that the Table of Specification should include the total number of test items and assessment tasks and the percentage allocated to each objective and each area of content. Schmeiser and Welch (2006) indicated that the process should begin by

defining explicitly, the linkage between test purpose and the criterion that defines the test domain.

All of the decisions regarding what are to be tested and how they are to be tested are made by subject matter experts (SME), educators and test developers, the principal decision makers. These decisions are prioritized based on the depth and breadth of the particular domain being tested. Once all major decisions are made, item writing begins, which is usually done in isolation. Messick, (1994) in noting the counter-productivity of this practice asserts:

> One cannot simply construct "good tasks" in isolation, however, and hope that someone down the line will figure out "how to score them." One must design a complex assessment from the very start around the inferences one wants to make, the observations one needs to ground them, the situations that will evoke those observations, and the chain of reasoning that connects them. (p. 40).

Numerous criticisms have been levelled at traditional test blueprinting. These include inadequate content validity, which does not provide any direct evidence to help in the interpretation of scores or inferences drawn from observable test performances (Messick, 1989). This is further compounded by the fact that most content coding schemes are inherently fallible and lack a firm system of rules or concrete indicators which will allows for consistently writing or coding items to the content categories (Messick, 1989; Luecht, 2012, 2013).

Traditional table of specifications often fail to provide any guidance about the intended difficulty and task complexity attributes that contribute to the difficulty of an item within a particular content category. In instances, where the cognitive specifications

are vaguely outlined, they are seldom interpreted or used. Thus, it is never known how complex an item is until after it has been administered to large numbers of examinees and estimates of difficulty computed (Luecht, 2013, 2012). Webb, (1999) noted that in determining the cognitive complexity level of items, it usually involves classifying items into defined categories after the items have been developed. Hence, changes in difficulty are usually attributed to the random outcome of item writers or largely due to the misunderstood features or content that one item writer chooses to include or exclude in the item writing process. There is also no way of comparing the cognitive complexity across items or to determine if that level of complexity is what was intended to challenge the examinees (Mislevy, 2006; Luecht, 2013). Statistical specifications are not usually considered alongside the content blueprint; rather, test forms are built to be statistically parallel. Test forms built from such specifications usually produce scores with only reasonable psychometric quality (Mislevy, 2006; Luecht, 2013).

## Cognitive Task Modeling and Test Development

The cognitive task modeling process provides a different and far more detailed approach to design test specifications. It embodies a system of elaborate cognitive specifications that details every assessment task in a way that takes into account depth-of knowledge, required cognitive skills, cognitive load, auxiliary information and subject-specific content and contexts, to describe a family of items that present similar challenges to the examinee and behave in a statistically equivalent manner (Burke, Devore & Stopek, 2013; Luecht, 2013). Each Task Model is very clearly and carefully outlined so as to assess the observable behaviors that are informative about the latent characteristics

of the examinee along the proficiency scale (Hendrickson, Huff and Luecht, 2012). According to Luecht, (2013):

> Task models incorporates relevant procedural skill requirements of the task, content representation and declarative knowledge load, and location along a scale, it is an integrated specification for a family of items that presents the same type of task challenges to every examinee under the same conditions, and where every item generated from the tsk model has approximately the same relative difficulty (p. 65).

In designing the task model, it is important to meaningfully capture the purpose, use and constraints of the assessment task; the ways feature variations affect task complexity and item difficulty, construct relevance, and evidentiary focus (Luecht, 2013). Due consideration is also given to the content domain and cognitive levels, feature that connect a task with specific knowledge and skills, that describes the context in which examinees response provide data about what they know or can do. This is very important as it provides (a) detailed guidelines to item writers, such that items intended to measure the same content and skills have similar specifications and statistical properties and (b) the inferences made about students are supported by the chain of reasoning from claim, to evidence, to task model, to template, and to item (Luecht, 2013).

A task model can be represented using a task model grammar. Task model grammars give a description of domain-specific explanations of cognitive complexity. The task model grammars control complexity by incrementally changing the difficulty level through the use of cognitive skills statements, which contain the action verbs.

When sources of cognitive complexity are identified, test developers can use them to create items that cover the range of ability within a given achievement level or to create separate task models for each targeted level of complexity (Henderson, Ewing, Kaliski & Huff, 2013). Task Model Grammar statements are designed to replace the traditional dual or three-part content and statistical specifications or traditional content coding currently used by test developers. They provide a formal description of; (a) the combinations of defined declarative knowledge and skills needed to solve the task; (b) the information density and complexity of the task components; (c) auxiliary information or tools that facilitate or complicate the task and; (d) other relevant attributes associated with each component and or set of relations that might affect item difficulty (Luecht, 2006a,2006 b, 2007,2008; Luecht, Burke & Devore, 2009; Luecht, Dallas, Steed, 2010).

Assessment engineering cognitive task models replace traditional content specifications through careful empirically based quality control mechanisms, whereby large numbers of items can be generated manually or automatically, sharing exactly the same cognitive task complexity specifications and perform as statistical isomorphs (Luecht, 2013). A task model test specification that is well-developed and thorough, should translate the construct into items that provide meaningful information about candidates' levels of proficiency.

Gierl and Haladyna (2013) expound that task models serve two important functions. They provide concrete representation of the knowledge and skills that are specified in the construct map; and they define classes of tasks to be performed, by laying

out the knowledge, skills and abilities required to solve any type of task within a specific class.

## The Development of Cognitive Task Models

Cognitive task models development is an iterative process that is based on the use of either strong or weak theory. Strong theory gives a detailed description of the variables specified in the theory that affects examinee performance, and specifies item difficulty features. Weak theory presents a very practical approach to cognitive task modeling and uses design guidelines rather than design principles garnered from a combination of experience, luck, intuition, and research for its development (Drasgow, Luecht, & Bennett, 2006). Gierl and Lai, (2013) have successfully used weak theory to generate item models for Grade 9 Science and Grade 3 Mathematics assessments.

There are two approaches to developing cognitive task models. They are the Construct mapping approach or top-down and the reverse-engineering approach or bottom-up. In the top-down approach, cognitive task models are created directly from evidence models. Hendrickson, Huff, & Luecht (2010), provide the following guidelines for building cognitive task models.

- Specify claim(s) and feature(s).
- Evidence statements inherent in claim(s) are targeted at intended levels of cognitive complexity or achievement levels.
- Task models are created from evidence statements targeted at different levels of achievement, to maintain the idea of ordered task models.

- Decisions pertaining to the items are finally addressed

All ambiguity in the claims are removed and redefined as observable evidence. The level of cognitive complexity or proficiency of the claim at the evidence statement level is identified to determine where the task models for assessing the claim will be located. Task models developed for a certain claim or evidence statement can be nested within an achievement level to reflect those claims and evidence, targeted at that specific achievement level. This allows score interpretation for the assessment to be richer and supported by a stronger validity argument (Henderson, Ewing, Kaliski & Huff, 2013).

When the features that impact the complexity of the tasks are articulated and the sources of cognitive complexity identified, test developers can use them to create items that cover the range of ability within a given achievement level or to create separate task models for each targeted level of complexity (Henderson, Ewing, Kaliski & Huff, 2013).

In developing cognitive task models using reverse-engineering or bottom-up approach as utilized in this research, the models are created using cognitive task model grammars. The task models grammars are described as cognitively oriented statements that incorporate all of the salient procedural, declarative, and contextual aspects of a task model Luecht (2006a, 2006b, 2007, 2008; Luecht, Burke and Devore, 2009). Cognitive task model grammar statements are used descriptively to give an explicit description of: (a) combination of cognitive skills needed to solve the task; (b) declarative knowledge components used to challenge examinees in that region of the scale; (c) information density and complexity of the task components; (d) auxiliary information, resources or

tools that facilitate or complicate the task; and (e) other relevant properties and/or set of relations that might affect item difficulty (Luecht, Dallas & Steed, 2010). The task model skills statements are specified using action verbs. The action verbs are carefully selected, unambiguously and meaningfully defined to give a description of the procedural skills that the examinee uses to complete a task. Action verbs must also maintain their location relative to other action verbs along the trajectory or scale.

The following steps outlined by Luecht, (2006, 2007, 2008); Luecht, Burke & Devore, (2009), are essential for developing reverse engineering or bottom-up cognitive task models through the use of cognitive task model grammars.

1. Develop task descriptive statements that contain appropriate action verbs which must be unambiguously defined.
2. The creation of task model grammar statements is next. The statement are iteratively generated to reflect four cognitive dimension, which constitute the foundation for building the cognitive task models. They are: (a) the relative cognitive level and extent of actions required to complete each task (2) the information density, including the amount and complexity of the data to be considered, classified, manipulated, or analyzed; and (3) the complexity of the context and auxiliary information, tools that might facilitate or hamper the examinee in completing the task, and
3. The count of apparent cognitive actions required to complete the task.
4. Derived task-models are iteratively built to reflect the cognitive complexity of the task.

As cognitive complexity increases along any of the four dimensions, there will be corresponding increase or decrease in difficulty and operating characteristics of the assessment tasks. The four cognitive dimensions are linked to the empirical location of a particular task model on a scale. The location of the tasks along the scale can be changed by systematically altering the cognitive dimensions through manipulation of the task-model template features and components. Ultimately, the task model acquires a central location on a scale by specifying and controlling the cognitive level of the action(s) or manipulation(s) required by each task model, controlling information density and context complexity by manipulating the number and complexity of knowledge objects for each task, as well as identifying key properties of the objects relevant to the task and constraining the number and properties of the relationships among objects, and by constraining the use of auxiliary tools and facilitative task components. Features that do not relate to changes in location are ignorable and exchangeable. Task model grammar can be used to represent each assessment task based on the four cognitive dimensions (Luecht, 2007, 2008a, 2008b, 2009).

A task model grammar has five components. They are: (a) actions or skills; (b) variables or variable sets used in the problems; (c) properties of the values assigned to variables or constants; (d) design factors for the distractors; and (e) verbal/information load of the context or problem. These task model grammar components range from simple to complex and requires continual review and refinement content experts and item designers to ensure understanding, consistent use, and the ongoing utility of the elements (Luecht, 2013).

Task model grammars can be represented as procedural skills component and declarative skill components. Procedural skill component uses (a) primitive actions verbs or skills, and have a fixed interpretation of the required skill location along the scale and (b) skill constructions or complex skills that combine lower-level primitive clauses to form higher-order clauses. Declarative knowledge components can range from simple concepts to complex systems of knowledge components that are linked together by complex relations. To make the task more challenging for the examinee, and increase the overall information density, it is imperative to increase the cognitive processing loads in working memory, by using more knowledge objects, more complex knowledge objects, more complex relations, and offering fewer auxiliary tools or resources (Luecht, 2013).

A task model grammar can be represented in the predicate calculus form as:

Action $_2$[action $_1$(is.related(object$_1$,object $_2$),object =|context,aux.tools)]

The complexity of the 'actions' can differ in terms of ordered complexity or functionally nested. Complexity of the 'objects' can also differ by assigning more or less relevant properties; or in quantity, so that by including more objects, greater cognitive load and, correspondingly, greater complexity will result. In addition, 'objects' can be made more complex by incorporating 'relations' that can vary in type. Properties can also be added that can affect the complexity of the 'relations'. The 'context' and 'auxiliary tools' have important properties and controls that affect the complexity of the task in predictable ways (Luecht, 2009, 2013). Therefore task model grammars replace content blueprints by providing an integrated specification for a family of items that

present similar challenges to every examinee under the same conditions, with every item generated from the task model having approximately the same difficulty (Luecht, 2013).

## Rules for Building Cognitive Task Models

Luecht (2012, 2013) postulates the following rules for building cognitive task models.

- Task models should be incrementally ordered by complexity.
- Task models that are located at the same proficiency level must reflect conjunctive performance
- Higher performance assumes that lower level knowledge and skills have been successfully mastered.

## Task Model Maps

Task model maps represents the distribution of the task models along the proficiency scale, by incorporating difficulty and complexity. The number and concentration or density of task models along the scale is directly proportional to the amount of measurement precision that is needed most at that location along the scale. This provides richer interpretations of performance in areas of the scale that correspond to the regions on the construct map (Luecht, Dallas & Steed, 2010).

Task models differ in location (difficulty) along the proficiency scale. Task model maps that places the concentration of task models closer to the lower end of the map as seen in Figure 4, may denote minimum competency, conversely, those at the higher end of the map represent mastery decisions or competencies. Task models near the center of

the proficiency scale, either maximizes precision of the mean or of mastery decisions representing average performance. Each dot potentially represents large families of items (Luecht, Dallas & Steed, 2010).


Figure 2.1 Task Model Map of Scale Construction

**Building Task Models That Control Difficulty and Dimensionality**

Task models are developed with inherent features that control difficulty and dimensionality. These specific features includes objects and their properties, nature of the relationships among objects, and cognitive level of the action(s) required by the task (Zhou, 2009). This suggests that: (a) task models can order themselves, thereby controlling difficulty with respect to the construct (b) extraneous nuisance dimensionality is controlled; (c) each task model is capable of creating multiple item models and, in turn, to create multiple items; (d) what information is scored as well as which scoring evaluators are used (Luecht, 2007; Zhou, 2009).

Luecht, (2013) outlined the following steps for building task models that control difficulty and dimensionality:

- Control the number and complexity of key knowledge objects for each task
- Identify the key properties of the objects relevant to the task (facilitative or distractive)
- Control the number of objects to be acted upon or manipulated
- Constrain the number and nature of the relationships among objects
- Specify and control the cognitive level of the action(s) or manipulation(s) required by the task
- Constraining the use of auxiliary tools and facilitative task components
- Explicitly define the nature and nesting of relations among objects
- Explicitly define the nature and hierarchical sequencing of functional clauses

Complexity of the task is impacted by many factors such as the number of communication tasks, more topics; Information density: higher structural density of text or speech samples. Luecht, (2008, 2009; Luecht, Dallas and Steed, 2010) recommend increasing task complexity based on controlling and constraining the knowledge objects; relations; and auxiliary tools or resources.

Empirical studies on how cognitive complexity affects differences in the difficulty of assessments are emerging, such as Mosenthal and Kirsch (2007), who defined three classes of variables that correlate with task difficulty. They are the (a) length and organizational complexity of the materials to which document tasks refer (b) length and organizational complexity of task directives, and (c) difficulty of the task solution process. These features accounted for about 80% of the variance of the IRT task

difficulty parameters. The details of such analyses can help item writers control the difficulty of the tasks they develop (Embretson, 1985).

Scheuneman, Gerritz, and Embretson (1991), examined item difficulty in reading, and found that about 65% of the variance in item difficulties in the reading section of the National Teacher Examination related to variables built around syntactic complexity, semantic content, cognitive demand, and knowledge demand.

### Methods of Comparing Cognitive Task Model Derived Difficulty Statistics

Assessment engineering is well suited to promote and extend the concept of item families, cluster and bundles in its analysis for its task models which are located at different regions of the construct map measurement scale. The different approaches for modeling data involving item families are developed for dichotomous and polytomous data, in order to gain a better understanding of examinee responses. A number of statistical procedures have been identified to calibrate the item models including the linear logistic test model (Embretson & Daniel, 2008), the 2PL-constrained model (Embretson, 1999), the hierarchical IRT model (Glas & van der Linden, 2003), the Bayesian hierarchical model (Johnson &Sinharay , 2005; Sinharay Johnson, & Williamson, 2003), the expected response function approach (Mislevy, Wingersky, & Sheehan, 1994), and the linear item cloning model (Geerlings, van der Linden, & Glas, 2011). These models are used with automatically generated items, which can be divided into several item families whose primary goal is to estimate the family-level model parameters by accounting for the dependence among the items within the families (Sinharay & Johnson, 2013). Item group score have been known to provide a more stable

aggregate score and more theoretically meaningful scoring unit than the individual item (Comrey, 1984).

The Related Siblings Model (RSM), (Glas & van der Linden, 2003, 2001; Johnson & Sinharary, 2005) is highly recommended for multiple choice items, and gives an indication of the variability of sibling items within families. Johnson and Sinharary, (2005), extended the RSM model to incorporate polytomous item families in their formulation, a Bayesian hierarchical model that assumes a separate item response function for each item but relates the siblings' item parameters within a family using a hierarchical component (Glas & van der Linden, 2001). Multiple Choice items were modeled using the 3PL model and Constructed Response items were modeled using the Generalized Partial Credit Model (Muraki, 1992).

The Identical Sibling Model (Hombo & Dresher, 2001) assumes a single response function for all items in a family, where all of the items are treated as a single entity. One criticism is that it is restrictive and does not allow for variation within the item family.

Gierling, Glas & van der Linden, (2011) recommended the use of the Linear Item Cloning Model, which accounts for the association among responses to a common sibling. The model uses the Response Sibling Model for the first two levels and then adds a Linear Logistic Test Model structure for the expected value of the item difficulty parameter of each family.

Wilson & Adams (1995) recommend the use of Rasch models for item bundles where the clusters or bundles of test items are identified by a common stimulus materials, common item stems, common item structures, or common item content.

Research conducted by Olson, Olsen and Smith (2010) for within family or item cluster groups revealed that for each of three test forms, on average, the mean correlations were slightly higher within the cases groups than across the cases by about .04. This can potentially pave the way for calibration of item families and allows for generation of items on the fly from the family structure. Item families should be sufficiently defined so that item difficulty, discrimination and model misfit parameters be applicable to each sibling item drawn or generated from the item family or cluster. When calibrating item family models, the dependency structure inherent among the items from the same item family should be taken into account (Olsen, Olsen & Smith, 2010; Sinharaay, Johnson & Williamson, 2003; Johnson & Sinharay, 2005).

**Studies Comparing Cognitive Task Model Difficulty and Dimensionality**

There are very few empirical studies that have been conducted to compare the results of item parameter statistics for cognitive task model processes. Research conducted by Luecht, Burke and Devore (2009) compared the relative difficulty of the task models with empirical difficulty indicators, based on examinee responses. The results showed a high correlation of .92 between the task-model ordering based on the complexity ratings and coding and the Partial Credit Model (Masters, 1982) proficiency score estimates.

Zhou, (2009) in reviewing select assessment engineering principles for a Certified Public Accountant Examination, noted that the cognitive model development phase, driven by task and item modeling, facilitated accurate and efficient item development.

The benefits lie in its facility to control content representation and to give an indication of the difficulty levels of items (Zhou, 2009).

Luecht, (2013) in utilizing assessment engineering task models and templates for developing Computerized Adaptive Performance Tasks, was able to manipulate various template components in order to correspondingly change the task complexity, information density, context complexity, and facility and utility of auxiliary tools and information. According to Luecht, (2013), assessment engineering is a powerful and very practical approach for implementing self-adapting, complex performance assessment tasks that embodies multiple perspectives of test development, computer systems and software design, as well as psychometrics.

Luecht, Dallas & Steed,( 2010) in mapping out ordered sequence of fixed specification cognitive task models for multidimensional high school formative algebra and reading comprehension assessments, successfully developed task models, through the use of task model grammars, which later helped in developing templates for the items.

These studies confirm the feasibility that cognitive task modeling can potentially be a better way of designing high-quality, replicable, and scalable assessment task with the specification that specifically incorporates ordered complexity (Luecht, Burke & Devore, 2009).

## Psychometric Analysis of Cognitive Task Models

In calibrating the cognitive task models, psychometric models are used in a confirmatory manner to ascertain how well the measurement information supports the scales. The dataset can be analyzed at the item level or at the task model family structure

level. Assessment engineering psychometric calibration procedures for task models and/or templates has obvious advantages which according to Luecht, (2007) are (a) less pretesting, (b) robust parameter estimation, and (c) misfit is minimized if the families are well formed.

The Rasch model is frequently used to analyze data. It is mathematically specified as:

$$Pr\{X_{ni} = 1\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}}$$

$\beta_n$, is the ability of person $n$ and $\delta_i$ is the difficulty of item $i$. $\Pr\{X_{ni} = 1\}$, is the probability of a correct response. Modeling the locations of tasks along the scale, gives an indication of what a person at a given level of proficiency might be expected to do in terms of requirements of the tasks. This is important as it allows for greater interpretation and adds to the meaning of a score. From a cognitive task modeling perspective, this can reduce or even eliminate pretesting meant to estimate item parameters (Mislevy, Sheehan, & Wingersky, 1994).

Geerlings, Glas and van der Linden (2011) used the Linear Item Cloning Model (LICM), a hierarchical IRT model for the calibration of a dataset consisting of items generated by rule-based cloning algorithms. The parameters of the model were estimated using Bayesian framework, with a data-augmented Gibbs sampler. These rules or radicals were modeled as fixed effects whereas the joint effects of all irrelevant item features also called incidentals were modeled as random effects. They concluded that the

model is applicable to situations where items within families are generated by the same rules.

Glas and van der Linden (2003) also conducted three different studies in which families of cloned items were calibrated and administered under the multilevel 3PL IRT model. The results of the study of the item pool calibration accurately confirmed that it is advantageous to model the family structure in data from cloned items by a two-level IRT model with different parameter distributions for each family. They noted that using a hierarchal model accounts for item model structure without ignoring variation among instances. Both Glas and van der Linden (2003) and Sinharay and Johnson (2008) found that the hierarchical model is a better fit to the data. Gierl, Leighton and Hunka (2007) successfully showed how the Attribute Hierarchical model (AHM) can be used to classify examinees responses using the 2PL IRT model into a set of structured attribute patterns associated with different components specified in a cognitive model of task performance. The model facilitated the interpretation of student response patterns with respect to a cognitive model of task performance, and lends itself to modeling examinees performance in quantitative domains where learning is cumulative within topics.

Johnson and Sinharay, (2005) calibrated polytomous item families, using the Related Siblings Model (RSM) and found that the model provides a reasonable way to calibrate item families that allowed for some variation. They used the Markov Chain Monte Carlo (MCMC) algorithm for the Bayesian model, the family score function, and the approximate Bayes factors. Hierarchical Bayesian framework is often used to estimate the task model parameters, by employing hyper-parameters (Luecht, 2013).

Sinharay, Johnson and Williamson (2003) also applied the Related Siblings Model (RSM), to fit the hierarchical model using the Markov Chain Monte Carlo (MCMC) algorithm (Gelman, Carlin, Stem, & Rubin, 1995; Gilks, Richardson, & Spiegelhalter, 1996). The hierarchical model assumes that for each item family an expected response function (FERF) gives the probability of a correct response to an item randomly generated from the item family for a given examinee ability. They found that it may be possible to calibrate the item family once without calibrating those items in the future. The RSM took into account the dependency among the items belonging to the same item family.

Williamson, Johnson, Sinharay & Bejar, (2002b) explored the application of hierarchical model calibration as a means of reducing, if not eliminating, the need for pretesting of automatically generated items from a common item model prior to operational use. They applied the Related Siblings Model to mathematics item data to explore the application of the model for calibrating operational data incorporating multiple items generated both from automatic item generation (AIG) and manual item generation. While some item families demonstrated some variability in Item characteristic curves (ICC's), many others were very similar and approximated the ICC consistency observed in families that used the same item repeatedly on each form. In cases where the variations in ICCs for item families were consistently similar to those obtained from recalibration of the same multiple-choice item over repeated administrations , the evidence suggests that model generated item models can be

leveraged to produce multiple parallel items that have highly similar statistical properties (Williamson, Johnson, Sinharay and Bejar, 2002b).

Michel (2007), used cluster analysis to assist in the interpretation of quantitative item models. The item models were based on items from an operationalized GRE test and generated instances calibrated using a 3PL IRT model. The dendograms produced in the analysis can be used to identify related clusters of instances which might not be easily identified by simple inspection of the item parameter estimates.

Luecht, ( 2009) demonstrated through the use of hierarchical Bayesian calibration framework that multiple instances of nodes can be created, calibrated and scored from a particular template and used interchangeably, for the Computerized Adaptive Performance Task (CAPT). Each node or collection of nodes can be calibrated as a class or family, having similar underlying distribution of item parameter estimates. The cognitive task models and templates, used were manipulated to produce changes in task complexity, information density, context complexity, and facility and utility of auxiliary tools and information. The logically determined complexity indices have correlated strongly with item difficulty for the CAPT. The nodes provided a convenient way of linking the instantiation of a template to IRT item parameters.

Olsen, Olsen and Smith (2010) modeled and analyzed data from an information technology credentialing and certification test for designing computer databases. The analysis is based on the Rasch, Masters partial credit and confirmatory factor analysis. The results of the correlation analysis within and across case clusters showed slightly larger correlations (0.04) within the cases than across the cases. The Master‟s partial

credit Rasch analysis indicated that six of the nine case clusters displayed acceptable fit between the empirical and modeled item response functions. A comparison of the average Rasch measures at the item level and at the case group level indicated that four cases with individual item average Rasch measures were within measure values of 0.20 of the case group average Rasch measures. The unidimensional component accounted for approximately 25% of the variance in the item responses.

## The Role of Cognitive Task Models in Large Scale Testing

Cognitive task modeling forms the basis for template and item development. Numerous items can be created from the cognitive models which are significantly cost-effective. Isomorphs are created efficiently and effectively in a timely manner. This is particularly advantageous for large scale test developers as item banks can be created quickly which will minimize item exposure, through test administration because larger pools of operational items are available for each test administration.

Large-scale testing companies should have at their disposal a rich collection of high quality items, in an attempt to keep their expenses minimal. Cognitive task modeling can facilitate this process by improving item quality and guide the production of items with similar conceptual and statistical properties, automatically or manually (Bejar, 1996; Johnson & Sinharay, 2005). More recently, research in educational measurement has been given significant attention directed at methods to ensure an adequate and secure supply of items for item pools, particularly for continuous testing environments (Williamson, Johnson, Sinharay & Bejar, 2002b).

The systematic and strategic development approach of assessment engineering will greatly reduce the cost per item, since multiple instances per model are created rather than single instance per content specialist, and the model is continually re-used to yield many test items.

As Luecht, 2012 observed:

> When we stop treating items as high-cost, low use commodities with a limited shelf life, long-term costs decline. For example, if a testing program has historically spent on average $300 US per item (factoring in item writing, pilot testing publication and processing costs), and an assessment engineering template costs $600 US, the latter may not seem worthwhile. However, if the controlled production of items for a task model templates eventually eliminates much of the pilot testing of items and the associated template can generate 400 items, item exposure risks go down by an order of magnitude and costs per item drop to $1.50 US. (p. 34)

The errors that are usually common in traditional item development, such as omissions or additions of words, phrases, or expressing as well as spelling, punctuation, capitalization, item structure, typeface, formatting, and language problems can be reduced or avoided altogether as only specific elements in the stem and options are manipulated across large numbers of items (Luecht, 2013). Extensive field testing of items can be minimized, if not eliminated, since instances generated from the parent model are pre-calibrated and, thus, do not need to be field tested (Gierl, Zhou & Alves, 2008; Zhou, 2009).

By implementing empirically-based cognitive models of learning or assessment engineering item, task modeling and analysis, testing companies will be able to overcome fundamental problems that continue to plague them and so support inferences about

examinee thinking processes. It is only by so doing, that testing companies will be able to create large banks of well validated items in a cost effective manner, with effective and efficient test assemble (National Research Council, 2001; Snow & Lohman, 1989; Leighton & Gierl, 2007).

## Contemporary Approaches to Test Design and Development

### Evidence Centered Design

Evidence centered design (ECD) is an approach to test development that is based on the principles of evidentiary reasoning in the production and delivery of assessments (Mislevy, Steinberg, & Almond, 2003). As a result of Advances in technology and cognitive science, ECD has evolved and made it possible to use evidence-based arguments or claims to describe examinees proficiency, behaviors or more complex performance which can validly demonstrate proficiencies which are to be included in the assessment (Luecht, 2013; Mislevy, Almond & Lukas, 2003). This radical shift in the design models, has produced assessments that are coherent, geared towards gathering complex data, from which inferences about complex students models will be ground out and to gauge complex learning or evaluate complex programs, built on a sound chain of reasoning from observation to inference (Mislevy, Almond and Lukas 2003),

In designing the work product or task, the underlying knowledge and purposes of the test are of paramount importance, as specialists, test developers, statisticians, delivery-process developers, and interface designers all collaborate and coordinate together to produce a conceptual design framework for the elements of the coherent

assessment, at such a level of generality that a broad range of assessment types can be supported (Luecht, 2013; Mislevy, Almond & Lukas, 2003).

The main evidence centered models for design and delivery are the Conceptual Assessment Framework (CAF), whose models lay out the blueprinting for the operational elements of an assessment, and coordinates substantive, statistical and technical details. Thus CAF wields assessment arguments into blueprints for items and tasks, facilitated through the related student models, evidence models and task models (Luecht, 2013, Mislevy, Almond & Lukas, 2003). The four processes delivery architecture of the delivery system select and administer tasks, interact with the examinee to present materials, capture work products and evaluate responses from each task and accummulate evidence across them (Mislevy, Almond & Lukas, 2003). Statistical features of items such as their difficulty are also included and governed by the evidence model.

The ECD task models also represent a families of tasks, and specify the environment in which the student will say, do or produce something and the specifications for the work product (Luecht, 2013; Mislevy; Steinberg & Russell, 1999). Task Models are systematically developed to manipulate the evidential and statistical parameters of the items in an assessment (Luecht, 2013). Tests developers use task features to deliberately manipulate or change the psychometric properties of the assessment formats and the complexity of the tasks. Thus task feature variables play an important role in determining item difficulty, characterizing proficiency and producing

task variants (Almond, Kim, Velasquez, & Shute, 2014). This manipulation is important as it allows the task model feature or variable to share similar features and provide similar evidence regarding claims about the examinees to be determined by the assessment evidential features of the task (Mislevy; Steinberg & Almond, 1999).

The ECD evaluation framework is related to item statistics and intended complexity. Parameter estimates are determined after pilot testing of items generated from a task model. Item difficulty statistics are computed by fitting the IRT models to the data, which should generally align with items intended to assess a specific level of achievement or cognitive complexity (Hendrickson, Ewing, Kaliski & Huff, 2013). Although some variability in the item difficulty statistics for items written to a particular task model is usually expected.

Among the many benefits of ECD, is that it allows for the continuous updating of items and parameters and attempts to drive the assessment process with the construct, rather than by item types (Luecht, 2013). However, software development in which to execute the approach is still not developed. In addition, there is no discussion of design specifications and piloting items (Luecht, 2013).

**Automatic Item Generation**

Automatic item generation (AIG) is a process of test development, facilitated through the use of computer technology to generate test items. It uses a combination of cognitive and psychometric theories to efficiently produce tests that contain high quality items created using computer technology. (Gierl & Haladyna, 2013; Embretson & Yang,

2007; Irvine & Kyllonen, 2002; Luecht, 2013; Gierl, Fung, Lai & Zheng, 2013). (Drasgow, Luecht, & Bennett, 2006). By using a cognitive theory, or strong theory approach, cognitive features are identified at a small grain size or in such details that item features that predict test performance are specified and controlled (Gierl & Lai, 2012). Thus, psychometric models and computer technology developed through AIG are used for predicting item parameters or calibrating automatically generated items, and for implementing these processes (Drasgow, Luecht, & Bennett, 2006).

Modeling the difficulty based on the features of the items, can be accomplished using scale-up, which creates item templates or prototypes with slots that can be filled with exemplars of features, known to elicit particular knowledge and skills in examinees. Such item templates would be expected to allow the on-the-fly generation of many varied items with predictable properties (Glas & van der Linden, 2001; Sinharay & Johnson, 2008).

Test development specialists identify the content, design and create item models, templates or prototypes that highlight the features or elements in the assessment task. The item model elements are then manipulated to generate new items through the use of computer-based algorithms. Thus thousands of new high-quality items can be produced from a single item model, in a short period of time (Gierl & Haladyna, 2013).

According to Drasgow, Luecht and Bennett, (2006), in describing the AIG process, the determinants of item difficulty must also be clearly understood so that each of the generated instances will not need to be calibrated individually. Item parameter

estimates such as the difficulty parameter can be accurately predicted using underlying item-generative rules. This allows for items that are generated based on the same set of rules to have the same psychometric characteristics. These isomorphic items contain comparable content and are exchangeable psychometrically. Hence, the properties that makes an item difficult or easy and what construct the item is supposed to be measuring are clearly understood (Luecht, 2013).

Item statistics such as the item difficulty parameter estimates can be obtained by pilot testing items as part of an operational test administration, with a small sample of examinees or alternatively, by accounting for the variation among the generated items in an item model and, using this information, to estimate item difficulty with a statistical procedure (Drasgow, Luecht and Bennett, 2006). Statistical procedures used to calibrate the item parameters without the need for extensive field or pilot testing includes the linear logistic test model (Fischer, 1973; Embretson & Daniel, 2008), the 2PL-constrained model (Embretson, 1999), the hierarchical IRT model (Glas & van der Linden, 2003), the Bayesian hierarchical model (Sinharay & Johnson, 2005; Sinharay, Johnson, & Williamson, 2003), the expected response function approach (Mislevy, Wingersky, & Sheehan, 1994), and, the linear item cloning model (Geerlings, van der Linden, & Glas, 2011).

Although items that are automatically generated from the same parent model can vary significantly in terms of their psychometric properties. However, items can also be created with almost identical difficulty parameter estimates, or within an item pool with a wide range of item difficulty (Bejar, 1993).

Automatic item generation has many potential benefits, which include more targeted specification sources and levels of item difficulty and the production of large item banks, which improve cost effectiveness and to generate additional items with good psychometric properties automatically (Luecht,2013; Gierl & Haladyna, 2013).

## Evaluating Sources of Item Difficulty Scales

### Cognitive Task Model Item Difficulty

Item difficulty in assessment engineering is accomplished through an iterative process, and takes into account a list of cognitive processes or skills that can be evaluated statistically and iteratively. Item difficulty is a property of both the item and the examinee, and results from the interaction between the respondents and the item. Gorin, (2006) noted that both the respondents and their response time are observable entities that can be statistically evaluated.

Embretson and Daniel (2008) applied the Linear Logistic Latent Trait Model (LLTM) to operational test items, in order to understand their sources of complexity/difficulty. The results showed that the LLTM approach supports the validity of the cognitive model, and accounted for about half the variance of item difficulty. The level of prediction associated with the model was sufficient in selecting items for operational use without further tryout or pretesting (Embretson and Daniel, 2008).

Among the advantages associated with the LLTM is that construct validity at the item level is established Messick (1995). As in cognitive task modeling, validity is supported when the cognitive complexity features are built into the items in the test

design stage and is empirically supported as predicting item psychometric properties, such as item difficulty. In addition, items with different sources of cognitive complexity were generated by varying item features systematically, based on the cognitive model. This can also be accomplished through computer programs which can generate large numbers of items with predictable psychometric properties (Embretson, 1999; Adrensay, Sommer, Gittler & Hergovich, 2006; Luecht, 2013). Finally, score interpretations can also be linked to expectations about an examinee's performance. As with the Rasch Model, item psychometric properties and ability are measured on a common scale, hence, expectations that the examinee solves items with particular psychometric properties can be given.

Gorin, (2005) also investigated the Linear Logistic Latent Trait Model (LLTM; Fischer, 1973) parameter estimates of experimentally manipulated items in order to verify the impact of encoding and decision processes on item difficulty. The results indicated that manipulation of some passage features, such as increased use of negative wording, significantly increases item difficulty in some cases, while altering the order of information presentation in a passage, did not significantly affect item difficulty.

**Item Difficulty Modeling**

Item difficulty modeling is facilitated in assessment engineering through it cognitive task modeling process. Task model grammars make it possible to systematically describe the features of items and their relationship with difficulty. The task model grammars are the fundamentals on which the cognitive task model component that affect changes in item difficulty are built. Thus task models incorporate all the

features that contribute to difficulty and represent the cognitive sources of difficulty of the items. Through the use of reverse assessment engineering, multiple choice items from two large scale assessments were cognitively decomposed using specific guidelines in order to arrive at the item difficulty. The difficulty component could be manipulated through an iterative process in order to exert some control over the difficulty in order to produce reliable and valid difficulty estimates. Research conducted by Gorin & Embretson (2006) identified features of paragraph comprehension items found in the Graduate Record Exam (GRE) that were responsible for changes in item difficulty. The modeled relationship between the features that contribute to item difficulty and how they relate to the model was captured in a simple regression equation.

Item difficulty models can be tested by defining the observable features of an item that can be systematically coded and entered into statistical analyses. The level of these features determine the portion of processing complexity that should drive the difficulty level of the item. Applications of the difficulty model to reading comprehension items from standardized achievement tests explained between 35% to 72% of the variance in item difficulties (Embretson & Wetzel, 1987a; Gorin & Embretson, 2006). The difficulty modeling process is often iterative as item features are added to or removed from the difficulty model based on their contribution to the explanatory power of the model.

**Item Mapping**

Item mapping is a strategy used to identify and describe what students at specified levels of achievement know and are able to do. It has been widely used in educational assessment in areas of standard setting (Wang, Wiser & Newman, 2001; Wang, 2003),

scale anchoring (eg., Gomez, Nash, Schedl, Wright, & Yolkut 2006), and score reporting (e.g., Kirsch, Jungeblut, Jenkins & Kolstad, 1993; Hambleton, 1997). Item mapping is achieved when there is a high degree of 'alignment', which Webb (1999) defined as "the degree to which expectations and assessment are in agreement and serve in conjunction with one another to guide the system towards student learning what they are supposed to know and do" (p. 4).

By Item mapping in assessment engineering is accomplish through its cognitive task models, which outlines what students know and are able to do, in terms of determining the knowledge and skills that students possess. This takes student's performance into alignment by requiring a clear definition of what students know and can do as evidenced by their actual performance. Task models locates items along the test score scale, based on specific criteria outlined by subject matter experts. The IRT Rasch models used in item mapping have student achievement levels and item difficulties on the same scale. Thus allows for items within the examinee's proficiency level along the test score scale to be identified. Hence, having located the items, SMEs can describe in detail, the knowledge and skills required for examinees along the score scale to demonstrate in order to give correct responses to these items. Thus differences in accomplishment or mastery of students at different performance levels across the score scale can be identified. In determination the performance levels and their descriptions, the content being assessed in the test and inferences to be drawn from the scores must be considered.

## Scale Anchoring

Beaton and Allen (1992) describes scale anchoring as statistically using specific characteristics of items to distinguish between successive points along the proficiency scale, and traditionally when subject matter experts use specific items to provide an interpretation of what students at or close to selected scale points know and can do.

Assessment engineering through cognitive task modeling accomplishes scale anchoring by providing explicit descriptive information of what students can or cannot do. This is done by subject matter experts and test developers. Anchoring the scale allows the items to stay in position. In addition, it provides an understanding and interpretation of what students at various specific anchor points along the scale know and can do based on their responses. In the anchoring process, because examinees located at high score levels generally know and can do more than those located at lower levels, items are examined to see how well they differentiates between successive anchor points. The items are reviewed between adjacent anchor points to see whether or not the specific tasks can be generalized to describe the level of proficiency at the anchor points. Two approaches to scale anchoring, identified are the direct method and the smoothing method. These are facilitated through assessment engineering cognitive task modeling and assume that a scale can be generated, either by traditional psychometric or item response theory (IRT) methods. Results of study on Scale anchoring are used to further inform educational experts about what students have learned, and to statistically identify items of interest that provides data for experts to consider and the implications of the identified items (Beaton and Allen, 1992).

## Cognitive Task Analysis

Cognitive *task* analysis (CTA) relates to understanding the different cognitive processes and skills that are required in order to perform a specific tasks. It involves mapping of the task, identifying the critical decision points, clustering, linking, and prioritizing them, and characterizing the strategies used (Klein, 1998).

Cognitive task analysis is conducted in assessment engineering through its cognitive task models where the task performance are captured in detail. It outlines the cognitive knowledge and processes that must be expended by the examinee in order to achieve the required performance. Subject matter experts are involved in providing extensive descriptions, which allows them to develop and analyze the tasks effectively. Detailed information about the procedural and declarative knowledge, thought processes and goal structures that underlie observable task performance are given. According to Chipman, Schraagen, & Shalin, (2000), it captures information about both overt observable behavior and the covert cognitive functions behind it in order to form an integrated whole. Some of the methods used for conducting a cognitive task analysis are the Applied Cognitive Task Analysis (ACTA), the Critical Decision Method (CDM), Skill-Based CTA Framework, Task-Knowledge Structures (TKS), Hierarchical Task Analysis and the Cognitive Function Model (CFM) (Clark, Feldon, van Merriënboer, Yates & Early (2006).

Cognitive Task Analysis in assessment engineering is also performed a priori, before the design of instruction and/or tests. The descriptions are then used to develop expert systems, tests to certify job or task competence, and training for acquiring new and

complex knowledge for attainment of performance goals (Chipman, Schraagen, & Shalin, 2000; Jonassen, Tessmer, & Hannum, 1999).

## The Role of Validity in Assessment Engineering

The validity of score interpretation, use and consequences play an important role in assessment engineering. It provides the basis for investigating Kane's (2013) Interpretive Use Argument, which specifies the proposed interpretations and uses of the test scores. This can be accomplished firstly through the fundamental processes of construct mapping and evidence modeling which specifies claims that form the basis for creating cognitive task models. The claims, which ultimately produce scores or evidence about examinees performances are clearly stated to support score-based interpretation. In developing the assessments, the purpose, which according to Bachman & Palmer, (2010), guides the development of both the test and the Interpretive Use Argument is given primacy.

Secondly the different item types generated, from templates provide vital scorable useful information (Luecht, 2013). Through the different controls and attributes used in item selection, different statistical models such as IRT model are used to summarize the relationships between test taker ability, as indicated by performance on a sample of items, and expected performance on other items that also fit the model, from which item parameters are known (Zumbo, 2007). Under the argument-based approach to validation, generalizability of almost all test score interpretations is vital as almost all of these interpretations involve generalization over some universe of generalization that goes beyond the observations actually made (Kane, 2013).

The validation of a proposed interpretation must be evaluated for coherence, clarity, plausibility and completeness of the IUA, in addition to the evaluation of the plausibility of its inferences and assumptions. Assessment engineering cognitive task modeling is an iterative process which allows for modifications to be made to the design process. During the developmental process, modifications are made to the test design so that the IUA becomes more specific. Modifications are made to resolve any discrepancies, through the iterative process of development and revision which continues until the fit between the test and IUA is acceptable (Kane, 2013). Mislevy (2009), coined the term "assessment design argument" to emphasize the design choices that must be made during test development, all of which are influenced by the proposed interpretation and use.

Construct validity is supported when extensive descriptions are given through the cognitive task models of the cognitive processes and hypothesized relationships among these processes that can provide a stronger foundation for test development and score interpretation (Embretson, 1994; Mislevy, 1994; Messick, 1995). Ferrara and DeMauro (2006) further suggest that specifications of content and procedural knowledge, having a measurement plan describing the nature of the assessment tasks, and hypotheses and evidence of the nomological network of the construct should also be included. Thus when the processes measured by items and tests are those that were intended by the researcher, construct validity can be achieved (Gorin, 2005).

# CHAPTER III

# METHODOLOGY

The Methodology Chapter is presented in four sections: (a) Reversed Assessment Engineering Cognitive Task Model Grammar Development; (b) Cognitive Task Model Derived Difficulty (c) Statistical Analysis of Empirical Data and (d) Analyses of Cognitive Task Models.

## Reversed Assessment Engineering Cognitive Task Model Grammar Development

Reverse engineering cognitive task models are developed by decomposing existing operational items that have been administered to examinees, from two nationally-based large scale examinations. A similar process was used by Luecht, Burke & Devore (2009) to build cognitive task models for a Complex Computer-Based Performance Exercises. For purposes of this research, the Advanced Placement (AP) English Language and Composition and the Calculus BC examinations, are used. The instrument for the Calculus BC consists of forty-five multiple choice questions. The English language and composition assessment, has fifty-four multiple choice items. Both the Calculus BC and English Language and Composition exercises will be subjected to a type of cognitive analysis, in which all of the items will be reverse engineered in order to generate plausible cognitive task models for each task.

The reverse engineering cognitive task models process begins by creating a list of relevant task model grammars or action verbs that are subject specific. The task model grammars (TMG'S) describes the procedural skills and actions that will be manipulated by the examinee to provide evidence about their expected performance at specific regions of the construct map.

Having built a list of all plausible action verbs, the type of skills associated with each verb is explicitly, meaningfully and unambiguously defined. This is important to ensure that each verb always represents the same constellation of skills across tasks associated with the particular construct, and that they maintain their relative difficulty along the scale (Luecht, 2013).

The action verbs are then ranked and categorized into groups labelled low (1), medium (2), and high (3) based on their level of apparent cognitive complexity of the skills contained in each action verb or the task model grammar complexity design components. For example the action verb 'Identify' is classified as 'easy' , given a rating of '1', whereas the action verb 'Interpret' is considered to embody moderate cognitive skill description, and 'Evaluate' is of a more complex description and given a rating of '3' (See Appendix A).

After the action verbs have been defined and categorized, they are meticulously reviewed by the small group of SME's, in order to ensure that their position and categorizations along the scale was plausible and fixed. They then form the basis for developing the cognitive task models.

The next step involves the development of task descriptive statements. These statements describes each task in terms of the required actions of the, the amount and type of data or information that is being manipulated or used and information related to the context and features of the task that might affect item difficulty. The task descriptions are also suggestive of the complexity of each task and the complexity of the information being used/manipulated by the examinees to complete each task (Luecht, Burke & Devore, 2009). The relevant cognitive task model grammars or action verbs are contained in these task descriptive statements. Each descriptive statement was given a difficulty rating that is representative of the level of cognitive complexity of the various required actions and skills contained in each task. In order to ascertain the difficulty of the items, the SME's used a criteria that gave consideration to the explicitness and implicitness of the relations among the stem, the key and the distractors; the semantics and syntactic features such as vocabulary represented in the task or passage; the number of unknowns; the number of steps or count of actions; the information density and the context complexity of the tasks. The difficulty/complexity value so obtained will represent the cognitive complexity of the various actions that will be deduced from the task model grammars.

Table 3.1

Subject Matter Experts Task Descriptive Statements for Calculus.

| **Task Descriptive Statements** | **Perceived Difficulty** |
|---|---|
| Recognize geometric sequence ⎪ Apply formula for geometric sum ⎪ Compute sum ⎪ Simplify fraction | Moderate |
| Recognize implicit equation ⎪ Calculate derivative implicitly ⎪ Substitute pair ⎪ Simplify ⎪ Use point slope | Moderate |
| Rewrite structure ⎪ Apply basic rules for derivatives | Easy |
| Apply definition of absolute value ⎪ Re-draw graph ⎪ Calculate area | Moderate |
| Recognize separable diferential equation ⎪ apply rules for exponent ⎪ Apply rules for integration ⎪ Substitute condition to solve ⎪ Solve equation | Hard |

The entire task modeling process is facilitated by Subject Matter Experts (SME's) or Content Specialists, who meticulously evaluate the cognitive skills that will be needed by the examinee to solve each task correctly, and so determine the level of difficulty, and rate the items according to the features or skills that are associated with the task. For this research a small group of two SME's were used.

## Computation of Cognitive Task Model Difficulty Score

Generally, in order to compute the cognitive task model difficulty score, the task model coding schema and scoring indices are done in which task model and difficulty indices are generated and the coding/scoring of the tasks for cognitive task complexity information density and context complexity must be iteratively created and computed.

The coding schemes and scoring indices will be used to classify each task in terms of three cognitive analysis framework, which are, the required task actions, information

density and context complexity. The coding process for the English and Composition included encoding linguistic features of the task including semantic and syntactic features such as vocabulary and propositional density of the passage. The calculus encoding system followed similar processes of reasoning levels, vocabulary and computation. First, the required task actions represents the cognitive complexity of the various actions that were extrapolated from the SME's task descriptive statements. The difficulty/complexity value is a simple index of apparent cognitive complexity/difficulty given the features of the task where 1=easy/simple, 2=medium and 3=difficulty/complex.

Second, the information density represents how dense the material is that is included in each tack. Scoring specific information on each task is also taken into consideration in deriving the classifications. The difficulty/complexity index here represents the different levels of information that must be managed by the examinee, where '1' for example, reflects relatively simple, uncomplicated entries, values and '3' represents complex sets of features or a system of values to be analyzed. Finally, the lower section shows the complexity of the tasks context, ranging from 1=simple to 3=complex.

The coding/scoring of the tasks uses assigned difficulty/complexity scores following a series of four iterative steps. The four ratings are the average task complexity, count of actions, information density and context complexity. The following equation is representative of the process

$$b_{ij} = \beta_{1i}\,\mathrm{ATC} \times \beta_{2i}\,\mathrm{COA} \times \beta_{3i}\,\mathrm{ID} \times \beta_{4i}\,\mathrm{CC}$$

Where ATC is average task complexity, COA is Count of actions, ID is information density and context complexity.

The average cognitive task complexity is computed by averaging the ratings of the action verbs that were extracted from the SME task descriptive statements across all measurement opportunities of the skills required to complete each task.

The required actions is representative of the various actions embedded in the tasks and extrapolated from the action verbs or the task descriptive statements. The scores or measurement opportunity awarded to each item will also be taken into consideration. The count of actions or apparent steps needed to complete each task, gives the total number of actions that the examinee must take in order to respond correctly to an item.

The information density value represents the different levels of information that must be managed by the examinee. The levels of information range from simple to complex sets of operating systems or values to be analyzed by the examinee. The scores or measurement opportunity awarded to each item will also be taken into consideration.

The final coding deals with the context complexity of each task. The context complexity is computed judging from the context in which the examinee has to work. Whether there are context clues etc.

The complexity score is computed by multiplying the four previous rating values of the average task complexity, count of actions, information density and context complexity together. This computed value is essentially a predicted item difficulty or

complexity parameters. The complexity scores have the effect of positioning each task model in a multidimensional cognitive complexity space. The higher the complexity score, the more complex the item. This Coding scheme replaces formal predicate calculus statements (See Appendix C)

The multiplicative formula used in calculating the derived task model difficulty, the product of the independent variable or predictive components, conceptually allows these components to be interconnected after centering or standardizing the values. Hence the likelihood of having several independent variables together is increased. This greatly expands the understanding of the relationships to the dependent variable in the model.

A difficulty/complexity index score was next computed. This score was used to re-score the examinees. It is computed by multiplying the average cognitive task complexity by the count of actions. The average cognitive task complexity and count of actions are previously computed as indicated above by averaging the ratings of the action verbs that were extracted from the SME task descriptive statements and the count of steps needed to complete each task. This computed value is essentially a predicted item difficulty and represents the overall cognitive complexity of each task. This limits the number of dimension to three.

## Psychometric Analysis of Empirical Data

The empirical response data, is statistically calibrated in order to generate item parameter estimates for each item using the Rasch model computed in Winsteps. The Rasch model is represented mathematically as:

$$Pr\{X_{ni} = 1\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}}$$

Where $\beta_n$represents the ability of person $n$, and $\delta_i$ represents the difficulty of item $i$. The resulting 'b' parameter estimates will be categorized into three pre-determined groups, based on the computed level of the item 'b' parameter estimates.

**Analysis of Cognitive Task Models**

After the cognitive task mode derived difficulty for each of the items is computed, the anchor calibration of the cognitive task models is done in which the task model derived difficulties is subjected to several iterations in order to compute the Pearson Product Moment Correlations Coefficient and allow for comparisons to the empirical parameter estimates. The anchor calibrations involves centering or standardizing the 'b' parameter estimates and then running the anchor calibration. The fixed task model difficulty is then anchored and calibrations are done with the new data and comparisons made relative to the empirical estimates.

The Rasch model is used to test how well the model fits the data to ascertain the goodness-of-fit of the data to the model using the Chi-square statistic ($\chi 2$) which analyzes the fit of the cognitive task model and the Rasch model using both Infit and Outfit statistics computed in Winsteps. The Infit, which is sensitive to misfit of items well-matched to examinee proficiency, and the Outfit, sensitive to outliers or items located further away from examinees proficiency are computed and compared. Both the Infit and the Outfit have expected values of 1.0. Values of between .7 and 1.30 are

considered, to be "acceptable." Of the two statistics, Infit is generally considered to be more relevant since it signals potential misfit for those items best suited for an examinee at a particular proficiency level. The standard deviation of the mean square infit and outfit will be examined and compared. This research proposal addresses two research questions.

## Statistical Analysis

1. To what extent can AE cognitive task model derived 'b' parameter estimates be compared to empirical Rasch Model 'b' parameter statistics?

This research question is based on the premise that cognitive task model derived difficulty indices should yield comparable 'b' parameter estimates as the empirical statistics. Using the Rasch model, to calibrate all of the examinee response data in order to estimate the difficulty parameters of each item. Cognitive task models for the calibrated items with empirical difficulties following different categorizations are created. The empirical difficulties estimates will be regressed unto the task model difficulty design components. Comparison of parameter estimates will then be made of: (a) Difficulty parameter estimates based directly from empirical response data; (b) difficulty parameter estimates computed from the task models following several modifications in order to cross validate the task models.

Other statistics that will be computed are the Pearson Product moment correlation, to analyze the correlation between the assessment engineering task-model and the empirical difficulties; R-square; and the standard errors.

If the cognitively derived task models difficulty yield similar results as the empirical estimates, then the potential exists for assessment engineering to greatly impact the need for large examinee samples for item calibration. This may be feasible since assessment engineering is based directly on intentional and principled designs (Luecht, 2012). Thus, if a task model difficulty estimate can be used for all items generated within the task model family it is no longer necessary to pretest every item and a natural quality control mechanism of managing variation within task model is possible.

2. Can AE cognitive task model derived difficulty estimates replace the Rasch Model 'b' parameter estimates in scoring examinees?

This research question seeks to examine the impact and potential utility of using the task model logically derived difficulty estimates to estimate examinees thetas. The person proficiency estimates will be computed using estimates of each item difficulty based on the empirical response data and derived task models complexity scores.

The examinee's theta estimates will be computed using the Rasch Model in Winsteps. The examinees response data is rescored, based on the overall cognitive task model complexity parameters. Theta estimates are computed and compared. With the Rasch model, both the person ability and the item difficulty are on the same logit scale, and gives an indication if items on the assessments were targeted very well to the examinee population who took the tests. These analyses will also give an indication as to whether there is any difference in scoring examinees empirically or with the task model complexity index scores.

# CHAPTER IV

# RESULTS

This research dissertation was designed to investigate whether Assessment Engineering cognitive task model derived difficulty parameters is comparable to the empirical Rasch model difficulty parameter estimates, using data from two large-scale assessments. In addition, this research seeks to examine the extent to which cognitive task model parameters estimates could be used to score examinee's data. The analyses intend to obtain answers to two research questions: (1) To what extent can cognitively derived task model difficulty parameters be compared to empirically based 'b' parameter estimates? (2) Can AE cognitive task model derived difficulty estimates replace the Rasch Model 'b' parameter estimates in scoring examinees?

Organizationally the results are presented in four sections (a) Computation of difficulty parameters (b) correlation and r-squared evaluation (c) evaluation of model fit and (d) the impact of cognitive task modelling on examinee proficiency scores. Importantly, the 'b' parameters of the cognitive task models are known a-priori and this allows for the manipulations of the test design structure, which can directly influence the difficulty parameters. Essentially, if the cognitive task model derived difficulties matches the Rasch model, it provides substantial evidence in support of continued research and the viability of generating thousands of items without the use of pretesting

and data hungry psychometric models. The benefits and implications to educational testing organizations are immeasurable.

## Computation of Difficulty Parameter Estimates

The empirical Rasch model difficulty parameter estimates were calibrated using Winsteps (Linacre, 2009). Winsteps was also used to compute the fit indices and the examinees proficiency scores. Formal task model grammars, which concretely describe the knowledge and skills contained in the tasks, were used as the foundation for developing the task model coding scheme and coding indices, created to represent the cognitive complexity of the required actions, for which a difficulty/complexity value was assigned. Next the information density or complexity of the materials included in each task, representing the levels of information managed by the examinee and finally the complexity of the context. The cognitive task models complexity scores were computed by multiplying the average cognitive task complexity, the count of actions, the information density and the context complexity. The following equation represents the process

$$b_{ij} = \beta_{1i}\,\mathrm{ATC} \times \beta_{2i}\,\mathrm{COA} \times \beta_{3i}\,\mathrm{ID} \times \beta_{4i}\,\mathrm{CC}$$

Where AC,is average task complexity, COA, Count of actions, ID is information density and CC is context complexity. Anchor calibrations are then performed.

Finally, the cognitive complexity index score, used to compute the proficiency scores, were computed by multiplying the values of the average cognitive task complexity by the number of actions. Anchor calibrations are then performed.

**Relationship between Cognitive Task Model and Empirical Difficulty Parameter Correlation and R-squared Evaluation**

Table 4.1

Correlation and R-Squared between Task Model and Empirical Difficulty

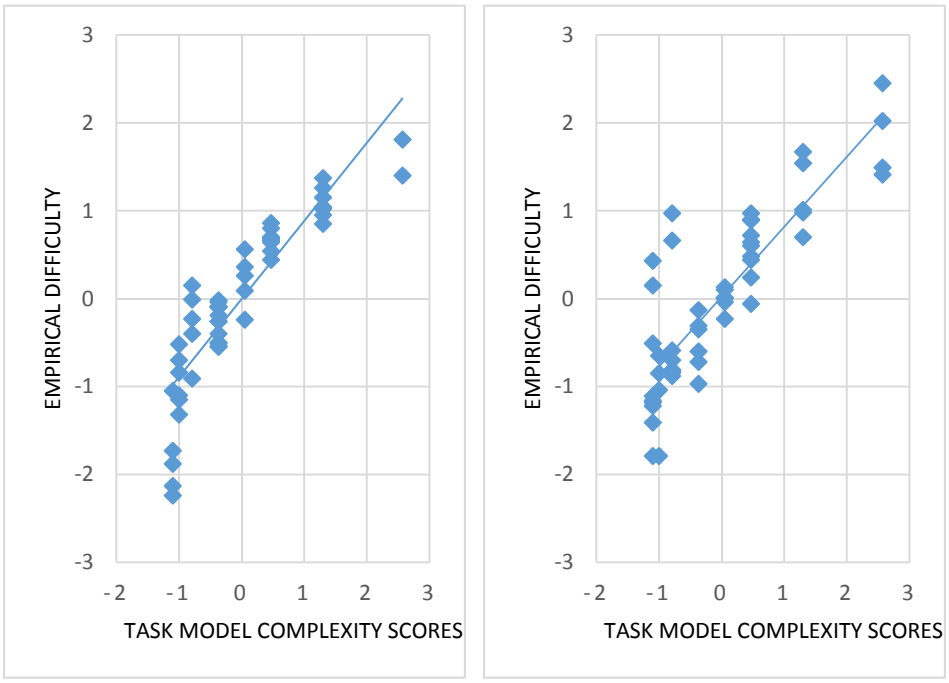
| Assessments | Variables | R | $R^2$ |
|---|---|---|---|
| English 2012 | Task Model<br>Empirical | 0.891 | 0.794 |
| English 2013 | Task Model<br>Empirical | 0.854 | 0.729 |
| Calculus 2012 | Task Model<br>Empirical | 0.886 | 0.785 |
| Calculus 2013 | Task Model<br>Empirical | 0.831 | 0.691 |

Correlation significance level is .001

Table 4.1 shows the Pearson Product Moment Correlations Coefficients and the R-Squared statistics for the cognitive task models and the empirical estimates for English Language and Calculus. The results indicate significantly (p=.001) positive correlational relationships between the empirical difficulties and the task models complexity scores.

The largest correlations is English 2012 (r=0.891) and the smallest Calculus 2013, (r= .831).

The R-Squared values suggest that the model accounts for approximately 70% of the proportion of the variance for all assessments. The subject with the largest proportion of explained variance is English 2012, $r^2$.794 and the smallest, Calculus 2013, $r^2$.691




English 2012 English 2013

Figure 4.1 Scatterplot of Item Difficulty Estimates for Empirical and Cognitive Task Model for English Language.

The scatterplots of item difficulty parameters shown in Figure 4.1 indicate a positive correlational relationship between the empirical difficulty and the

cognitive task model complexity scores. Thus the empirical difficulty and the task model are highly correlated, as most of the parameter difficulty scores align relatively close to the trend line, indicating a general ordering of parameter estimates.


Figure 4.2 Scatterplot of Difficulty Estimates for Empirical and Cognitive Task Model for Calculus.

The scatterplot in Figure 4 depicts a positive relationship between the empirical difficulty and the cognitive task model difficulty complexity scores. The task-model difficulty complexity measures for the Calculus is more varied, however, despite the slight difference the impact on scores differentials is still negligible. The task model

composite difficulties was able to perform in a similar manner as the empirical difficulties as it relates to rank ordering the examinees as shown in Figure 5.

**Rasch Model Fit of Task Model and Empirical Difficulties for English Language 2012**

Chi Square statistics are useful for quantifying the fit of the data to the Rasch Model. Model fit is a statistical procedure used for model validation, and summarizes the discrepancy between observed values and the values expected under the model (Wright & Stone, 1979). Winsteps (Linacre, 2009) computes two fit statistics as part of its calibration process. They are Infit and Outfit (Wright & Stone, 1979). Mean Square Outfit is the sum of squared standardized residuals, and is outlier sensitive, that is, it is more sensitive to unexpected observations by persons on items that are relatively very easy or very hard. Mean Square Infit equals the sum of information-weighted mean square. It is sensitive to inlier-pattern or to unexpected patterns of responses, targeted near the person's ability. Of the two statistics, Infit is more relevant as it indicates potential misfit for items best suited for an examinee at a particular proficiency level. The expected values of the mean-square Infit and Outfit statistics are 1.0. Values between .7 and 1.30 are generally acceptable.

Table 4.3 summarizes the model fit results for the cognitive task model and the empirical Rasch difficulties.

Table 4.2

Summary Statistics across Items for Empirical and Cognitive Task Model for English Language 2012.

| Variables | | N | MEAN | SD | MAX | MIN |
|---|---|---|---|---|---|---|
| Empirical | MS INFIT | 55 | 1.00 | 0.08 | 1.20 | 0.85 |
| | MS OUTFIT | 55 | 0.99 | 0.15 | 1.26 | 0.68 |
| | SE | 55 | 0.05 | 0.01 | 0.09 | 0.05 |
| Task Model | MS INFIT | 55 | 1.05 | 0.22 | 1.78 | 0.49 |
| | MS OUTFIT | 55 | 1.08 | 0.31 | 2.15 | 0.37 |
| | SE | 55 | 0.05 | 0.01 | 0.06 | 0.05 |

Table 4.2 shows that the MS Infit and MS Outfit fit statistics for the empirical Rasch model parameters are close to or equal to 1.0 and within the acceptable range. The MS Infit and MS Outfit values for the cognitive task model exceed 1.0, signaling that there are some random` noise in the data not modelled in the empirical. The higher standard deviation for the task models indicate a much wider variability in scores around the mean, than the empirical. The maximum and minimum scores for the empirical are within acceptable ranges, while those of the cognitive task models are outside, suggesting misfits in the data. All items for the empirical measure fit the expectations of the Rasch model.

Generally, the model standard error (*SE*) values for both the empirical and the task model parameter estimates are small and nearly identical. This indicates that the models have small amounts of error associated with the estimates of item model fit.

Table 4.3

Summary Statistics across Persons for Empirical and Cognitive Task Model for English Language 2012.

| Variables | | N | MEAN | SD | MAX | MIN |
|---|---|---|---|---|---|---|
| Empirical | MS INFIT | 2172 | 1.00 | 0.11 | 1.47 | 0.69 |
| | MS OUTFIT | 2172 | 0.99 | 0.25 | 4.73 | 0.42 |
| | SE | 2172 | 0.35 | 0.09 | 1.02 | 0.3 |
| Task Model | MS INFIT | 2172 | 1.06 | 0.13 | 1.5 | 0.67 |
| | MS OUTFIT | 2172 | 1.08 | 0.26 | 2.92 | 0.35 |
| | SE | 2172 | 0.35 | 0.09 | 1.03 | 0.30 |

According to Table 4.3, the MS Infit and MS Outfit for the 2172 examinees, for the cognitive task model exceed 1.0 indicating person misfits. The empirical values are within the acceptable range. The standard deviation values for both the cognitive task model and the empirical are relatively close and indicate a wide distribution in persons along the scale. The maximum MS Outfit and MS Infit values for both the task model and the empirical are very high, suggesting large misfits in persons in the data. The

minimum values are within the acceptable range. Generally, all items for the empirical measure fit the expectations of the Rasch model.

The standard error (*SE*) values for both the empirical and the task model person measures are relatively small, indicating that there are some noise associated with the precision of the person estimates of the models.

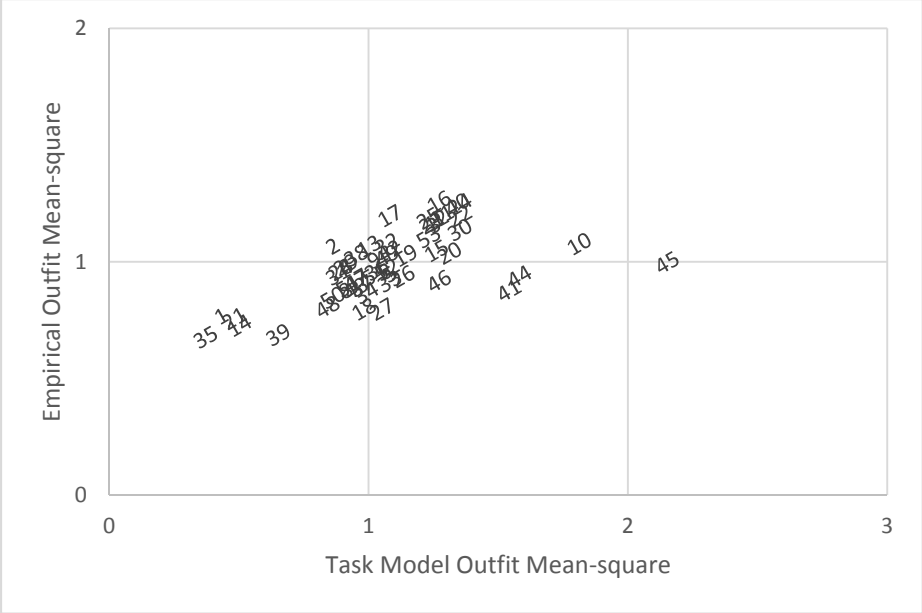



Figure 4.3 Scatterplot of Misfitting MS Outfit Items for Empirical and Cognitive Task Model for English Language 2012.

The scatterplot in Figure 4.3, shows that the main outliers for the cognitive task model are items, 35, 1, 21, 14, 39, 41, 44, 10 and 45, with very high or very low MS Oufit values. All of the MS Outfit values for the empirical are within acceptable ranges.

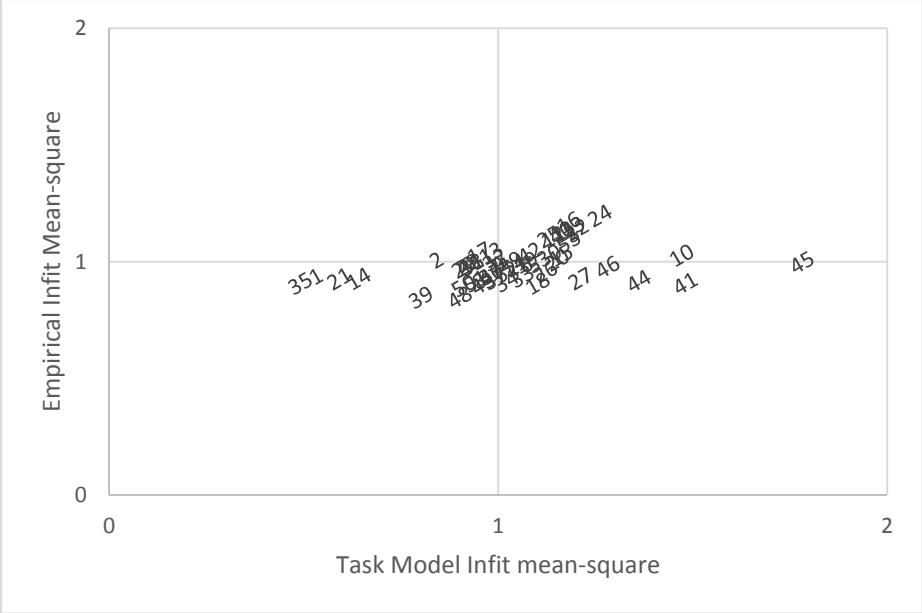



Figure 4.4 Scatterplot of Misfitting MS Infit Items for Empirical and Cognitive Task Model for English Language 2012.

According to Figure 4.4, the most misfitting items with extreme MS Infit values for the cognitive task model are 45, 10, 44 and 41, highlighted in the scatterplot above. All of the MS Infit values for the empirical are within acceptable ranges.

Table 4.4

Summary Statistics of Misfitting items for English Language 2012.

| Items | P Value | MS Infit | MS Outfit |
|---|---|---|---|
| 10 | 0.69 | 1.47 | **1.81** |
| 20 | 0.58 | **1.15** | **1.31** |
| 22 | 0.51 | **1.2** | **1.35** |
| 24 | 0.48 | 1.26 | **1.35** |
| 30 | 0.7 | **1.13** | **1.35** |
| 40 | 0.56 | 1.16 | **1.34** |
| 41 | 0.66 | 1.48 | **1.54** |
| 44 | 0.33 | **1.36** | **1.58** |
| 45 | 0.41 | **1.78** | **2.15** |

Table 4.4 shows the 9 cognitive task model misfitting items with high and extreme MS Infit and MS Outfit values, which exceed 1.30 and are outside of the acceptable range. There are four items with high MS Infit values. These items have p-values ranging from .69 to .33.

**Description of Misfitting Item for English Language 2012**

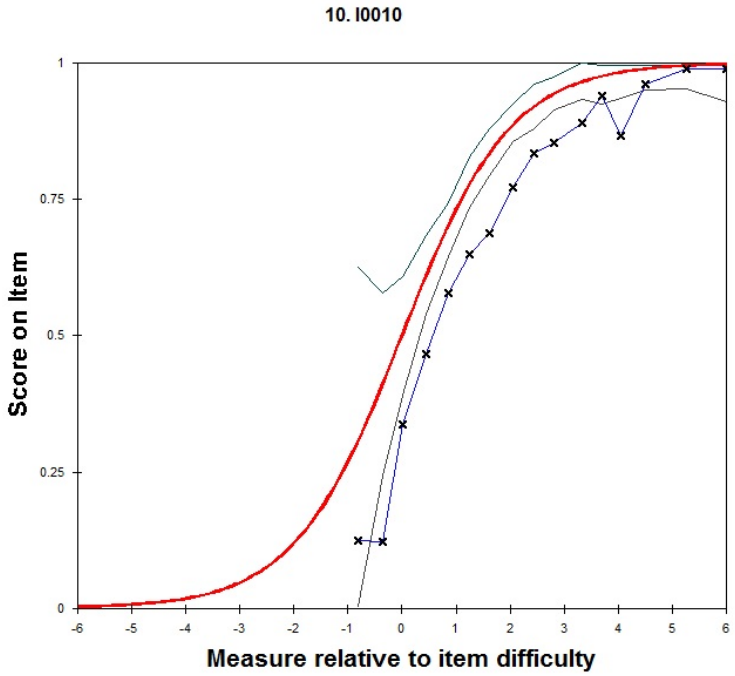



Figure 4.5 Empirical Conditional Mean Scores and Expected Response Function Item10.

Figure 4.5 shows the expected response function and the task model difficulty conditional on the empirical scores for item 10. The examinees are generally tracking the model line, however misfit is apparent in the upper regions of the scale at approximately +3.5 logits, where high performing examinees were unexpectedly scoring this item incorrectly.

These appear to be random examinees, as the most misfitting persons and those with the most misfitting responses were different from those with the most unexpected responses. The proportion of examinees to score this item correctly is 69%.

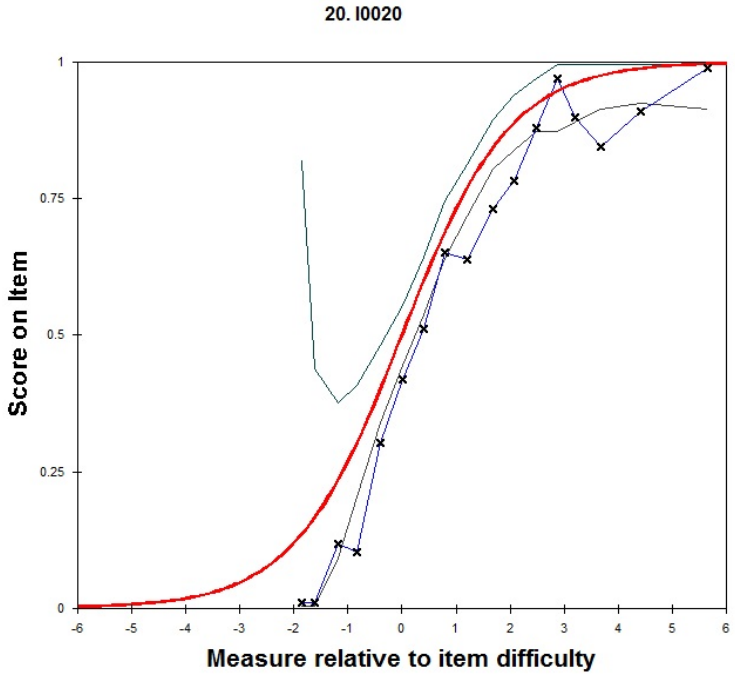



Figure 4.6 Empirical Conditional Mean Scores and Expected Response Function Item 20.

Figure 4.6 shows the expected response function and the task model difficulty conditional on the empirical scores for item 20. The examinees are roughly tracking the model line. In this figure, the misfit is apparent in the upper regions of the scale at approximately 2.8 logits, where high performing examinees predicted to score this item correctly were unexpectedly scoring it incorrectly. Some of the most misfitting examinees, with the most misfitting response strings also had the most unexpected responses to this item, while others were random examinees. The error may have been due to careless mistakes. 58% of examinees scored this item correctly.

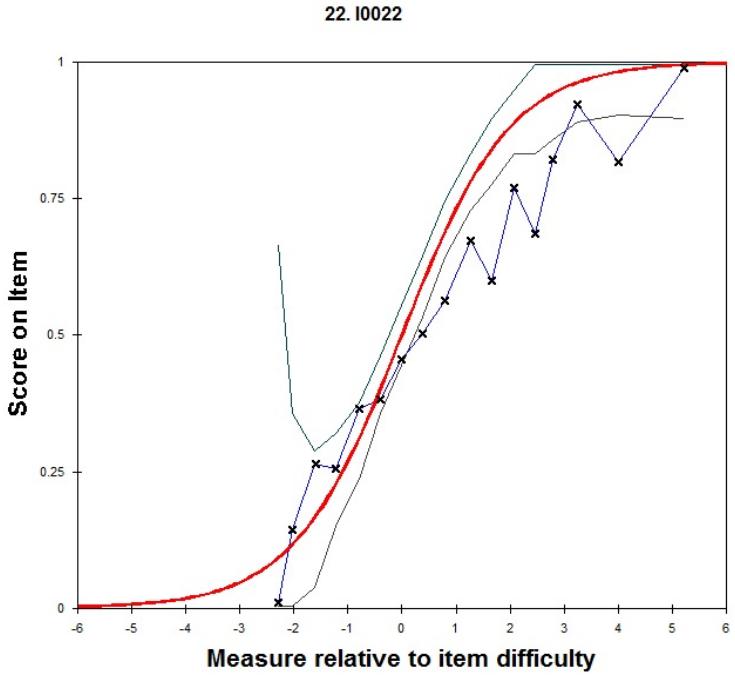



Figure 4.7 Empirical Conditional Mean Scores and Expected Response Function Item 22.

Figure 4.7 shows the expected response function and the empirical scores conditional on the estimated empirical scores for item 22. The misfit is pronounced in the upper regions of the scale, above +1.0 logits where high performing examinees are expectedly scoring this item incorrectly. These appear to be random examinees scoring this item incorrectly. Some of these were the most misfitting persons, with the most misfitting response string and gave the most unexpected responses to the item. However, most were random examinees. This suggest that persons were making careless mistakes. The proportion who scored correctly is 51%.

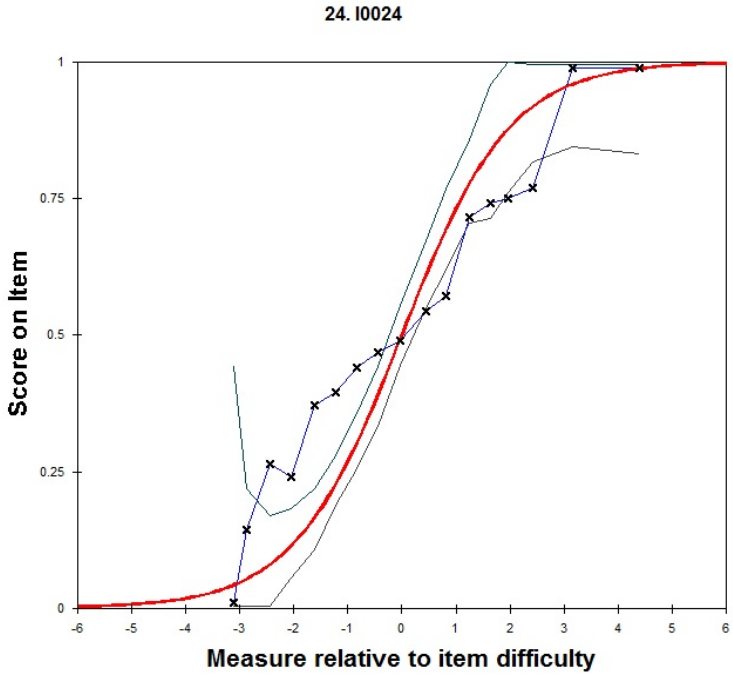



Figure 4.8 Empirical Conditional Mean Scores and Expected Response Function Item 24.

Figure 4.8 shows the expected response function and the empirical scores conditional on the estimated empirical scores for item 11. The misfit is pronounced at the upper regions of the scale at approximately 2.5 logits and above, where high performing examinees are unexpectedly scoring this item incorrectly. Misfit is also evident at -0.2 logits where examinees were unexpectedly scoring this item correctly, who were predicted to score it incorrectly. Overall, these were random examinees scoring this item as the most misfitting examinees, with the most misfitting response strings, who gave the most unexpected responses were different. This item is not differentiating between high and low performers. Approximately 48% of examinees scored correctly on this item.

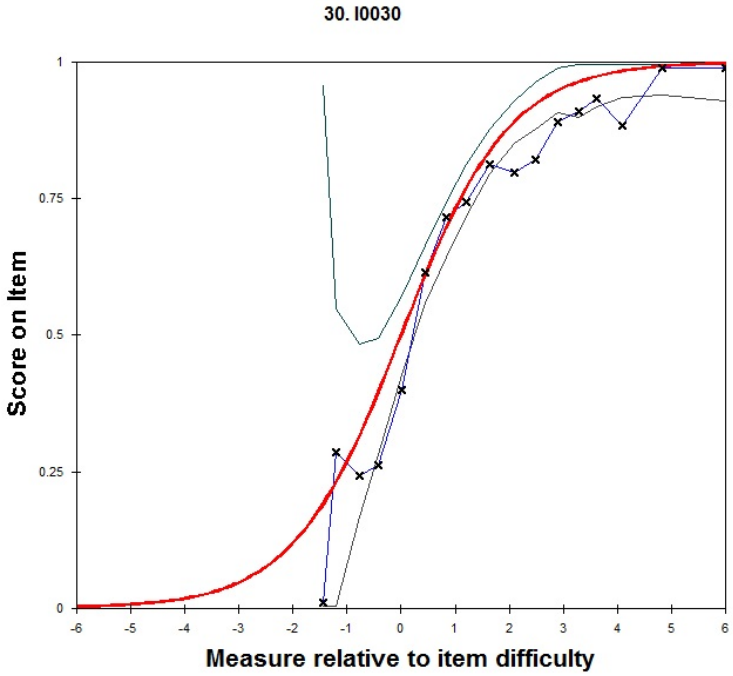



Figure 4.9 Empirical Conditional Mean Scores and Expected Response Function Item 30.

Figure 4.9 shows the expected response function and the empirical scores conditional on the estimated empirical scores for item 30. The misfit is pronounced in the upper regions of the logit scale, at approximately +1.0 and +3.5 logits, where high performing examinees are unexpectedly scoring this item incorrectly. Misfit is also evident in the lower regions of the logit scale at approximately -1.0 logits, where low performing examinees are unexpectedly scoring the item correctly. The same examinees who were the most misfitting, also give the most misfitting responses, but not the most unexpected responses. 70% of examinees scored this item correctly.

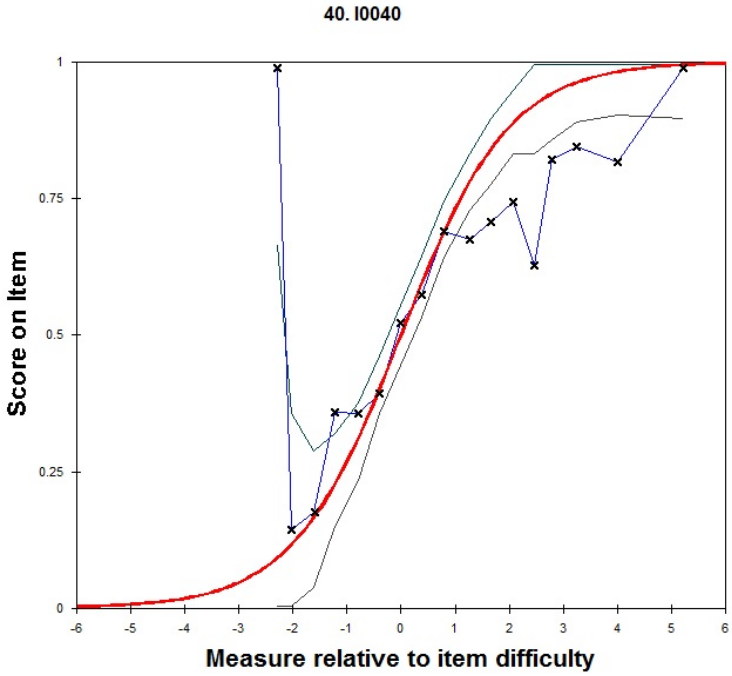



Figure 4.10 Empirical Conditional Mean Scores and Expected Response Function Item 40.

Figure 4.10 shows the expected response function and the empirical scores conditional on the estimated empirical scores for item 40. The misfit is evident in the upper regions of the scale at about +2.3 logits, where high performing examinees are unexpectedly scoring this item incorrectly. The misfit is also observed in the lower regions of the scale, at approximately, -1.0 logits where low performing examinees are unexpectedly scoring the item correctly. Some of the same examinees who were most misfitting with the most misfitting responses also had the most unexpected responses. This could be as a result of examinees carelessness, or guessing. The proportion of examinees who scored correctly on this item is 56 %.

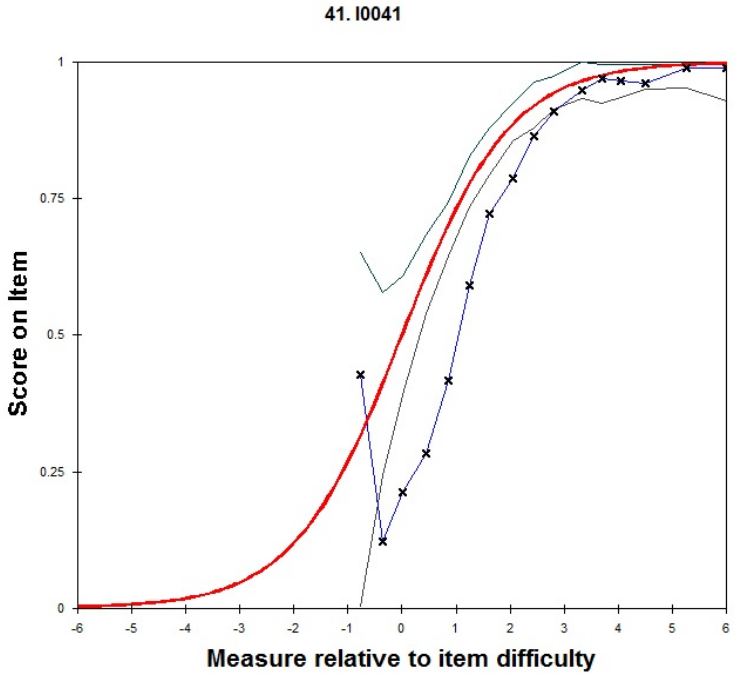



Figure 4.11 Empirical Conditional Mean Scores and Expected Response Function Item 41.

Figure 4.11 shows the expected response function and the empirical scores conditional on the estimated empirical scores for item 41. The examinees are generally following the model line. However, misfit is evident at approximately -0.05 logits, where examinees are unexpectedly scoring this item incorrectly. The most misfitting examinees with the most misfitting response string' gave the most unexpected responses. 56% of the proportion of examinees scored the item correctly item.

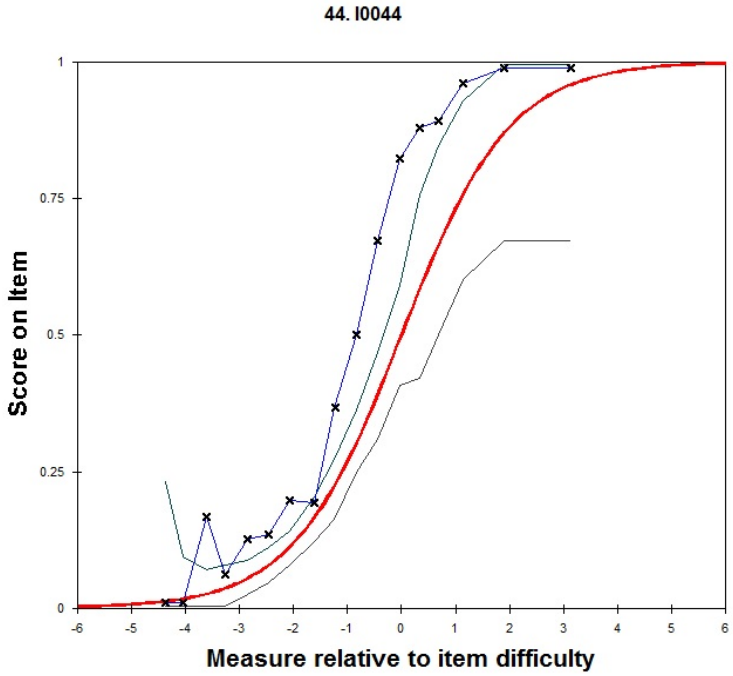



Figure 4.12 Empirical Conditional Mean Scores and Expected Response Function Item 44.

Figure 4.12 shows the expected response function and the empirical scores conditional on the estimated empirical scores. Generally, the examinees are tracking the model line. However, the misfit is pronounced in the lower regions of the scale at approximately -3.2 logits, where low performing examinees were unexpectedly scoring this item correctly. The same examinees had the most misfitting response string and gave the most unexpected responses. The proportion who scored it correctly is 66%.

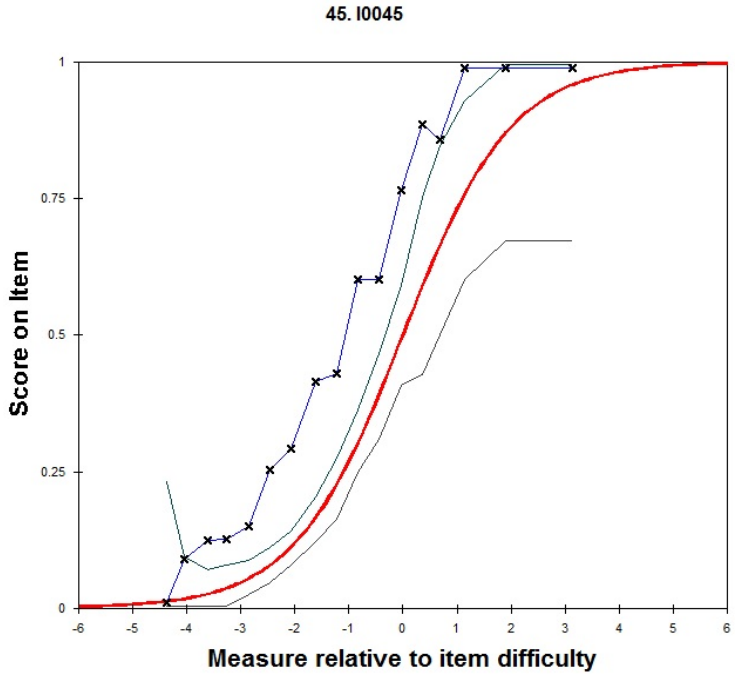



Figure 4.13 Empirical Conditional Mean Scores and Expected Response Function Item 45.

Figure 4.13 shows the expected response function and the empirical scores conditional on the estimated empirical scores. While misfits is evident along the scale, the misfit is very apparent in the lower regions of the scale at approximately -2.8 logits, where low performing examinees are unexpectedly scoring this item correctly. Some of the examinees with the most misfitting responses also had the most unexpected response strings, while others were random examinees. The proportion of examinees to score this item correctly is 66 %.

**Rasch Model Fit of Task Model and Empirical Items for English Language 2013**

Table 4.5

Summary Statistics across Items for Empirical and Cognitive Task Model for English Language 2013.

| Variables | | N | MEAN | SD | MAX | MIN |
|---|---|---|---|---|---|---|
| Empirical | MS INFIT | 54 | 1.00 | 0.07 | 1.18 | 0.86 |
| | MS OUTFIT | 54 | 0.99 | 0.15 | 1.65 | 0.65 |
| | SE | 54 | 0.04 | 0.00 | 0.05 | 0.03 |
| Task Model | MS INFIT | 54 | 1.11 | 0.36 | 2.26 | 0.66 |
| | MS OUTFIT | 54 | 1.16 | 0.54 | 2.94 | 0.55 |
| | SE | 54 | 0.04 | 0.00 | 0.05 | 0.03 |

According to Table 4.5, the MS Infit and MS Outfit values for the empirical Rasch model are close to or equal to 1.0 and within the acceptable range. The MS Infit and MS Outfit values for the cognitive task model exceed 1.0, which suggest irregularities in the data, such as outliers or unexpected response patterns near the examinee's ability level along the latent scale. The standard deviation reveals wide variability in scores for the task model and closely distributed scores for the empirical. The maximum MS Infit and MS Outfit scores are high and outside of the acceptable range for the task model. The empirical has a high MS Outfit value, which is also unacceptable. The minimum scores are within acceptable ranges.

Generally, the model standard error (*SE*) values for both the empirical and the task model person measures are small and similar indicating that misfit of the data to the model is small in relation to the precision of measurement.

Table 4.6

Summary Statistics across Persons for Empirical and Cognitive Task Model for English Language 2013.

| Variables | | N | MEAN | SD | MAX | MIN |
|---|---|---|---|---|---|---|
| Empirical | MS INFIT | 4299 | 1.00 | 0.13 | 1.54 | 0.62 |
| | MS OUTFIT | 4299 | 0.99 | 0.24 | 2.75 | 0.33 |
| | SE | 4299 | 0.34 | 0.05 | 1.02 | 0.30 |
| Task Model | MS INFIT | 4299 | 1.08 | 0.14 | 1.75 | 0.64 |
| | MS OUTFIT | 4299 | 1.16 | 0.30 | 3.92 | 0.33 |
| | SE | 4299 | 0.34 | 0.05 | 1.03 | 0.30 |

According to Table 4.6, the MS Infit and MS Outfit for the 4299 examinees for the empirical analysis has values close to or equal to 1.0, and within the acceptable range. The MS Infit and MS Outfit values for the task models just exceed 1.0, indication some misfits in the data. The standard deviation for both the task model and empirical indicate a comparable distribution of persons along the scale. The maximum values for both the

empirical and the task model are outside of the acceptable range and indicate some misfits of persons. The minimum values are acceptable.

Overall, the standard error shows that there are some noise associated with the precision of the person measures.

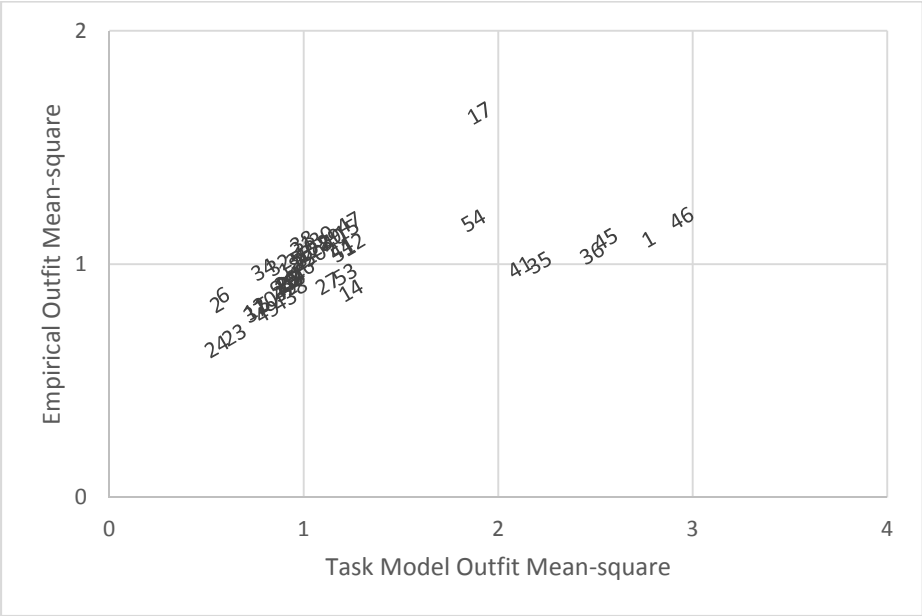



Figure 4.14 Scatterplot of Misfitting MS Outfit Items for Empirical and Cognitive Task Model for English Language 2013.

Figure 4.14, scatterplot shows the MS Outfit items for the empirical and task model for English Language 2012. The empirical has item 17 with extreme values. The items for the cognitive task model with extremely high or low MS Outfit values are 46, 1, 45 36 35, 41 and 17.

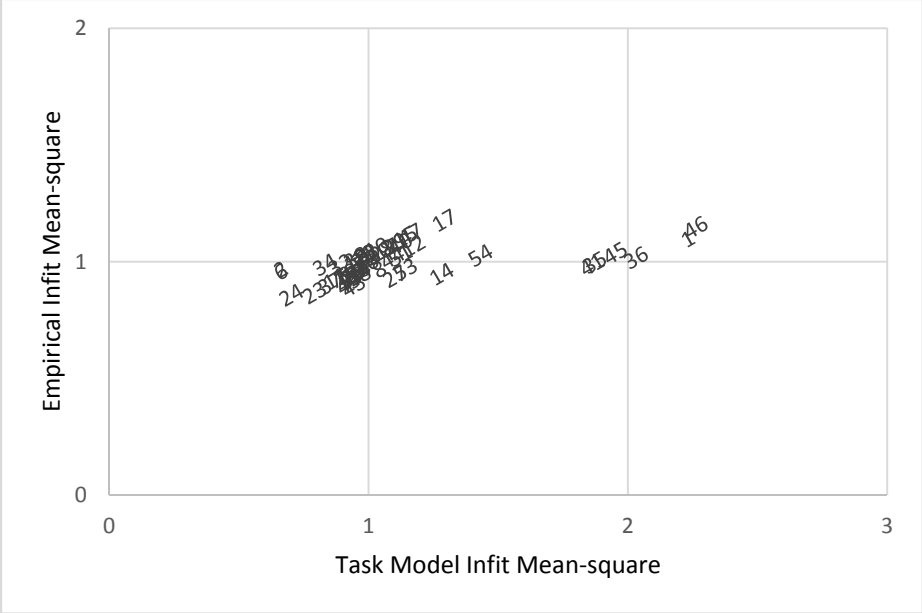



Figure 4.15 Scatterplot of Misfitting MS Infit Items for Empirical and Cognitive Task Model for English Language 2013.

According to Figure 4.15, the scatterplot shows the MS Infit values for the empirical and cognitive task model. The most misfitting items with extreme MS Infit values for the cognitive task model are 46, 1, 45, 36, 35, 41 and 17. These items have extremely high or low MS Infit values that signals misfits in the data. There are no items or outliers for the empirical with extreme MS Infit values.

Table 4.7

Summary Statistics of Misfitting items for English Language 2013.

| Items | P Value | Infit | Outfit |
|---|---|---|---|
| 1 | 0.55 | **2.23** | **2.77** |
| 17 | 0.17 | **1.29** | **1.9** |
| 35 | 0.61 | **1.87** | **2.21** |
| 36 | 0.34 | **2.03** | **2.48** |
| 41 | 0.5 | **1.86** | **2.11** |
| 45 | 0.33 | **1.95** | **2.55** |
| 46 | 0.44 | **2.26** | **2.94** |
| 54 | 0.24 | **1.43** | **1.87** |
| EMP 17 | 17 | **1.18** | **1.65** |

Table 4.7 shows scores for the 8 misfitting items with extreme MS Outfit and MS Infit values for the empirical and cognitive task model analysis. The proportion correct range from .61 to .17. The extreme MS Infit values range from 2.26 to 1.43. The extreme MS Outfit values range from 2.94 to 1.87. Item 17 is the only misfitting item that is common to both the empirical and the task model.

**Description of Misfitting Items for English Language 2013**

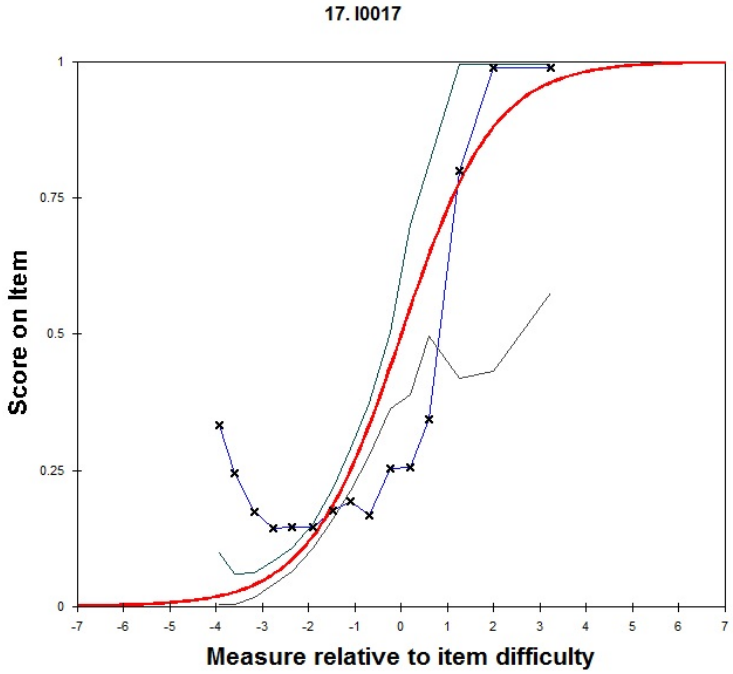



Figure 4.16 Empirical Conditional Mean Scores and Expected Response Function Item 17.

Figure 4.16, shows the expected response function and the empirical scores, conditional on the estimated empirical scores for item 17. The misfit is very pronounced along the lower regions of the scale, where the low performing examinees predicted to score this item incorrectly were scoring it correctly. The most misfitting examinees, with the most misfitting response string were giving the most unexpected responses. This item appear to be measuring more than one dimension, as persons who were expected to score incorrectly were not. Only 17% ($p_i = .17$) of examinees scored correctly on this item.

This item appear to have systematic error and should be considered to be revised or discarded. The options should be examined for ambiguity.

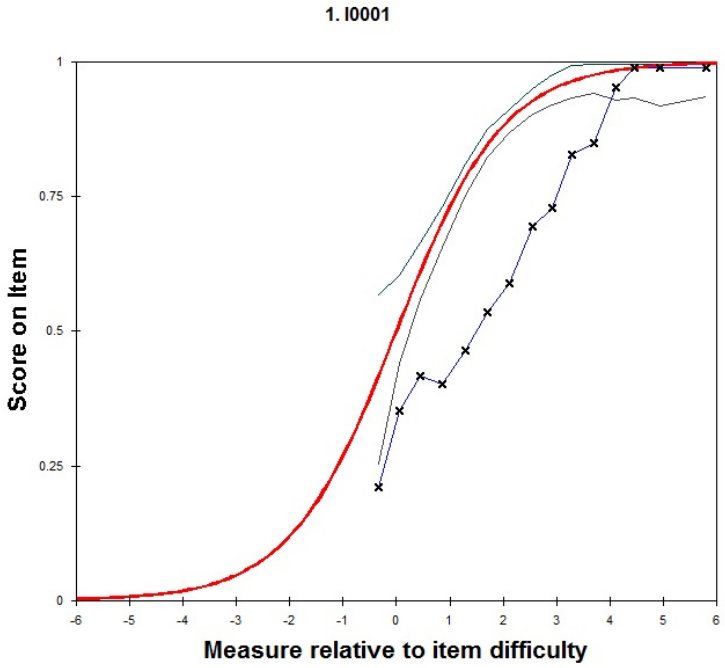



Figure 4.17 Empirical Conditional Mean Scores and Expected Response Function Item 1.

The misfit is pronounced in the upper regions of the scale at approximately +3.0 logits, where high performing examinees who were predicted to score correctly were unexpectedly scoring this item incorrectly. The examinees who were scoring incorrectly appear to be random as those who were most misfitting were different to those with the most misfitting responses and those who gave the most unexpected responses. These examinees may have made careless mistakes in selecting the correct answer. 55 % of examinees scored this item correctly.

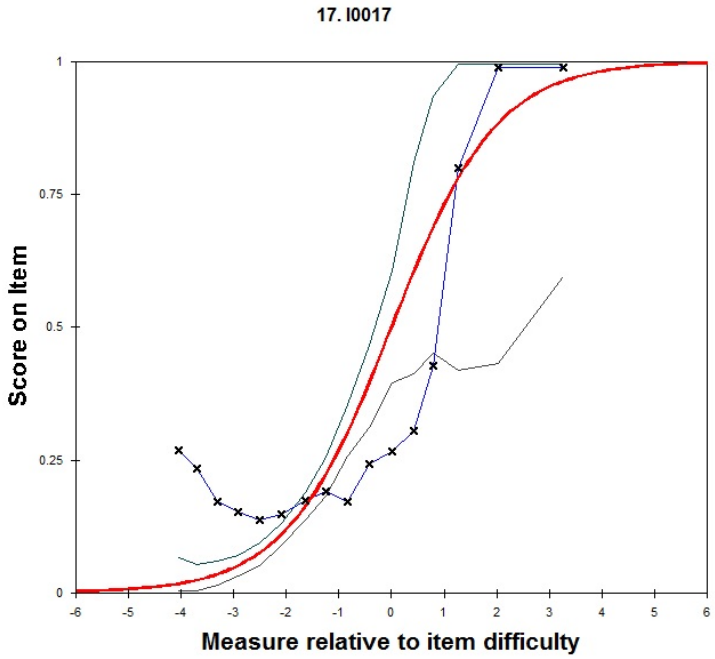



Figure 4.18 Empirical Conditional Mean Scores and Expected Response Function Item 17

Figure 4.18, shows the expected response function and the empirical scores, conditional on the estimated empirical scores for item 17 for cognitive task model. This item is showing great misfit along the scale. This item appear to be measuring more than one dimension, as persons who should be scoring correctly, unexpectedly are not and those predicted to score it incorrectly at the lower regions, are unexpectedly scoring it correctly. Only 17 % ($p_i = .17$) of examinees scored this item correctly.

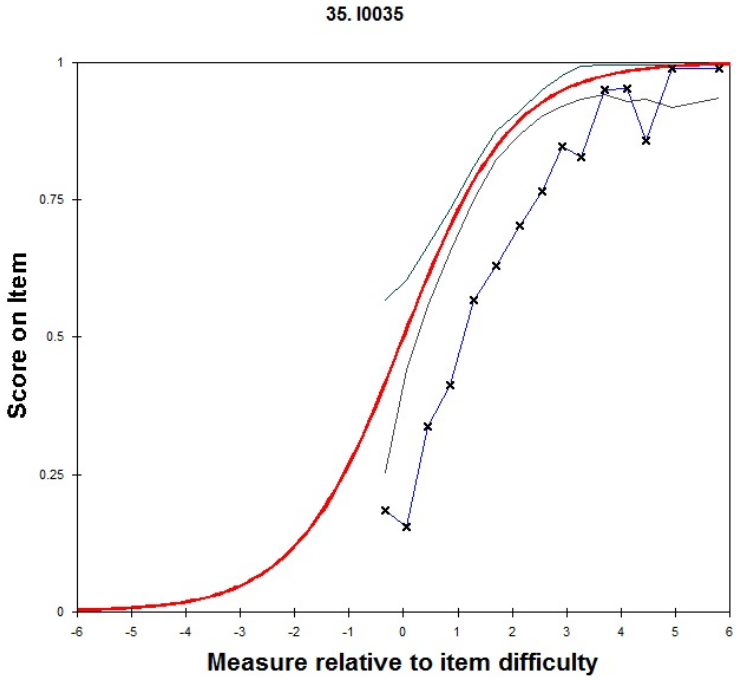



Figure 4.19. Empirical Conditional Mean Scores and Expected Response Function Item 35.

Figure 4.19, shows the expected response function and the empirical scores, conditional on the estimated empirical scores. The examinees appear to be roughly following the model line, however, misfits are evident in the upper regions of the scale, above +3.0 logits, where high performing examinees are unexpectedly scoring the item incorrectly. Some of these examinees had the most unexpected response string and the most misfitting response string. They appear to be mainly random examinees. 61% scored this item correctly.

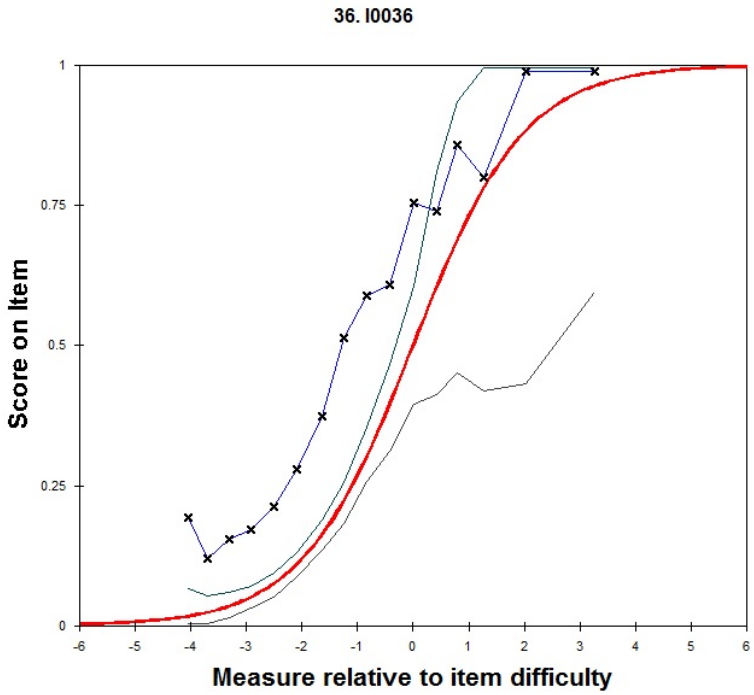



Figure 4.20 Empirical Conditional Mean Scores and Expected Response Function Item 36.

Figure 4.20, shows the expected response function and the empirical scores, conditional on the estimated empirical scores for item 36. The examinees are roughly following the model line, but misfit is evident in the lower regions of the scale, at about -3.8 logits, where low performing examinees are unexpectedly scoring this item correctly. Some of the most misfitting examinees, had the most misfitting responses and the most unexpected response string in addition to other random examinees. Some high performing examinees were unexpectedly scoring this item incorrectly as well. 34 % of the examinees scored correctly on this item. This item should be examined for the possibility that more than one option can be correct.

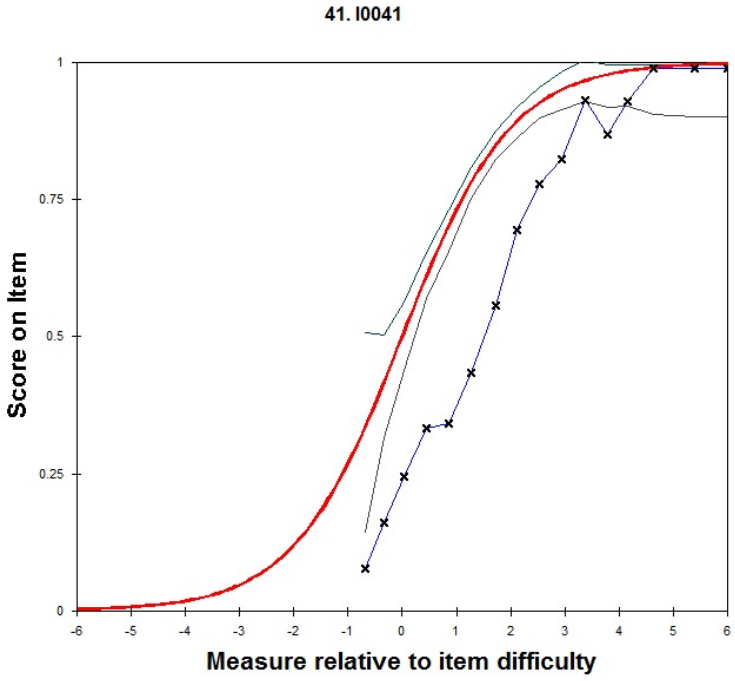



Figure 4.21 Empirical Conditional Mean Scores and Expected Response Function Item 41

Figure 4.21, shows the expected response function and the empirical scores, conditional on the estimated empirical scores for item 41. The item is misfitting in the upper regions of the scale, at approximately 3.0 logits, where high performing examinees are unexpectedly scoring this item incorrectly. These same examinees had the most misfitting response string and the most unexpected response string, while others were random examinees. 50 % of examinees scored correctly on this item. The examinees could have made careless mistakes.

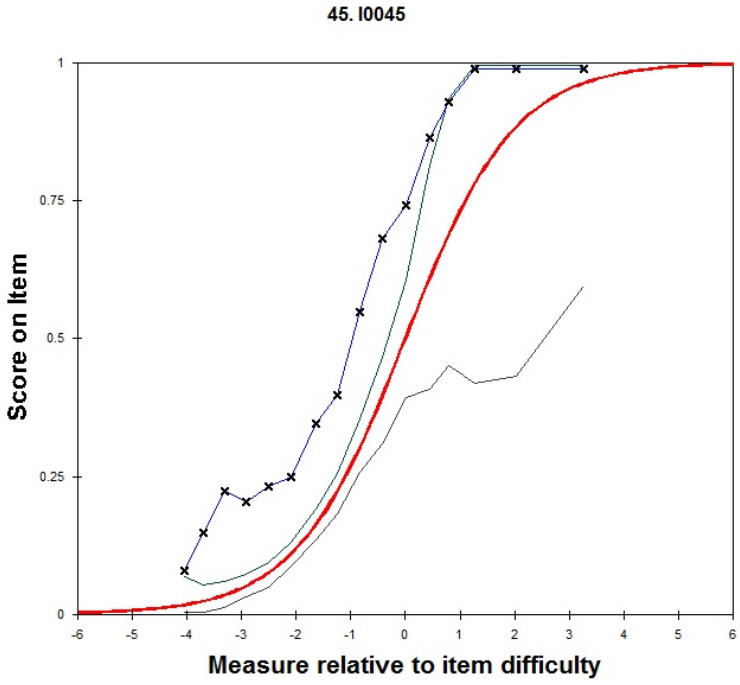



Figure 4.22 Empirical Conditional Mean Scores and Expected Response Function Item 45.

Figure 4.22, shows the expected response function and the empirical scores, conditional on the estimated empirical scores for item 45. This item is showing that examinees are roughly tracking the model line. However, misfits are evident in the lower regions of the scale, at approximately 3.0 logits, where a number of low performing examinees are unexpectedly scoring this item correctly. These were the most misfitting persons, with the most misfitting responses and the most unexpected responses to the item. Also included are some random examinees. 33 % of examinees scored correctly on this item. Some examinees may have guessed the correct response.

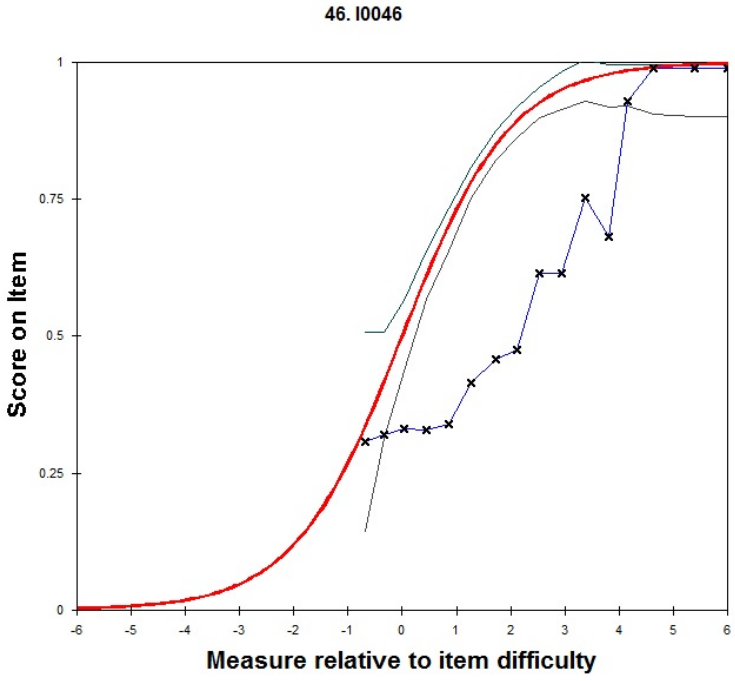



Figure 4.23 Empirical Conditional Mean Scores and Expected Response Function Item 46

Figure 4.23, shows the expected response function and the empirical scores, conditional on the estimated empirical scores for item 46. The misfit is pronounced in the upper regions of the scale above 2.0 logits, where high ability examinees are unexpectedly scoring this item incorrectly. Some of the same examinees had the most misfitting responses and gave the most unexpected responses. 44 % of the examinees scored correctly on this item.

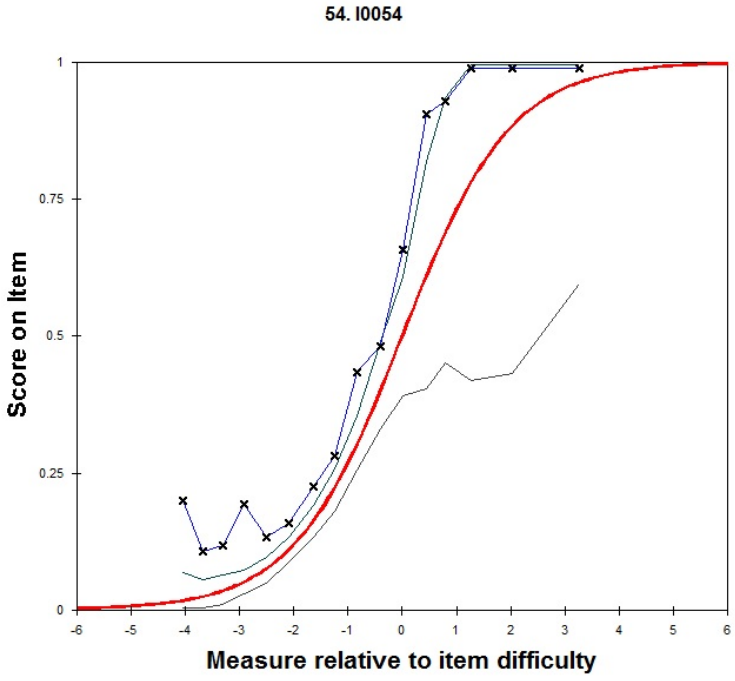



Figure 4.24 Empirical Conditional Mean Scores and Expected Response Function Item 54

Figure 4.24, shows the expected response function and the empirical scores, conditional on the estimated empirical scores. The examinees are roughly tracking the model line, however misfit is pronounced in the lower regions of the scale at approximately 2.5 logits, where low performing examinees are unexpectedly scoring this item correctly. Some of these same examinees had the most misfitting and unexpected responses. Examinees may have guessed the correct response. 54 % of examinees scored correctly on this item.

**Rasch Model Fit of Task Model and Empirical Difficulties for Calculus 2012**

Table 4.8

Summary Statistics across Items for Empirical and Cognitive Task Model for Calculus 2012

| Variables | | N | MEAN | SD | MAX | MIN |
|---|---|---|---|---|---|---|
| Empirical | MS INFIT | 45 | 1.00 | 0.06 | 1.11 | 0.89 |
| | MS OUTFIT | 45 | 0.99 | 0.09 | 1.17 | 0.80 |
| | SE | 45 | 0.04 | 0.00 | 0.05 | 0.03 |
| Task Model | MS INFIT | 45 | 1.12 | 0.25 | 2.19 | 0.80 |
| | MS OUTFIT | 45 | 1.17 | 0.38 | 2.83 | 0.74 |
| | SE | 45 | 0.04 | 0.00 | 0.04 | 0.03 |

Table 4.8 shows that the empirical MS Infit and MS Outfit values are close to or equal to 1.0, within the acceptable range, indicating relative fit of the data to the model. The high MS Infit and MS Outfit scores for the task model suggest that there are irregularities in the data. The small standard deviation for the empirical indicate that the scores are very compact, while those of the task model are broadly distributed around the mean. The maximum scores for the task model extreme and outside of the acceptable range. The maximum and minimum scores for the empirical are within the acceptable

range. Thus, Table 8 illustrates that all items for the empirical measure fit the expectations of the Rasch model.

The model standard error (*SE*) values for both the empirical and the task model person measures are small, indicating that misfit of the data to the model's precision of measurement is small.

Table 4.9

Summary Statistics across Persons for Empirical and Cognitive Task Model for Calculus 2012.

| Variables | | N | MEAN | SD | MAX | MIN |
|---|---|---|---|---|---|---|
| Empirical | MS INFIT | 4248 | 1.00 | 0.10 | 1.40 | 0.73 |
| | MS OUTFIT | 4248 | 0.99 | 0.20 | 2.76 | 0.31 |
| | SE | 4248 | 0.37 | 0.09 | 1.02 | 0.32 |
| Task Model | MS INFIT | 4248 | 1.11 | 0.14 | 1.68 | 0.67 |
| | MS OUTFIT | 4248 | 1.17 | 0.30 | 5.46 | 0.14 |
| | SE | 4248 | 0.38 | 0.09 | 1.03 | 0.33 |

According to Table 4.9, the MS Infit and MS Outfit for the empirical person measures are close to or equal to 1.0, within the acceptable range. The MS Infit and MS Outfit values for the task models just exceed the acceptable range. Generally, the

standard deviation for the task model and the empirical indicate a relatively broad distribution of persons along the scale. The extreme maximum values indicate that there are some random noise in the data. The minimum values are within acceptable range.

Overall, the standard error shows that there are random noise associated with the precision of the person model fit measures for the empirical and task model.

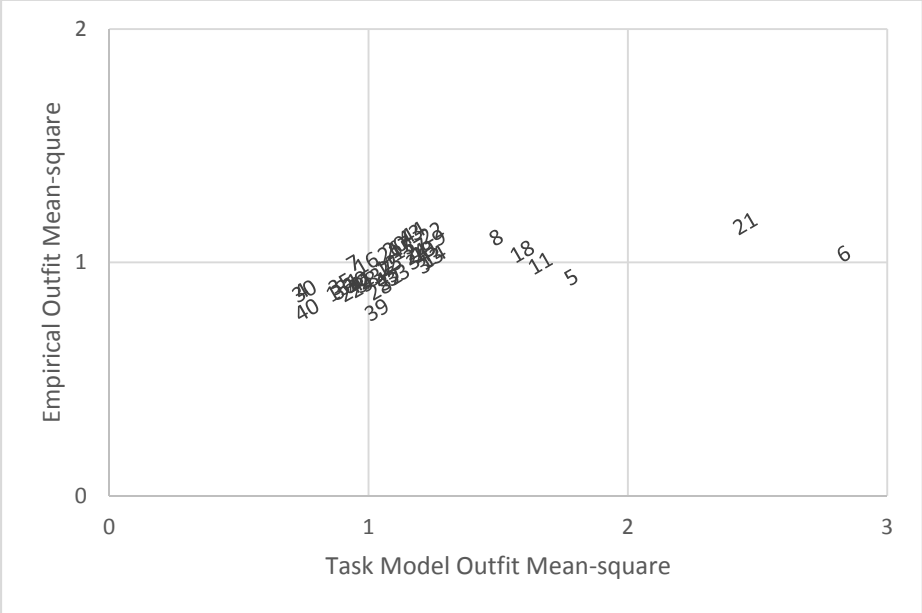



Figure 4.25 Scatterplot of Misfitting MS Outfit Items for Empirical and Cognitive Task Model for Calculus 2012.

Figure 4.25, scatterplot shows that the scores are generally closely aligned together. It highlights the extreme outliers for the cognitive task model. Items 6, 21, 5 and 11 have the highest or lowest MS Outfit values and are considered to be outliers. There are no outliers for the empirical in this dataset.

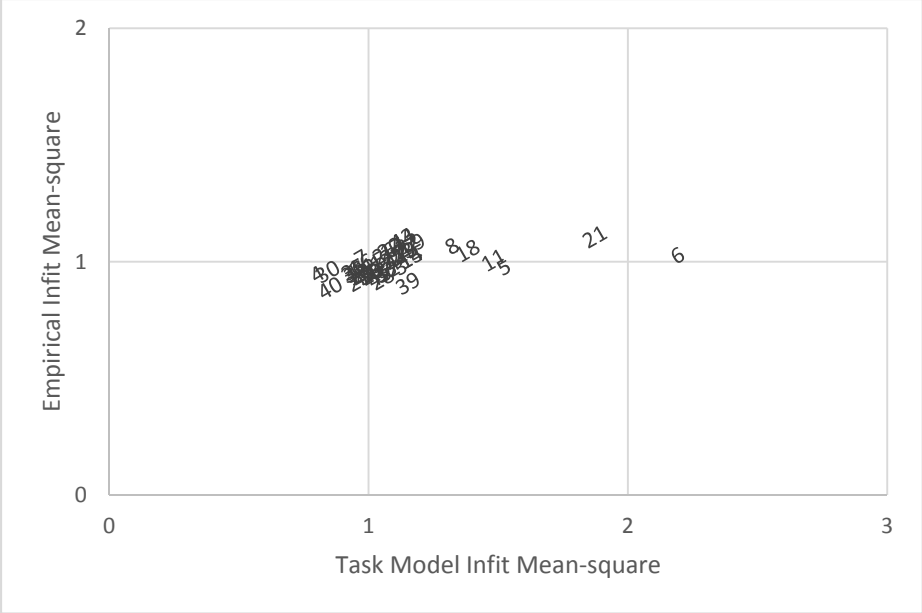



Figure 4.26 Scatterplot of Misfitting MS Infit Items for Empirical and Cognitive Task Model for Calculus 2013.

Figure 4.26, shows the scatterplot of the MS Infit items for the empirical and the task model. The most misfitting items with extreme MS Infit values for the cognitive task model are 6, 21, 15 and 11. There are no items or outliers for the empirical with extreme values.

Table 4.10

Summary Statistics of Misfitting Items for Calculus 2012

| Items | P Value | Infit | Outfit |
|---|---|---|---|
| 5 | 0.61 | **1.52** | **1.78** |
| 6 | 0.6 | 2.19 | **2.83** |
| 8 | 0.46 | **1.32** | **1.49** |
| 11 | 0.7 | **1.48** | **1.66** |
| 18 | 0.51 | **1.38** | **1.59** |
| 21 | 0.34 | **1.87** | **2.45** |

Table 4.10 shows 6 cognitive task model items with high or extreme MS Infit and MS Outfit scores. The MS Infit values range from 2.19 to 1.32. The MS Outfit values range from 3.83 to 1.49. These extreme scores are responsible for any misfit in the data. The most misfitting items are 6 and 21.

**Description of Misfitting Items for Calculus 2012**

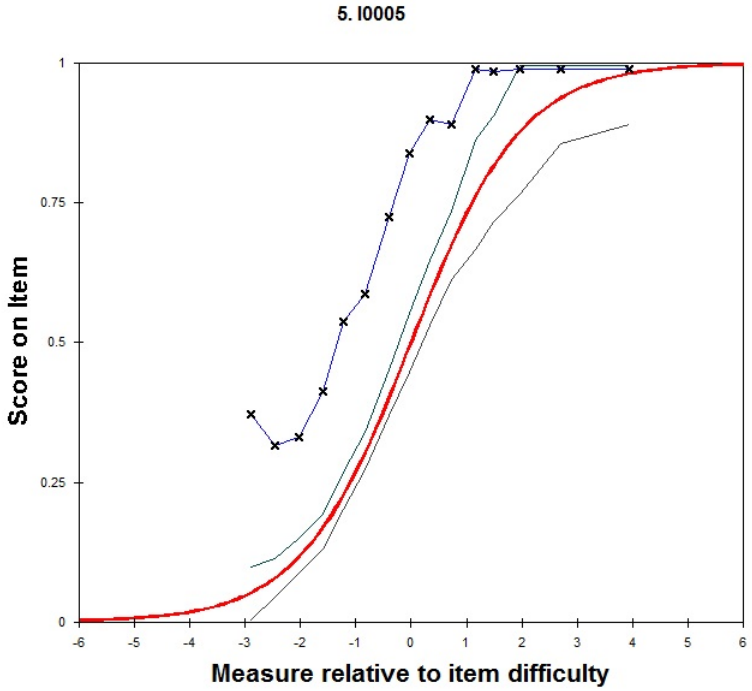



Figure 4.27 Empirical Conditional Mean Scores and Expected Response Function Item 5.

Figure 4.27, shows the expected response function and the empirical scores, conditional on the estimated empirical scores for item 5. Generally, the examinees are tracking the model line. The misfit is pronounced in the lower regions of the scale at approximately -2.8 logits, where low performing examinees were unexpectedly scoring this item correctly. These appear to be random examinees with the most misfitting response string. 61% of examinees scored this item correctly. Most likely examinees may have guessed the correct response.

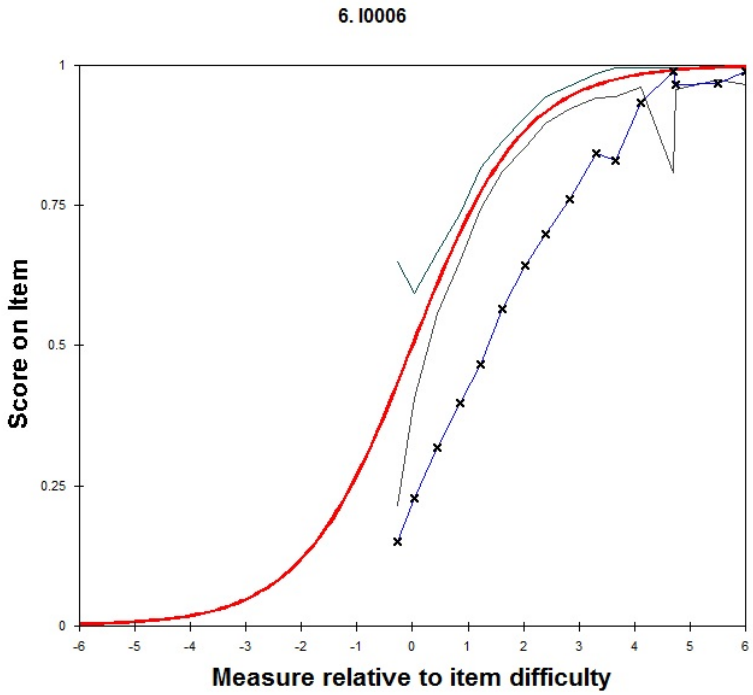



Figure 4.28 Empirical Conditional Mean Scores and Expected Response Function Item 6.

Figure 4.28, shows the expected response function and the empirical scores, conditional on the estimated empirical scores for item 6. The misfit is evident in the upper regions of the scale where high performing examinees predicted to score this item correctly were unexpectedly scoring it incorrectly. This could be as a result of careless mistake. 60 % of examinees scored this item correctly.

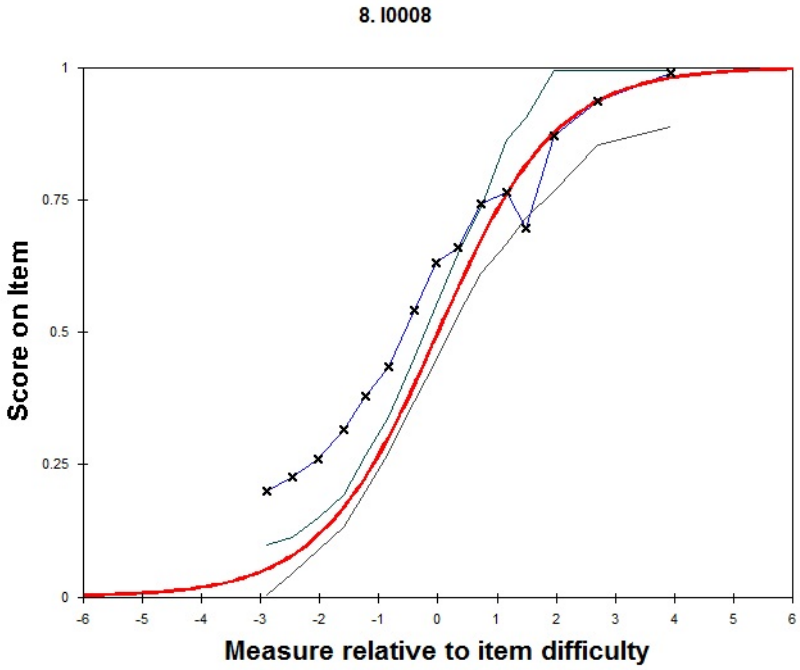



Figure 4.29 Empirical Conditional Mean Scores and Expected Response Function Item 8.

Figure 4.29, shows the expected response function and the empirical scores, conditional on the estimated empirical scores. This item is showing misfits and in the upper regions at about 2.0 logits, where high performing examinees were unexpectedly scoring this item incorrectly. Some low performing examinees were unexpectedly scoring this item correctly. These examinees comprised those who were most misfitting, gave the most misfitting responses and the most unexpected response string. The proportion of examinees to score this item correctly is 46%.

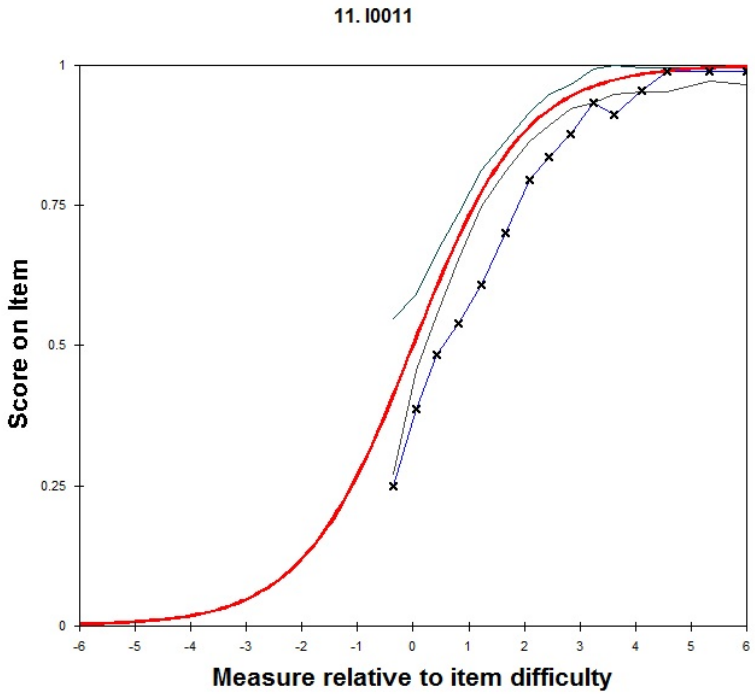



Figure 4.30 Empirical Conditional Mean Scores and Expected Response Function Item 11.

Figure 4.30, shows the expected response function and the empirical scores, conditional on the estimated empirical scores. The examinees are generally tracking the model line. This item is showing misfit in the upper regions of the scale, at approximately +2.8 logits where examinees were unexpectedly scoring this item incorrectly. 70% of examinees scored this item correctly.

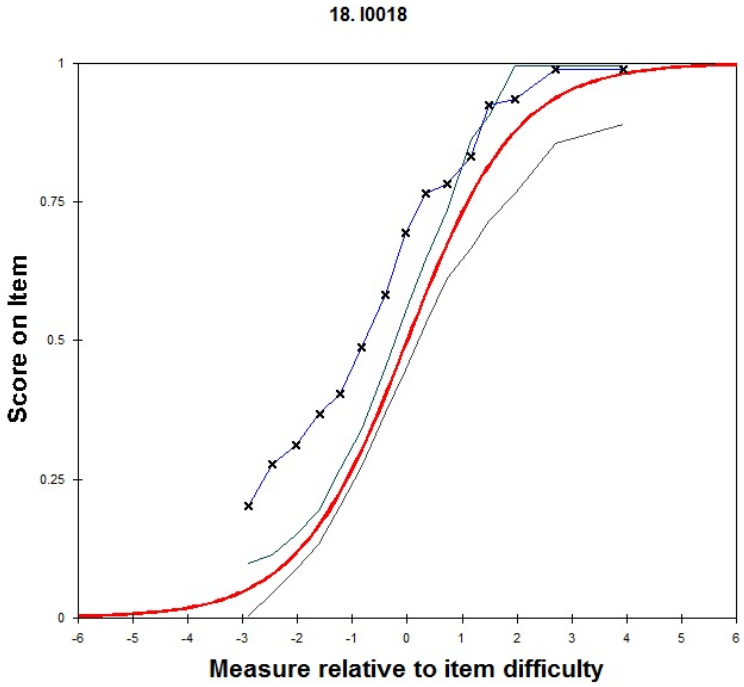



Figure 4.31 Empirical Conditional Mean Scores and Expected Response Function Item 18.

Figure 4.31, shows the expected response function and the empirical scores, conditional on the estimated empirical scores. The examinees are generally tracking the model line. However, the item is showing misfit in the lower regions of the scale -2.0 logits, where low performing examinees were unexpectedly scoring this item correctly. There were some random examinees in the upper regions of the scale who unexpectedly scored this item incorrectly. Some of the most misfitting examinees had the most misfitting responses, but not the most unexpected responses. 51 % of the proportion of examinees scored this item correctly.

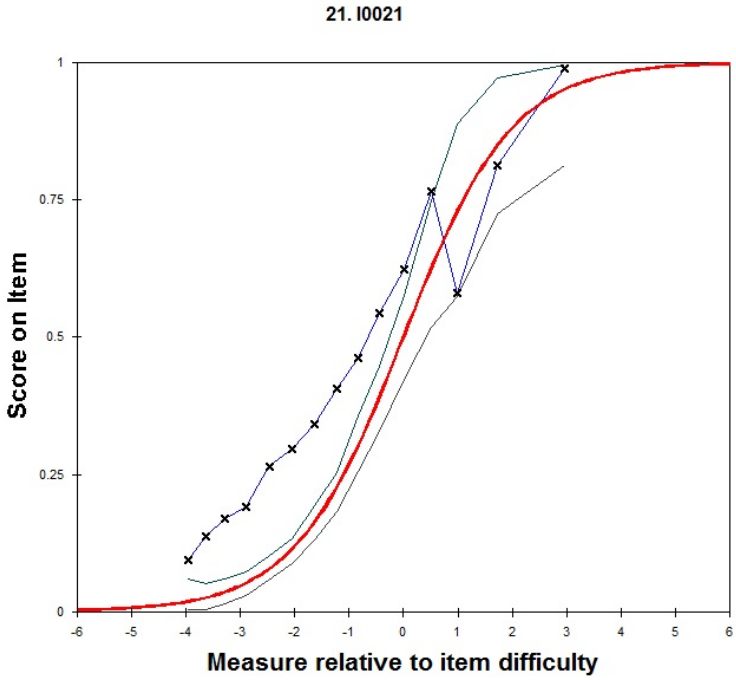



Figure 4.32 Empirical Conditional Mean Scores and Expected Response Function Item 21.

Figure 4.32, shows the expected response function and the empirical scores, conditional on the estimated empirical scores for item 21. This item misfit is pronounced in the region of the scale at approximately 1.0 logits, where high performing examinees are unexpectedly scoring this item incorrectly. These were some of the most misfitting examinees, with the most misfitting response strings and the most unexpected responses to this item. A number of examinees at the lower region of the scale unexpectedly scored this item correctly. 34 % of all examinees scored this item correctly.

**Rasch Model Fit of Task Model and Empirical Difficulties for Calculus 2013**

Table 4.11

Summary Statistics across Items for Empirical and Cognitive Task Model for Calculus 2013

| Variables | | N | MEAN | SD | MAX | MIN |
|---|---|---|---|---|---|---|
| Empirical | MS INFIT | 45 | 1.00 | 0.06 | 1.11 | 0.88 |
| | MS OUTFIT | 45 | 0.99 | 0.11 | 1.26 | 0.79 |
| | SE | 45 | 0.03 | 0.00 | 0.04 | 0.03 |
| Task Model | MS INFIT | 45 | 1.13 | 0.30 | 2.19 | 0.76 |
| | MS OUTFIT | 45 | 1.19 | 0.43 | 2.75 | 0.69 |
| | SE | 45 | 0.03 | 0.00 | 0.03 | 0.02 |

Table 4.11 shows that the empirical MS Infit and MS Outfit have relative fit, with values close to or equal to 1.0. The values of the task model have exceeded 1.0 and are outside of the acceptable range. The standard deviation for the empirical indicates scores that are closely distributed around the mean, while the task model are widely distributed. The maximum scores for the task model are high and exceed the acceptable range, while the maximum and minimum scores for the empirical and task model are acceptable. All items for the empirical measure fit the expectations of the Rasch model.

Overall, the model standard error (*SE*) values for both the empirical and the task model are small, indicating that the fit of the data to the models is associated with small amounts of random noise.

Table 4.12

Summary Statistics across Persons for Empirical and Cognitive Task Model for Calculus 2013

| Variables | | N | MEAN | SD | MAX | MIN |
|---|---|---|---|---|---|---|
| Empirical | MS INFIT | 7942 | 1 | 0.13 | 1.61 | 0.61 |
| | MS OUTFIT | 7942 | 0.99 | 0.26 | 5.23 | 0.36 |
| | SE | 7942 | 0.37 | 0.06 | 1.03 | 0.33 |
| Task Model | MS INFIT | 7942 | 1.12 | 0.15 | 1.74 | 0.59 |
| | MS OUTFIT | 7942 | 1.19 | 0.34 | 6.38 | 0.16 |
| | SE | 7942 | 0.37 | 0.07 | 1.04 | 0.34 |

According to Table 4.12, the MS Infit and MS Outfit values for the 7942 persons for the measures are close to or equal to 1.0, within the acceptable range. The MS Infit and MS Outfit values for the task models just exceed the acceptable range. The standard deviation for the task model and the empirical are very similar in that the persons are relatively widely distributed along the scale. The maximum and minimum values are acceptable.

The standard error shows that there are some noise associated with the precision of measurement of the persons.

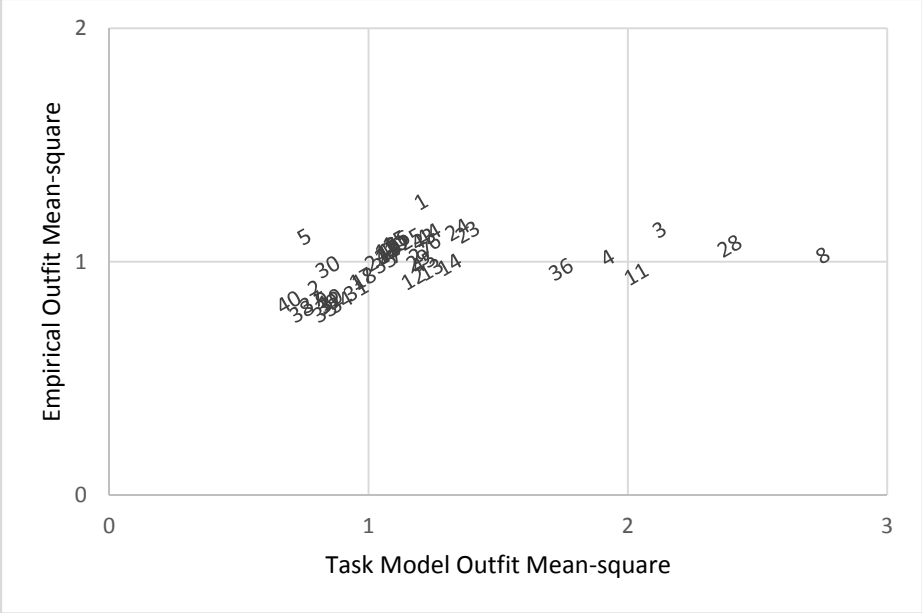



Figure 4.33 Scatterplot of Misfitting MS Outfit Items for Empirical and Cognitive Task Model for Calculus 2013.

Figure 4.33, shows the scatter plot of MS Oufit for the empirical and task model. The scatterplot reveals that items 8 28, 3, 11, 4 and 36 have the most extreme MS Outfit values and are outliers for the cognitive task model. There are no items or outliers for the empirical with extreme values and considered to be outliers.

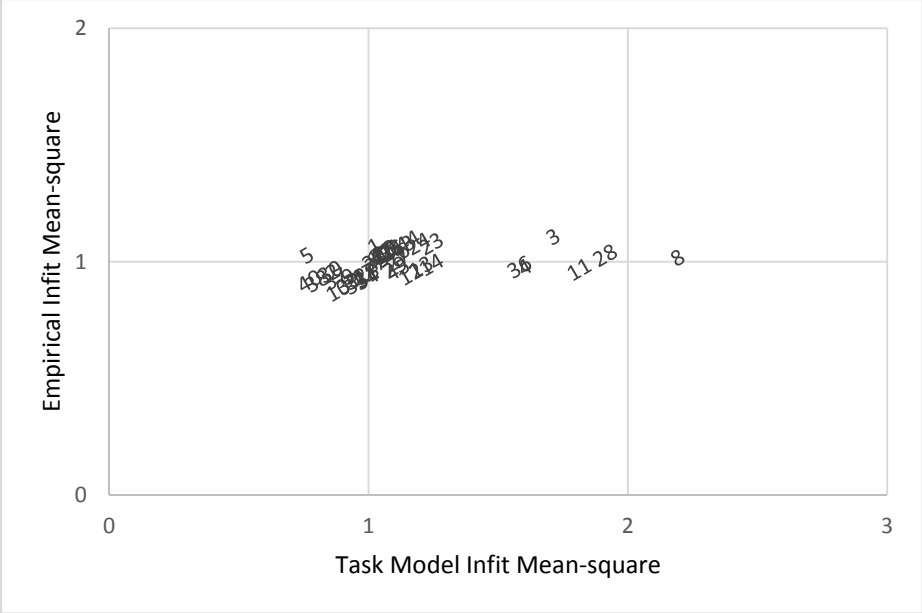



Figure 4.34 Scatterplot of Misfitting MS Infit Items for Empirical and Cognitive Task Model for Calculus 2013.

According to Figure 4.34, the four most misfitting items with extreme MS Infit values for the cognitive task model are 8, 28, 11, 3 4 and 36, which signals irregularities in the data. There are no items or outliers for the empirical with extreme values.

Table 4.13

Summary Statistics of Misfitting Items for Calculus 2013

| Items | P Value | Infit | Outfit |
|---|---|---|---|
| 3 | 0.53 | 1.71 | 2.12 |
| 4 | 0.69 | 1.60 | 1.92 |
| 8 | 0.76 | 2.19 | 2.75 |
| 11 | 0.57 | 1.81 | 2.03 |
| 14 | 0.42 | 1.24 | 1.31 |
| 23 | 0.37 | 1.24 | 1.38 |
| 24 | 0.35 | 1.19 | 1.34 |
| 28 | 0.32 | 1.91 | 2.39 |
| 36 | 0.68 | 1.58 | 1.74 |

Table 4.13 shows the 9 cognitive task model items with high MS Infit and MS Outfit values. The high MS Infit values range from 1.19 to 2.19. The high MS Outfit values range from 1.31 to 2.75. There are no extreme values for the empirical data.

**Description of Misfitting Items for Calculus 2013**

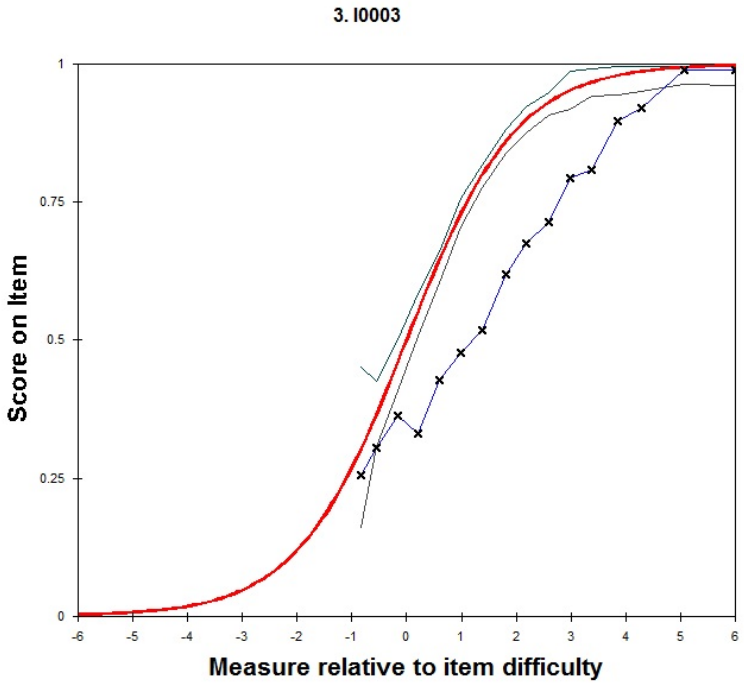



Figure 4.35 Empirical Conditional Mean Scores and Expected Response Function Item 3.

Figure 4.35, shows the expected response function and the empirical scores, conditional on the estimated empirical scores for item 3. This item is showing misfit particularly in the upper region of the scale where high performing examinees were unexpectedly scoring this item incorrectly. These examinees had the most misfitting response string and gave the most unexpected responses. However, they were not the most misfitting persons. 53 % of examinees scored this item correctly.

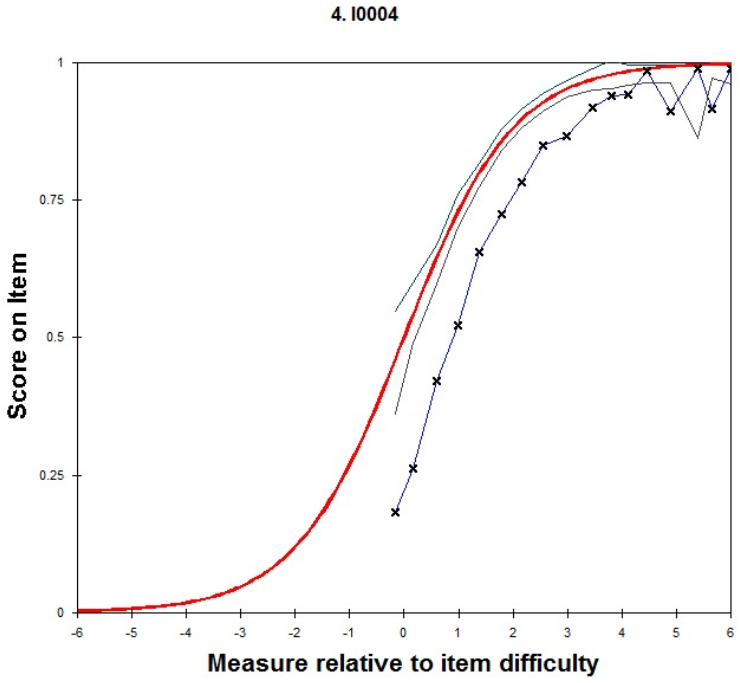



Figure 4.36 Empirical Conditional Mean Scores and Expected Response Function Item 4.

Figure 4.36, shows the expected response function and the empirical scores, conditional on the estimated empirical scores for item 4. The examinees is roughly tracking the model line. The item is showing misfit particularly at the upper regions of the scale at about 4.5 logits, where high performing examinees were unexpectedly scoring this item incorrectly. Some of these same examinees were the most misfitting, with the most misfitting response string and had the most unexpected responses to this item. The proportion of examinees who scored it correctly is 69 %.

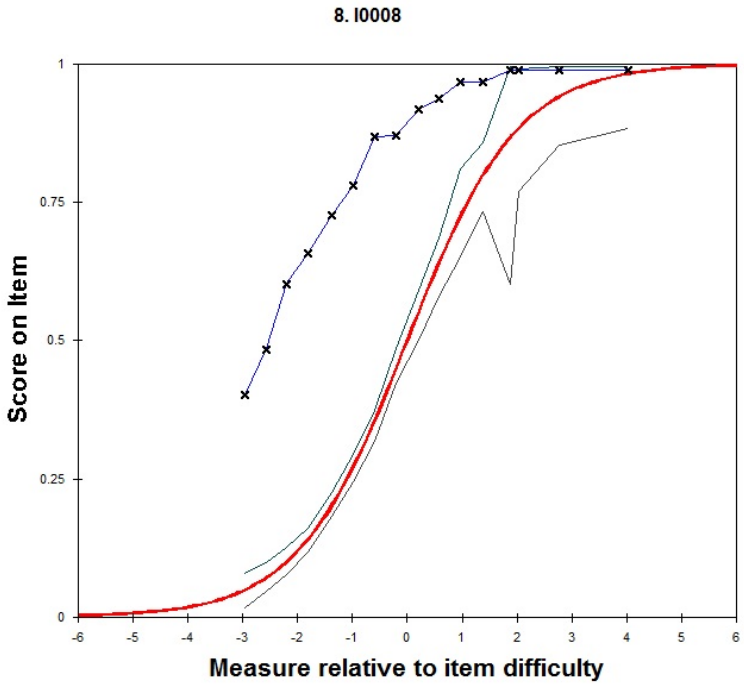



Figure 4.37 Empirical Conditional Mean Scores and Expected Response Function Item 8

Figure 4.37, shows the expected response function and the empirical scores, conditional on the estimated empirical scores for item 8. The examinees in the lower regions of the logit scale are misfitting the data, at about -2.0 logits where low performing examinees were unexpectedly scoring this item correctly. Examinees had the most misfitting and unexpected responses. The proportion of examinees who scored it correctly is 76 %.

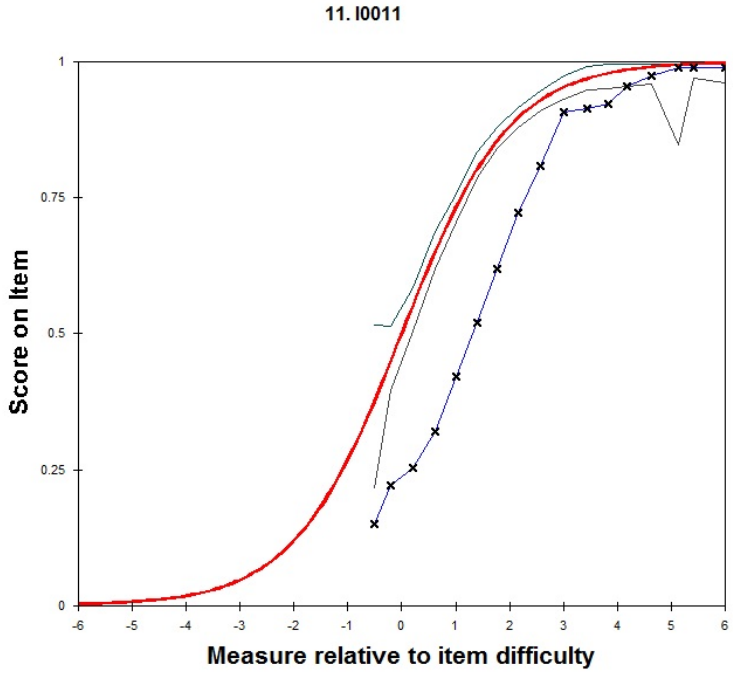



Figure 4.38 Empirical Conditional Mean Scores and Expected Response Function Item 11.

Figure 4.38, shows the expected response function and the empirical scores, conditional on the estimated empirical scores for item 11. The results show that the examinees are generally tracking the model line. The item is showing misfit at the upper region of the scale, at approximately 4.0 logits, where high performing examinees predicted to score correctly were unexpectedly scoring incorrectly. Some of these examinees were the most misfitting with the most misfitting responses. Others appear to be random examinees. Examinees may have made careless mistakes. 57 % of the proportion of examinees scored this item correctly.

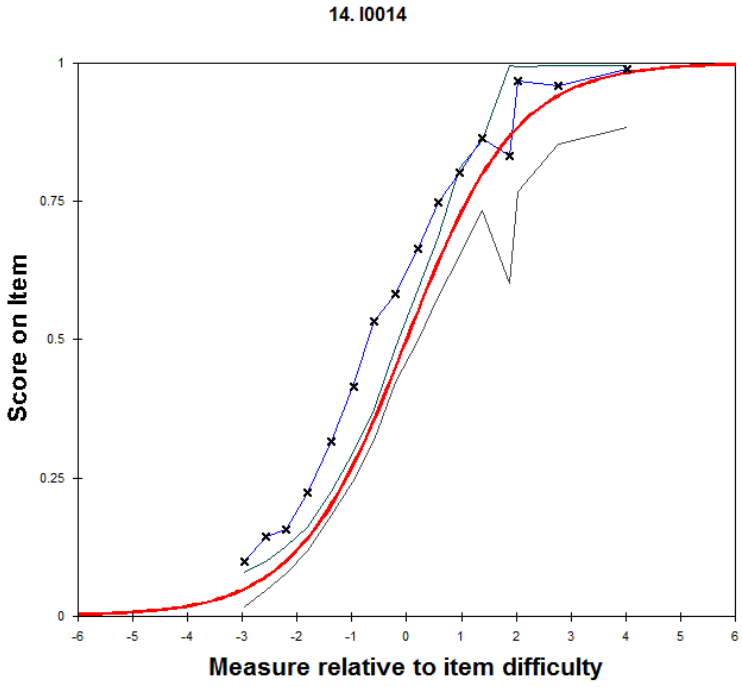



Figure 4.39 Empirical Conditional Mean Scores and Expected Response Function Item 14.

Figure 4.39, shows the expected response function and the empirical scores, conditional on the estimated empirical scores for item 14. The examinees are roughly tracking the model line. The item is showing misfit at the upper region at approximately 2.0 logits, where high performing examinees predicted to score correctly are scoring incorrectly. These same examinees had the most misfitting responses to this item. Misfit is also evident in the lower region of the scale at -2.0 logits where low performing examinees are scoring the item correctly. 42 % of the proportion of examinees scored this item correctly.

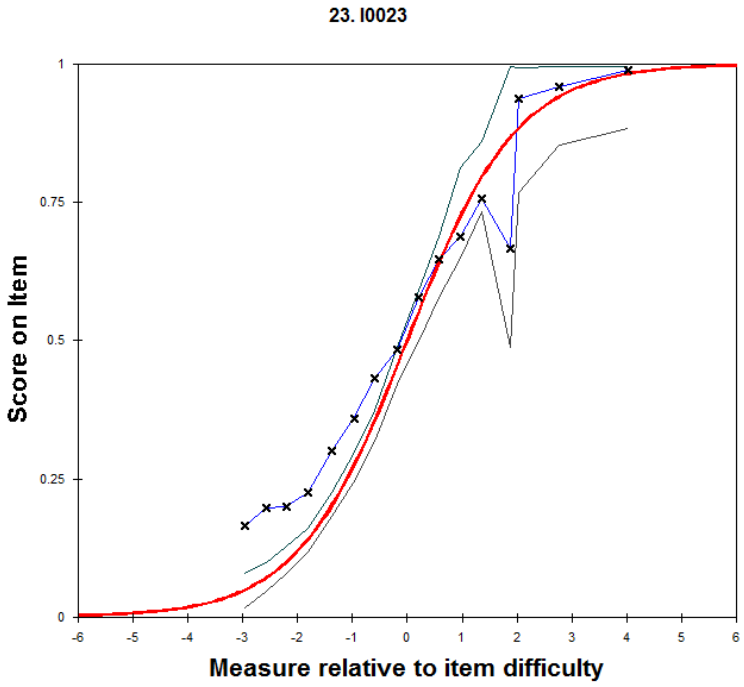



Figure 4.40 Empirical Conditional Mean Scores and Expected Response Function Item 23.

Figure 4.40, shows the expected response function and the empirical scores, conditional on the estimated empirical scores for item 23. This item is showing that generally, the examinees are tracking the model line. However, misfit is evident in the upper region where high performing examinees are unexpectedly scoring the item incorrectly. In the lower regions of the scale at about -2.0 logits, low performing examinees are unexpectedly scoring this item correctly. Some of the same examinees had the most misfitting and unexpected responses. 37 % of examinees scored this item correctly.

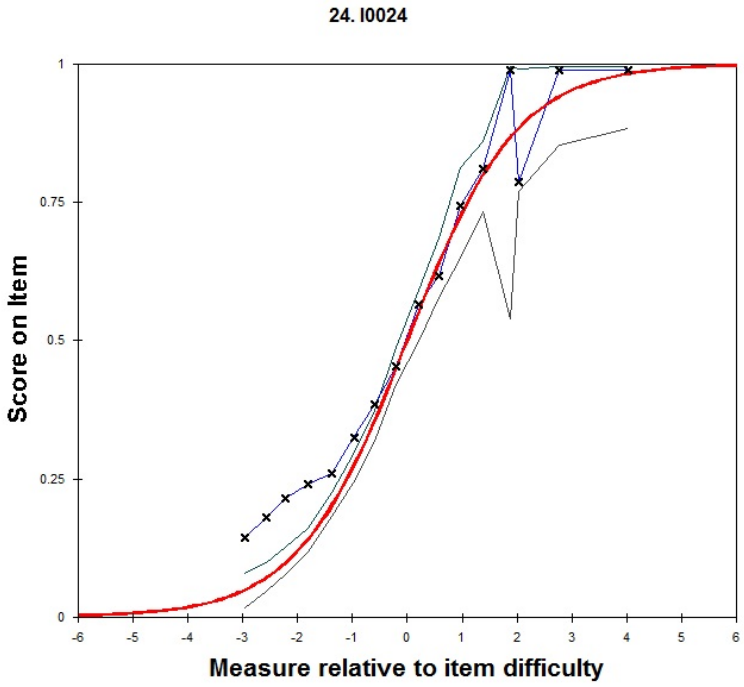



Figure 4.41 Empirical Conditional Mean Scores and Expected Response Function Item 24.

Figure 4.41, shows the expected response function and the empirical scores, conditional on the estimated empirical scores. The misfit is pronounced in the upper regions of the scale at about 2.0 logits, where high performing examinees were unexpectedly scoring incorrectly. In the lower region of the scale, some examinees were unexpectedly scoring the item correctly. The same examinees were the most misfitting, with the most misfitting response string and gave the most unexpected responses. Others were random examinees. This could be due to careless mistakes and guessing. 35 % scored this item correctly.

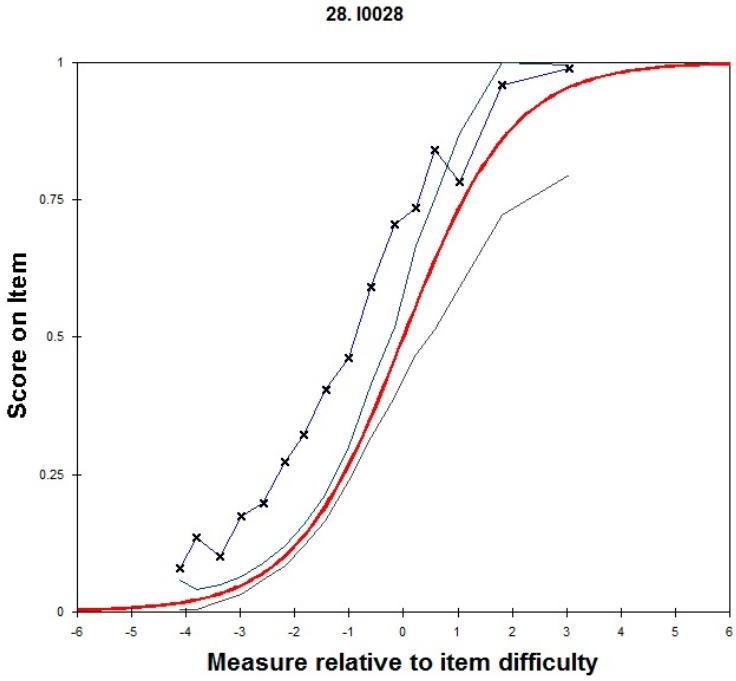



Figure 4.42 Empirical Conditional Mean Scores and Expected Response Function Item 28.

Figure 4.42, shows the expected response function and the empirical scores, conditional on the estimated empirical scores. This item is showing that examinees are roughly tracking the model line. However, misfits occur in the lower regions of the scale, about -3.2 logits, where low performing examinees are unexpectedly scoring this item correctly. Some of these examinees had the most misfitting response string and gave the most unexpected responses. Others were random examinees. There is evidence of misfit in the upper regions of the scale at about 1.0 logits where high performing examinees are unexpectedly scoring the item incorrectly. 32 % of the proportion of examinees scored this item correctly.

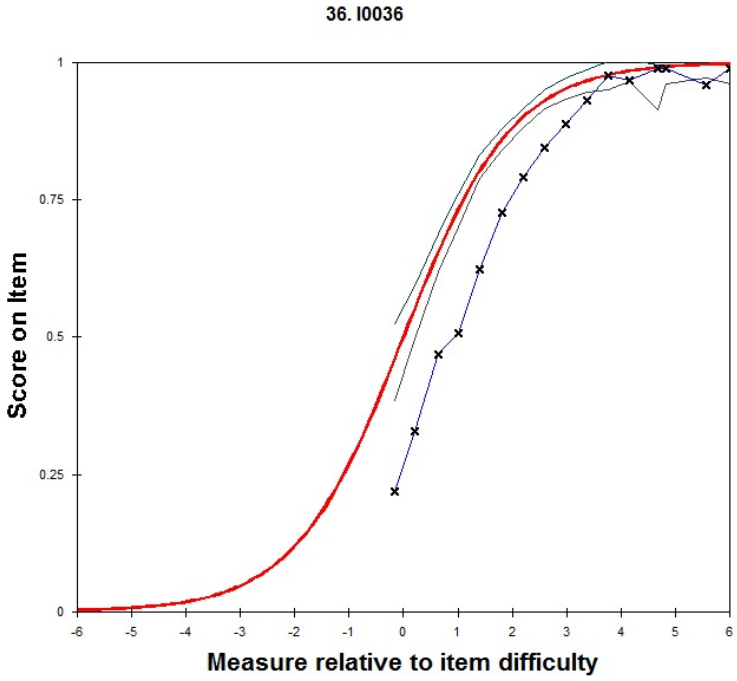



Figure 4.43 Empirical Conditional Mean Scores and Expected Response Function Item 36.

Figure 4.43, shows the expected response function and the empirical scores, conditional on the estimated empirical scores for item 36. This item is showing misfits in the upper regions of the scale, at about 3.5 logits, where high performing examinees were unexpectedly scoring this item incorrectly. These examinees had the most misfitting responses, and gave the most unexpected responses to this item. 68% of examinees scored this item correctly.

**The Effects of Cognitive Task Model on Proficiency Scores**

The second research questions relates to the impact of using cognitive task model derived complexity index scores to estimate examinees proficiency. This is important as the research findings can demonstrate the possibility of using assessment engineering cognitive task model to replace the need for pretesting items. This will have tremendous impact particularly in the reduction of the exponential cost incurred by testing companies for pretesting.

Table 4.14

Average Proficiency Estimates for Empirical and Task Model

| Subjects | Variables | N | $\bar{\hat{\theta}}$ | $SD\ \hat{\theta}$ | $SE\ \hat{\theta}$ |
|---|---|---|---|---|---|
| English 2012 | Empirical | 2172 | 0.952 | 1.025 | 0.022 |
| | Task Model | 2172 | 1.02 | 1.078 | 0.023 |
| English 2013 | Empirical | 4299 | 0.65 | 0.875 | 0.013 |
| | Task Model | 4299 | 0.677 | 0.911 | 0.014 |
| Calculus 2012 | Empirical | 4248 | 0.716 | 1.001 | 0.015 |
| | Task Model | 4248 | 0.737 | 1.069 | 0.016 |
| Calculus 2013 | Empirical | 7942 | 0.588 | 0.961 | 0.011 |
| | Task Model | 7942 | 0.532 | 1.058 | 0.012 |

According to Table 4.14, the average proficiency estimates for the cognitive task models and the empirical for each of the four assessment is quite similar. Overall, the average theta estimates of the examinees to perform on the assessments range from moderate to high. English 2012 assessment had the largest mean proficiency estimate, while the smallest mean proficiency estimate was with Calculus 2013. For both of these groups, the average ability ranges from moderately low to moderately high ability.

While all of the assessments had small standard error (SE) values, ranging from 0.011 to 0.023. The small mean standard error results indicate that proficiency estimates in these groups have less error associated with the values when computed than would higher SE. Calculus 2012 had the smallest SE 0.011. The highest SE values were estimated in English 2012, at 0.023.

The standard deviation (SD) for English 2012 is the largest, and indicates a large distribution of examinees across the scale for that subject. The lowest SD values was found with English 2013 and Calculus, 2013. All of the values are similar and close to 1.0.

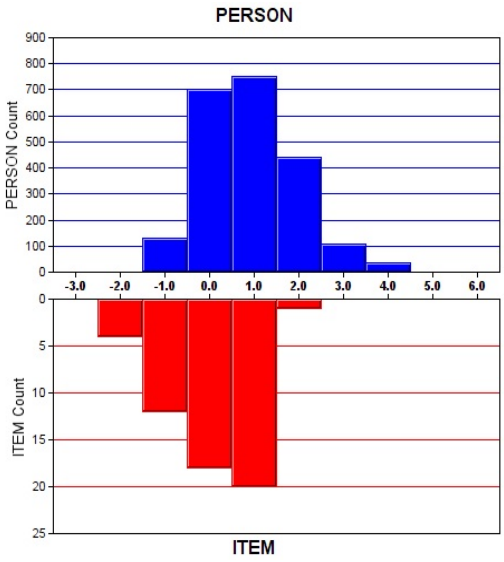



Figure 4.44 Empirical Item Person Map for English Language 2012

According to Figure 4.44, the person ability estimates range from approximately +4.5 to -1.5 logits, and the item estimates range from approximately +2.5 to -2.5 logits. The figure indicates that the items are not adequately targeting all of the examinees, as examinees located above 2.5 logits are not being targeted, which result in important information about them being lost. Overall, a large number of examinees are being covered by the items. Examinees will be able to score correctly the items that are located below their proficiency.

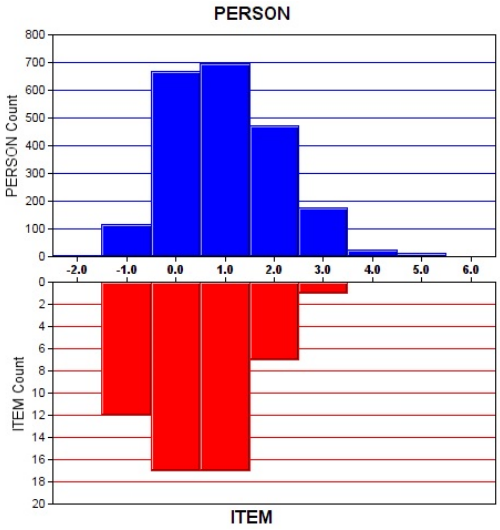



Figure 4.45 Cognitive Task Model Item Person Map for English Language 2012

According to Figure 4.45, the person ability estimates range from approximately +5.5 to -1.5 logits. Likewise, the item estimates range from approximately +4.5 to -1.5 logits. The items target most points along the logit scale. The distribution of items along the logit scale does not reflect the examinee population as some examinees above +4.5 logits were not catered for by the items, similarly examinees between +2.5 to +3.5 logits were excluded. The high ability examinees will miss some of the items but are able to score correctly all items below their proficiency. The frequency of easy items is highest.

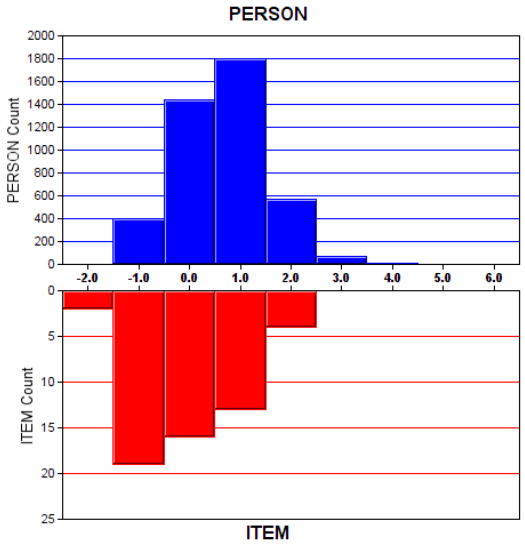



Figure 4.46 Empirical Item Person Map for English Language 2013

Figures 4.46 shows the person ability estimates range from approximately +3.5 to -1.5 logits. The item estimates range from approximately +2.5 to -2.5 logits. The items extend along the scale, disproportionately, and do not adequately cover those examinees above +2.5 logits. Generally, the items are targeting most of the examinees, even though it is disproportionate.

The frequency of easy items is high, and does not mirror the persons at that region of the scale. There is evidence of insufficient item targeting of items over the continuum to support examinee proficiency.

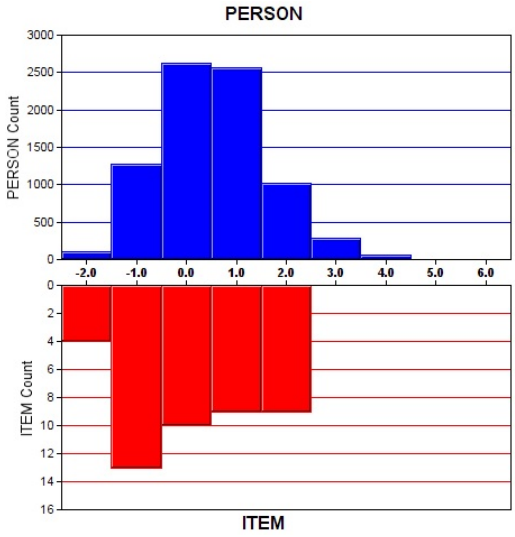



Figure 4.47 Cognitive Task Model Item Person Map for English Language 2013

According to Figure 4.47, the person ability estimates range from approximately +3.5 to -2.5 logits. The item estimates range from approximately +2.5 to -2.5 logits. The most able respondents on this measure are not being adequately targeted by the items and important information about these examinees is lost. It is also possible to predict the probability of a test taker's success on the rest of the items, which are below their ability. The highest frequency of items are between -1.5 and -0.5 logits.

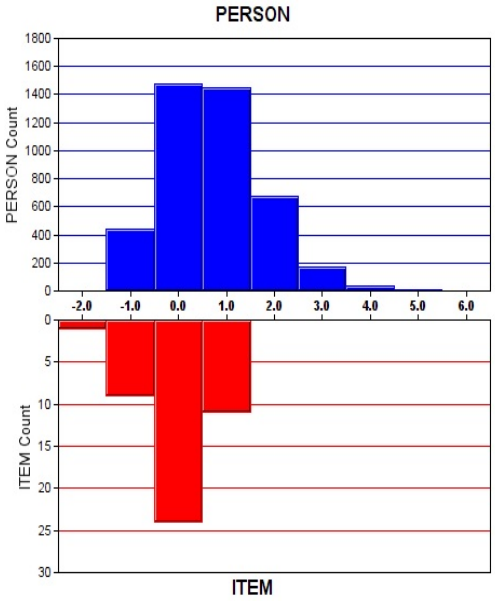



Figure 4.48 Empirical Item Person Map for Calculus 2012

Figure 4.48 shows the mirrored distributions of person proficiency estimates and item difficulties computed using for the empirical Calculus 2012 assessment for 45 items. According to Figure 4.48, the person ability estimates range from approximately +5.5 to -1.5 logits.

The item estimates range from approximately +2.5 to -1.5 logits. The items are not representative of the total target population, as not all items are targeted to examinees located above 1.5 logits on the scale. This result in valuable data about those high ability examinees being lost. In this example, the examinees overall are 'better' than the items. The examinees who are above 1.5 logits should be are able to score correctly on all of the

items below their proficiency. Likewise, the examinees should be able to score the items correctly that are below -1.5 logits, which is below their proficiency levels.

Most of the items are located between +0.05 and -0.05 logits. The person's ability does not match the item difficulty.

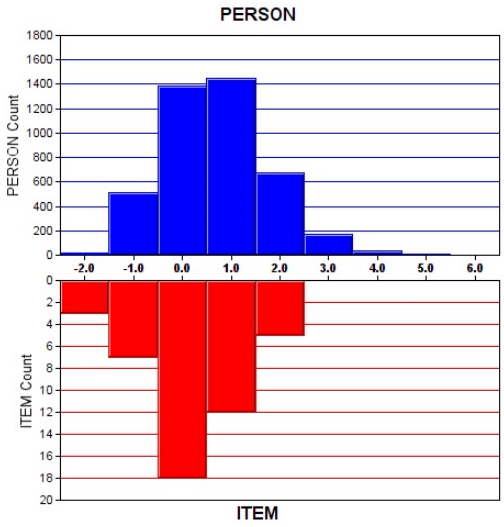



Figure 4.49 Cognitive Task Model Item Person Map for Calculus 2012

Figure 4.49 shows the mirrored distributions of proficiency scores and item difficulties for the task model Calculus items for 45 items. The person ability estimates range from approximately +5.5 to -2.5 logits, and the item estimates range from approximately +2.5 to -2.5 logits. The items do not target examinees located above +2.5 logits as too few difficult items provide adequate important information about the abilities of these high ability examinees. Conversely, the examinees on the lower levels

of the scale have easy items, which are not difficult enough to challenge them. Most of the items were located between +1.5 and -0.5 logits. There are no items above +2.5logits. For the cognitive task model, the person's ability appear to provide a more adequate match to the item difficulty on the English assessment.

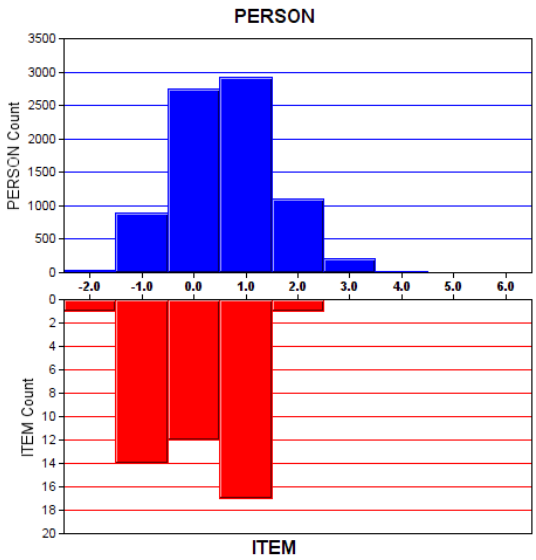



Figure 4.50 Empirical Item Person Map for Calculus 2013.

According to Figure 4.50, the person ability estimates range from approximately +3.5 to -2.5 logits. The item estimates range from approximately +2.5 to -2.5 logits. Thus, the items are not adequately measuring the most able respondents on this measure. There are no difficult items that would provide useful information about examinees abilities above +2.5 logits. Most of the items located between +0.5 to +1.5 are targeted to most of the examinees who mirror the items location. The hierarchy of items' difficulty show that the frequency of easy items was high, that of medium relatively high

and the items are the highest. This reflect the distribution of examinees in that region of the scale.

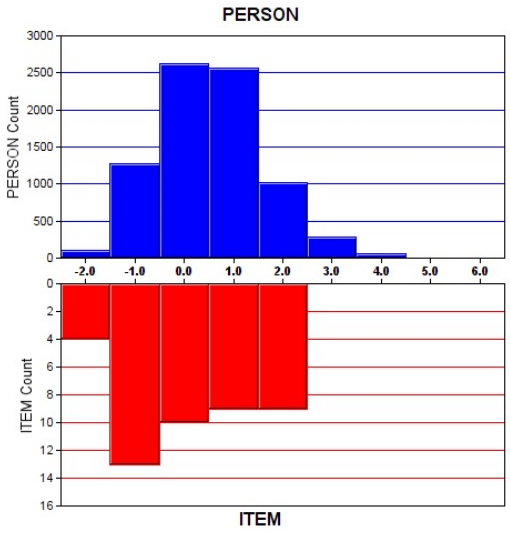



Figure 4.51 Cognitive Task Model Item Person Map for Calculus 2013

According to Figure 4.51, the person ability estimates range from approximately +4.5 to -2.5 logits. Likewise, the item estimates range from approximately +2.5 to -2.5 logits. There is a general spread of items across the scale, as almost all points along the scale are targeted. However, there are some examinees above 2.5 logits that are not being tapped by any item, which may result in important information about them being lost. It may be possible to predict the approximate result of a test taker's success on the given items that is below their ability level along the scale. The frequency of easy items are high as most items are located between -0.50 -1.50 logits.

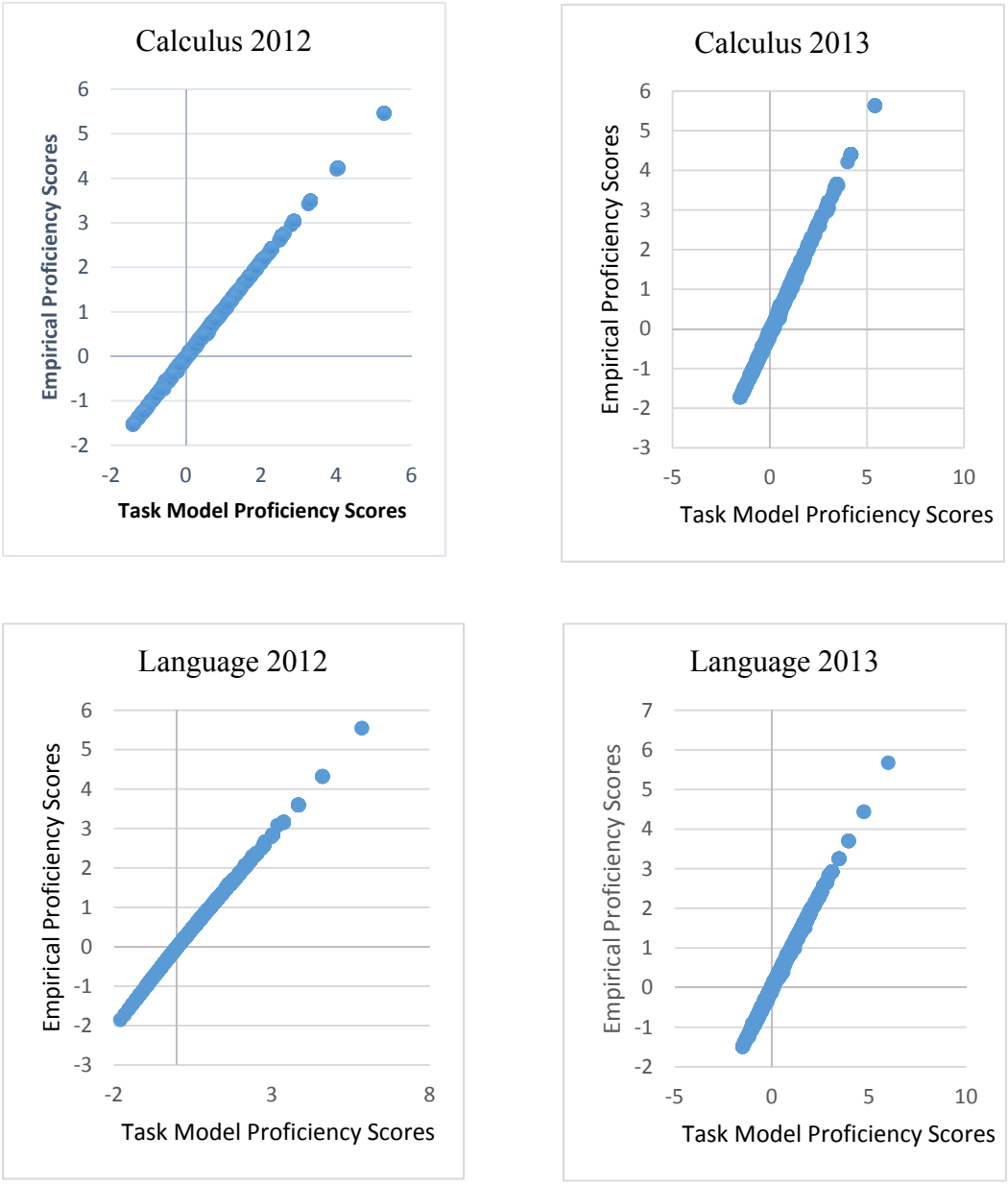



Figure 4.52 Scatterplot of Proficiency Scores for Calculus and English Language 2012 and 2013.

Figure 4.52, scatter plots display the relationship between the cognitive task model and empirical proficiency scores for the Calculus and English examinations. There is absolutely no difference in scoring the examinees using the cognitive task model complexity index score or scoring them with the empirical 'b' parameter scores. The examinees are therefore indifferent as to whether they are scored with the complexity scores or the empirical scores. This suggests that scoring examinees with either the empirical estimates or the cognitive task model estimates would yield similar results on these assessments.

# CHAPTER V

# CONCLUSIONS AND DISCUSSION

## Overview and Summary of Findings

The primary purpose of this dissertation is to investigate the comparability of Assessment Engineering Cognitive Task model derived parameter estimates with that of the statistical empirical Rasch Model. The cognitive task modeling design process makes it possible to estimate how difficult an item is, long before an examinee takes the test. This is usually accomplished without compromising the reliability and or validity of the test. Few studies have been conducted which compared the assessment engineering cognitive task modeling design process to the more traditionally estimated empirical difficulty parameters. Luecht, Burke and Devore (2009) had demonstrated that the task modelling process is capable of producing statistically similar results as the empirical estimates.

In this study, existing operational items from four criterion- referenced examinations were reversed engineered to develop cognitive task models or content blueprints that could potentially be used to develop thousands of multiple-choice items. The cognitive task modeling difficulty process is determined by careful design. The design process involved using task model grammars and making minor modifications to

cross validate the cognitive task models, and to evaluate the comparative and predictive power of the cognitive task model. Finally, task model grammar complexity index was computed to re-score the examinee data, which allowed for comparisons to be made between the task model proficiency scores and the empirical MLE scores. This chapter encapsulates the findings in reference to the research questions that were enunciated in chapter one, their implications, along with the limitations and suggestions for future research.

The first research question asked whether assessment engineering cognitive task model derived difficulty parameters and the empirical Rasch difficulty parameters will yield similar estimates. The high correlations between the cognitive Task model derived difficulty parameters and the empirical Rasch Model estimates across all four assessments as displayed in Table 1, have correlations that exceeds .800, which suggests that examinees maybe invariant as to whether the empirical or the task model parameter estimates are used. The R-squared values exceeded .600 suggesting that the variance between the task model and the empirical estimates are mostly accounted for.

Using the Winsteps fit statistics, as evidence in the mean square infit and mean square outfit, the results of this study demonstrates that at the item level, the cognitive task models had many misfitting items that were outside of the acceptable range, more than the empirical statistical model. Overall, the task modeling process worked well. In fact, a common misfitting item for both the task model and the empirical, item 17 displayed a larger MS Outfit value (1.65) for the empirical than for the task model (1.53).

This confirms that when misfits occurs, it is possible to redesign the features of the cognitive task model and so improve the task instead of discarding the item. In this way assessment engineering provides a viable alternative to test design through its innovative and detailed design approach to test blueprinting.

While this findings suggest that cognitive task modeling may be somewhat less accurate at producing difficulty parameters, as there was not a perfect relationship between cognitive task modeling and empirical examination data. The results from the four assessments show that the cognitive task modeling process is effective in providing very plausible insights into item location parameters along the scale. This finding also demonstrates the viability of using the cognitive task modeling approach for developing and incorporating difficulty and complexity into test items.

The second research question considered whether cognitive task model derived difficulty estimates can replace the Rasch model difficulty parameter estimates in scoring examinees. As shown in chapter IV, scoring the examinees with the task model derived difficulties appears to be as good as scoring the examinees with the empirical item difficulty estimates. For both the empirical and the task model, the SE values were found consistently low, indicating that both variables were stable and representative of the overall population. The cognitive task models were able to produce scores that were an accurate reflection of examinee's ability as the empirical in the English Language and Calculus examinations.

The results indicate that the cognitive task model captured proficiency just as effectively as the empirical data. This outcome implies that both the cognitive task

model and the empirical model are both equally effective at targeting proficiencies accurately. The empirical and cognitive task models were largely the same when proficiencies are compared. Assessment engineering cognitive task model difficulty can replace empirical difficulties for scoring examinees.

Overall, this study found that the cognitive task model difficulty parameters performed credibly well when compared to the empirically computed difficulty estimates, and that knowledge of the difficulty parameter a priori can result in reliable and valid estimates. In addition, a comparison of proficiency scores for the empirical and cognitive task model yielded similar results. The descriptive statistics showed that the amount of variation in item difficulties, that is, the standard deviations of the empirical Rasch model and the cognitive task model were generally similar. Overall, the task modeling process did not add any additional error or method variance that might detract from the quality of the item parameter estimates. The small size of the errors could be tolerated.
The result of this research dissertation, while useful to test developers and practitioners, cannot be generalized beyond this sample and the four assessments used for data collection.

## Practical Implications of the Results

Based on this study, the success of the cognitive task model derived difficulty parameters has proven to be informative. Knowledge of the cognitive task modeling process parameter difficulty estimates a priori can be used to guide test developers in the item selection process. Test developers can estimate how difficulty an item is before administering the test. This will provide them with considerably more flexibility in

distributing the test items throughout the test based on their level of difficulty. Hence, items would be selected at any point along the scale, which can provide the maximum amounts of information about examinee's ability and inform decisions regarding placements and interventions. Thus, items will be optimally chosen that will properly target each examinee. This will lead to increased accuracy between examinee and the difficulty of the test. Test designers and educators using assessment engineering cognitive task modeling design approach, will succeed in lining up or ranking items from the highest or most difficulty to the lowest or easiest along the scale, in order to minimize the mistakes of allocating too many easy items and too few difficult items on the test.

Test developers have a clear understanding of what they are measuring and how they are measuring it. This will impact the construct validity of the tests. Kane (2013) asserts that the valid interpretation and use of test scores requires a clear statement of the claims and assumptions a priori or that the evidence be adequately supported.

Through assessment engineering cognitive task models careful test design work, the model could specify the precise relationships between test items and ability scores, so that the intended outcomes can be realized.

In addition, test developers can now focus design features of the model instead of focusing on statistical models for item/test analysis. In this confirmatory approach, the design features built into the item a priori allow for difficulty estimates to be known without the use of data hungry psychometric models

The educator will have at their disposal, hundreds and thousands of items, uniquely crafted, and used without the fear of item exposure. This will result in cost effectiveness, as time and money that is usually spent on individual item pretesting can be eliminated. Thus the astronomical cost which are associated with item writing, and pilot testing can be reduced. In addition, the on-demand use of items for formative assessment with near to immediate feedback. Generally, the use of data hungry pscychometric, complex models will be reduced. Assessment engineering saves on cost of item development. This is very crucial today because of the high demand for more items especially in such areas as computer assisted testing (CAT) and other formative assessments. Hence the goal of assessment engineering which is to provide an extensive supply of low cost items will be facilitated.

## Limitations and Future Research

Assessment engineering cognitive task modeling is a very time consuming process. Designing the model and building the task models require a lot of up-front work.

More collaboration is needed between the subject matter experts and the test developers throughout the entire process. In addition the small numbers of subject matter experts provided a limitation in itself. As stated earlier, it is important for psychometricians, SME's and test developers to collaborate in the process.

For future research, it will be imperative that investigations be conducted to determine the relationship between the cognitive task model and the Rasch model given a finite amount of response data from which to estimate the model parameters. According

to Luecht (2013), the generation of assessment engineering item parameters do not rely on the need for data-hungry psychometric models.

To expand on the analysis to include distractor analysis, this will give a deeper understanding as to which option was not working and which was working best. It will also help to incorporate more qualitative feedback form the SME'S, which will help to deepen the research and provide much deeper information and interpretation into the item difficulty levels of misfitting items.

To experiment with different types of research designs such as experimental. By so doing, the researcher will be able to better determine the impact on students proficiency score gain over a specified period of time. In addition, administering the test to a control group and then re-administering the test after re-designing the task model for the misfitting items will provide more insight into the effectiveness of the task modeling process.

To use different types of test items in developing the task models, to include items such as matching, true and false and some short answer items. In addition, to include common items equating particularly for comparison between tests.

It is encouraged that the replicability of the task models be done with different datasets, particularly with a finite amount of response data from which to estimate the model parameters. Assessment engineering does not require large numbers of examinee data in order to compute item difficulties.

## REFERENCES


Arendasy, M.., Sommer, M., Gittler, G., & Hergovich, A. (2005). Automatic generation of quantitative reasoning items: A pilot study. *Journal of Individual Differences*, 27(1), 2-14.

Bachman, L. F., & Palmer, A. S. (2010). Language assessment in practice: Developing language assessment and justifying their use in the real world. Oxford UK. Oxford University Press.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics.* Vol. 17, No. 2, pp. 191 – 204.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Fredriksen, R. J.Mislevy, & I. I. Bejar (Eds.), Test theory for a new generation of tests (pp. 323-359). Hillsdale, NJ: Lawrence Erlbaum.

Bejar, I. I. (1996). Generative response modeling: Leveraging the computer as a test delivery medium (ETS Research Report No. 96-13). Princeton, NJ: Educational Testing Service.

Bennett, R, E. (2013). Preparing for the future: what educational assessment must do. Paper presented at the national center for research in evaluation standards and student testing. Los Angeles CA.

Burke, M., Devore, R., & Stopek, J. (2013). Implementing assessment engineering in the uniform certified public accountant (CPA) examination. *Journal of Applied Testing Technology,* 14, (1) 1-35.

Chipman, S. F., Schraagen, J. M., & Shalin, V. L. (2000). Introduction to cognitive task analysis. In J. M Schraagen, S. F. Chipman & V. J. Shute (Eds.), *Cognitive Task Analysis* (pp. 3-23). Mahwah, NJ: Lawrence Erlbaum Associates.

Chudowsky, N., & Pellegrino, J. W. (2010). Large-Scale Assessments That Support Learning: What Will It Take? Theory into Practice, 42(1), 75-83.

Clark, R. E.; Feldon, D. F.; van Merrienboër, J. J, G.; Yates, K. & Early, S. (2006). Cognitive Task Analysis.

Clauser, B. E., & Margolis, M. J. (2006). Book Reviews. In S. H. Irvine & P. C. Kyllonen (Eds.) Item Generation for Test Development. *International Journal of Testing*, *6*(3), 301-304.

Comrey, A. L. (1984). Comparison of two methods to identify major personality factors. *Applied Psychological Measurement, 8*, 397-408.

Drasgow, F., Luecht, R. M., & Bennett, R. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471-516). Washington, DC: American Council on Education.

Comrey, A. L. (1984). Comparison of two methods to identify major personality factors. *Applied Psychological Measurement, 8*, 397-408.

Embretson, S. E. (1985). Multicomponent latent trait models for test design. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics. New York: Academic Press.

Embretson, S. E. (1994). Application of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107–135). New York: Plenum Press.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. Psychological Methods, 3, 300-396.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*, 407–433.

Embretson, S. E., & Daniel, R. C. (2008). Designing cognitive complexity in mathematical problem solving items. Paper presented at the annual meeting of the American Educational Research Association. New York, NY: March.

Embretson, S. E., & Wetzel, C. D. (1987a). Component latent trait models for paragraph comprehension. *Applied Psychological Measurement*, 11, 175-193.

Embretson, S. E., & Yang, X. (2007). Automatic item generation and cognitive psychology. In C. R. Rao & S. Sinharay (Eds.) *Handbook of Statistics: Psychometrics, 26* (pp. 747-768). North Holland, UK: Elsevier.

Ferrara, S., & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education/ Praeger.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. Acta Psychologica, 37, 359-374.

Fulcher, G., & Davidson, F. (2007). Language testing and assessment: An advanced resource book. London and New York: Routledge.

Geerlings, H.; Glas, C. A. W., & van der Linden, V, J. (2011). Modeling rule-based item generation. *Psychometrika* 76, 337–359

Geerlings, H., van der Linden, W.J. & Glas, C. A. W. (2012). Optimal test design with rule-based item generation. *Applied Psychological Measurement* 37(2) 140–161.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). Bayesian data analysis. New York: Chapman & Hall.

Gierl, M. J., Fung, K. Lai, H., Zheng, B., (2013). Using Automated Processes to Generate Test Items in Multiple Languages Paper Presented at the Symposium "Advances in Automatic Item Generation" Annual Meeting of the National Council on Measurement in Education. San Francisco, CA.

Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). Automatic Item Generation: Theory and Practice. New York, NY: Routledge.

Gierl, M. J., & Lai, H. (2012). Methods for creating and evaluating the item model structure used in automatic item generation (pp. 1-30). Alberta: Centre for Research in Applied Measurement and Evaluation, University of Alberta.

Gierl, M. J., & Lai, H., (2013). Using weak and strong theory to create item models for automatic item generation: some practical guidelines and examples. In Gierl, m & Haladyna, T. (eds) Automatic Item Generation Theory and Practice. Routledge.

Gierl, M. J; & Leighton, J. (2010). Developing Construct Maps to Promote Formative Diagnostic Inferences Using Assessment Engineering. Paper Presented at the Annual Meeting of the National Council on Measurement in Education. Denver, CO, USA.

Gierl, M. J.; Leighton, J. P.; & Hunka, S. M. (2007). Using the attribute hierarchical method to make diagnostic inferences about examinees cognitive skills. In Leighton & Gierl (eds.) Cognitive Assessment for Education. Theory and Applications. Cambridge.

Gierl, M. J., Zhou, J., & Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *Journal of Technology, Learning, and Assessment, 7, 1-51.*

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Markov chain Monte Carlo in practice. London: Chapman and Hall.

Glas C, A. W., & van der Linden, W, J. (2001). Modeling variability in item parameters in CAT. Paper presented at the North American Psychometric Society Meeting, King of Prussia, PA.

Glas C, A. W., & van der Linden, W, J. (2003) computerized adaptive testing with item cloning. *Applied Psychological Measurement,* 27, 247–261.

Gomez, P. G., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test.

Gorin, J. S. (2005). Manipulating Processing Difficulty of Reading Comprehension Questions: The Feasibility of Verbal Item Generation. Journal of Educational Measurement, 42, (4), 351-373.

Gorin, J, S., (2006). Test Design with cognition in mind. *Journal of Educational Measurement: Issues and Practice*. Vol 25. Issue 4. P. 21-35.

Gorin, J, S., (2007). Test construction and diagnostic testing. In Leighton & Gierl (eds.) Cognitive Diagnostic Assessment for Education. Theory and Applications. Cambridge University Press.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, 30, 394-411.

Hambleton, R. K. (1997). Enhancing the validity of NAEP achievement level score reporting. Proceedings of achievement levels workshop. National Governing. Board, Washington, DC.

Hendrickson, A., Ewing, M., Kaliski, P., & Huff, K. (2013). Evidence-Centered Design: Recommendations for Implementation and Practice. *Journal of Applied Testing Technology*, 14, 1-27.

Hendrickson, A., Huff, K., & Luecht, R. (2010). Claims, evidence, and achievement-level descriptors as a foundation for item design and test specifications. *Applied Measurement in Education*, 23: 358–377.

Humbo, C., & Dresher, A. (2001, April). A simulation study of the impact of automatic item generation under NAEP-Like data conditions. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Irvine, S. H., & Kyllonen, P. C. (2002). Item Generation for Test Development. Hillsdale, NJ: Erlbaum.

Johnson, M. S., & Sinharay, S. (2005). Calibration of polytomous item families using Bayesian hierarchical modeling. *Applied Psychological Measurement*, 29, 369-399.

Jonassen, D. H., Tessmer, M., & Hannum, W. H. (1999). Task analysis methods for instructional design. Mahwah, NJ: Lawrence Erlbaum Associates.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*. 50(1), 1-73.

Kirsch. I., Jungebult, A., Jenkins, L., & Kolstad, A. (1993). Adult literacy in America: A first look at the findings of the National Adult Literacy Survey. Washington, DC: National Center for Education Statistics; US Department of Education.

Klein, G. (1998). Sources of power: How people make decisions. Cambridge., MA: MIT Press.

Lai, H., Gierl, M., & Alves, C. (2010, April). Generating items under the assessment engineering framework. Invited symposium paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Leighton, J. P. (2004). Avoiding misconceptions, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. Educational Measurement: Issues and Practice, 23 (4), 6-15.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26,* 3-16.

Linacre, J. M. (2009). *Winsteps Rasch*-model computer program (Program manual 3.69.0). Downloaded from www.winsteps.com.

Luecht, R. M. (2006a) Designing tests for pass-fail decisions using item response theory. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.

Luecht, R. M. (2006b). Engineering the test: Principled item design to automated test assembly. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

Luecht, R. M. (2007). An introduction to assessment engineering for automatic item generation. In Leighton & Gierl (Eds.). Cognitive Assessment for Education. Theory and Applications. Cambridge.

Luecht, R. M. (2007a). Assessment engineering in language testing: From data models and templates to psychometrics. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Luecht, R. M. (2007c,). Assessment engineering: An integrated approach to test design, development, assembly, and scoring. Keynote and workshop presented at the Performance Testing Council Summit, Scottsdale, AZ.

Luecht, R. M. (2008). Assessment engineering in test design, development, assembly, and scoring. Keynote presented at the Annual Meeting of East Coast Language Testing Organizations (ECOLT), Washington, DC.

Luecht, R. M. (2008a, February). *Assessment engineering*. Session paper at Assessment Engineering: Moving from Theory to Practice, Coordinated panel presentation at the Annual Meeting of the Association of Test Publishers, Dallas, TX.

Luecht, R. M. (2008b, February). The application of assessment engineering to an operational licensure testing program. Paper presented at the Annual Meeting of the Association of Test Publishers, Dallas, TX.

Luecht, R.M. (2009). Adaptive computer-based tasks under an assessment engineering paradigm. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.

Luecht, R. M. (2012). Computer-based and computer-adaptive testing. In K. Ercikan, M. Simon & M. Rousseau (Eds.), Improving large scale assessment in education: Theory, issues, and practice (pp. 91-114). New York: Taylor-Francis/Routledge.

Luecht, R., (2013. Assessment Engineering Task Model Maps, Task Models and Templates as a New Way to Develop and Implement Test Specifications. Journal of Applied Testing Technology, 14, 1-38.

Luecht, R. M., Burke, M., & Devore, R. (2009, April). Task modeling of complex computer-based performance exercises. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Luecht, R. M., Dallas, A., & Steed, T. (2010). Developing assessment engineering task models: A new way to develop test specifications. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Lueong, S.C. (2006). On varying the difficulty of test items. Paper presented at the 32nd Annual Conference of the International Association for Educational Assessment, Singapore.

Linn, R. L. & Gronlund, N. E. (2000). Measurement and assessment in teaching. Columbus, OH: Merrill.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

Masters, J. S. (2010). A Comparison of Traditional Test Blueprinting and Item Development to Assessment Engineering in a Licensure Context. Unpublished doctoral dissertation. University of North Carolina at Greensboro.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement. Washington, DC: American Council on Education.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13-23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741-749.

Michel, R. (2007). Cluster Analysis as a guide to the interpretation of quantitative m item models. Paper presented at the 52nd Annual Meeting of the Florida Educational Research Association in Tampa, Florida.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439–468.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257-306). Washington, DC: American Council on Education.

Mislevy, R. (2009). Validity from the perspective of model-based reasoning. In R. Lissitz (Ed.), The concept of validity (pp. 83–108). Charlotte, NC: Information Age Publishers.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. Research report. Educational Testing Service.

Mislevy, R J., Steinberg, L S., & Almond, R. (1999). Evidence-centered assessment design. Educational Testing Services. Princeton, NJ.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–67.

Mislevy, R. J., Wingersky, M. S., & Sheehan, K. M. (1994). Dealing with uncertainty about item parameters: Expected response functions (ETS Research Report 94-28-ONR). Princeton, NJ: Educational Testing Service.

Mosenthal, P.B., & Kirsch, I.S. (1991). Toward an explanatory model of document literacy. Discourse Processes, 14, 147–180.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Natar, C.; Zuelke, D.; Wilson, J & Yunker, B. (2004).The table of specification: insuring accountability in teacher made tests. *Journal of Instructional Psychology*, 31, 115-129.

National Research Council. (2001). Knowing what students know: The science and design of educational assessment. J.W. Pellegrino, N. Chudowsky, & R. Glaser (Eds.). Washington, DC: National Academy

Olsen, J, B., Olsen, J. A., & Smith, R.W. (2010). Investigating alternative approaches for analyzing item/task model. Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Denver CO.

Perea, L. (2011). Benefits of Teachers' Feedback to Reverse-Engineering Item Language Test Specifications from an Existing Item Bank. *Texas Papers in Foreign Language Education.* 15, 30-54.

Race, P. (2009). Designing assessment to improve physical sciences learning. A Physical Sciences Practice Guide. The Higher Education Academy. Physical Sciences Center.

Sinharay, S., & Johnson, M. S. (2008). Use of item models in a large-scale admissions test: A case study. *International Journal of Testing, 8,* 209-236.

Sinharay, S., & Johnson, M. (2013). Statistical modeling of automatically generated items. Automatic item generation: Theory and practice. Pp. 183-195.

Sinharay, S., Johnson, M. S., & Williamson, D. M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Statistics, 28,* 295-313.

Scheuneman, J., Gerritz, K. and Embretson, S. (1991). Effects of prose complexity on achievement test item difficulty. (ETS Research Report No. RR-91-43). Princeton, NJ: Educational Testing Service.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 307-353). Washington, DC.

Shu, Z., Burke, M., & Luecht, R. M. (2010). Some Quality Control Results of Using a Hierarchical Bayesian Calibration System for Assessment Engineering Task Models, Templates, and Items. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Snow, R. E., & Lohman, D. F. (1989). Implication of cognitive psychology for education measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). New York: Macmillan

Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. Educational Measurement: Issues and Practice, 18(1), 23–27.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item mapping method. *Journal of Educational Measurement*, *40*(3); 231-253.

Wang, N., Wiser, R., & Newman, L. (2001). Use of the Rasch IRT model in standard setting: An item mapping method. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Webb, N. L. (1999). Criteria for alignment of expectations and assessments in mathematics and science education. Research Monograph No. 6. Wisconsin Center for Educational Research. Madison, WI.

Williamson, D. M., Johnson, M. S., Sinharay, S., & Bejar, I. I (2002b). Applying hierarchical model calibrations to automatically generated items. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Wilson, M.; Adams, R. J., (1995). Rasch models for item bundles. Psychometrika. Vol 60, Issue 2, pp181-198. Wright, B., & Stone, M. (1979). *Best test design.* Chicagol, IL: MESA Press.

Zhou, J., (2009). A review of assessment engineering principles with select applications to the certified public accountant examination. A Technical Report for the American Institute of Certified Public Accountants. Centre for Research in Applied Measurement and Evaluation. Canada.

Zumbo, B. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), Handbook of statistics, Vol. 26: Psychometrics (pp. 45–79). Amsterdam. The Netherlands: Elsevier Science.

# APPENDIX A

## TASK MODEL GRAMMARS FOR ENGLISH LANGUAGE AND CALCULUS

Table 16

Task Model Grammars for English Language 2012 -2013

| TMG's | Definition | Complexity Ratings |
|---|---|---|
| **Easy** | | |
| Identify | To recognize or be able to name something or someone or to say what it is | 1 |
| Define | To give the meaning of a word or phrase | 1 |
| Explain | To make something clear or easy to understand | 1 |
| Infer | To form an opinion based on the evidence or facts given. To hint at something | 1 |
| Compare | To look at closely, to note the similarity or difference between things | 1 |
| Describe | To give an account of something | 1 |
| Imply | To suggest without saying or showing clearly or plainly | 1 |
| **Medium** | | |
| Interpret | To explain the meaning of something that is understood in a specified way | 2 |
| Explain | To make something clear or easy to understand, to expand | 2 |
| Infer | To form an opinion based on the evidence or facts given. To hint at something | 2 |
| Analyze | To examine critically or carefully in detail in order to identify the key factor | 2 |
| Compare | To look at closely, to note the similarity or difference between things | 2 |
| Assess | To judge or decide on the value or importance of something | 2 |
| Apply | To use a particular method or process or technique to be used on/for something | 2 |
| Distinguish | To notice or recognize differences between two or more things | 2 |
| Describe | To give an account of something | 2 |
| **Hard** | | |
| Evaluate | To make a judgment about something after considering carefully its value, Importance or originality | 3 |
| Summarize | To state or express in a concise form | 3 |
| Explain | To make something clear or easy to understand | 3 |
| Conclude | To form a final judgment about something or to reach a logical end based on evidence | 3 |
| Inference | An assumption or conclusion that is rationally and logically made, based on the given facts or circumstances | 3 |
| Assess | To judge or decide on the value or importance of something | 3 |
| Distinguish | To notice or recognize differences between two or more things | 3 |

Table 17

Task Model Grammars for Calculus 2012 -2013

| TMG's | Definition | Complexity Ratings |
|---|---|---|
| **Easy** | | |
| Recognize | To be able to identify a equation, formula, concept, principle or algebraic expression | 1 |
| Recall | To remember a mathematical operation from what was previously done | 1 |
| Identify | To be able to use mathematical procedures, operations or properties | 1 |
| Sum | To aggregate two or more numbers or total | 1 |
| Re-draw | To sketch or trace figures or lines | 1 |
| Use | To make use of mathematical operations in order to solve a problem | 1 |
| **Medium** | | |
| Apply | To put to a specific use in a given situation such as using a particular method Or formula | 2 |
| Calculate | To work-out or solve the value of something using a specific procedure | 2 |
| Interpret | To show how to extract the meaning out of an equation or formula or mathematical expression | 2 |
| Substitute | The replacement of a term of an equation or expression by another that is Known to have the same value in order to simplify the equation | 2 |
| Compute | Methods of solving a mathematical problem | 2 |
| Simplify | To reduce an equation, fraction etc to its simplest form | 2 |
| Rewrite | To write or expand a function or mathematical expression so as to show the products, sum etc of its factors | 2 |
| Analyze | A method of proving a problem by working backward to something that is known to be true based on the properties of numbers | 2 |
| Solve | To find or work out, use a solution to a problem | 2 |
| Determine | To specify, fix or define the position or limit | 2 |
| Classify/ Categorize | To put into groups based on common properties | 2 |
| Connect | Having a continuous path between two points, such that either it or its converse holds between two members of its domain Used of a curve, set or surface | 2 |
| **Hard** | | |
| Compare & Contrast | A strategy used to find a solution to a problem by examining the different methods for commonality in formulas, symbols etc. and difference and deciding on the best or most efficient one to solve the equation | 3 |

| | | |
|---|---|---|
| Evaluate | To find or determine the value of or to solve an expression or equation | 3 |
| Construct | To draw a line, angle or figure so that it meets specific requirements | 3 |
| Demonstrate | To provide mathematical prove or show how to operate or work an equation, formula etc. | 3 |

# APPENDIX B

## SUBJECT MATTER EXPERTS TASK DESCRIPTIVE STATEMENTS

Table 18

Subject Matter Experts Task Descriptive Statements for English Language 2012 - 2013

| TASK DESCRIPTION | PERCEIVED DIFFICULTY |
|---|---|
| Read passage\|Identify purpose\|Select Option | Easy |
| Identify meaning\|Make comparison\|Read sentences | Easy |
| Infer meaning from phrase\|Read\|Identify meaning | Moderate |
| Read passage\|Interpret phrase\|Identify meaning of Phrase | Moderate |
| Read paragraphs\|Compare paragraphs\| | Easy |
| Read sentence\|Interpret meaning in sentence\| | Moderate |
| Define word | Easy |
| Read paragraph\|Analyze options\|Draw conclusion\|Identify meaning | Moderate |
| Evaluate Sentences\|Analyze Assumptions\|Read sentences\| | Hard |
| Identify purpose of passage\|Read passage\|Examine meaning | Easy |
| Identify type of argument\|Define words\|Analyze passage\| | Moderate |
| Identify major points in passage\|Read passage\|Locate words | Easy |
| Define word\|Identify Words | Easy |
| Define words\|Differentiate between options\|Identify meaning | Easy |
| Make inference\|Idenfify meaning\|Define words\|Analyze options | Easy |
| Describe organization of passage\|Analyze passage\|Define words\| | Moderate |
| Infer meaning of words\|Read sentences\|Explain words | Easy |
| Explain paragraph\|Draw conclusion\|Understand paragraph | Moderate |
| Infer meaning from sentences\|Explain sentences\|Analyze sentence | Moderate |
| Identify relationship\|Identify meaning\|Infer meaning | Moderate |
| Explain meaning\|Identify option | Easy |
| Analyze paragraph\|Explain meaning of phrases\|Differentiate text | Moderate |
| Evaluate sentences\| Analyze sentences\| Identify meaning | Moderate |
| Analyze sentence\|Infer meaning\|Make conclusion\|Read sentences | Moderate |
| Make comparison\|Identify option\|Make suggestion | Moderate |
| Infer meaning\|define terms\|Read Endnote | Easy |
| Identify example\|Explain meaning | Easy |
| Make comparison\|Identify option | Easy |

| | |
|---|---|
| Identify argument\|Make comparisons | Easy |
| Analyze phrases\|Identify meaning\|Read critically | Moderate |
| Describe word\|Identify meaning | Easy |
| Read paragraphs\|Describe meaning\|Identify effect | Easy |
| Identify meaning\|Explain meaning\|Read sentences | Easy |
| Identify meaning\|Explain meaning\| | Easy |
| Define words\|Explain sentences\| | Easy |
| Identify meaning\|Explain words and phrases\|Read sentences | Easy |
| Interpret phrase\|Define words\|Explain terms\| | Easy |
| Identify argument\|Analyze responses\|Define terms | Easy |
| Infer meaning\|Read critically | Easy |
| Describe passage\|Define words and phrases\|Explain meaning | Moderate |
| Draw  conclusion\|Summarize passage\| | Moderate |
| Define word\|Explain phrases\|Read passage critically\| | Moderate |
| Define words\|explain sentences\|Critique sentence\|Describe text | Hard |
| Evaluate sentences\|Explain sentences\|Identify author's strategy | Hard |
| Evaluate passage\|Infer meaning from phrases\|Define terms | Hard |
| Define phrase\|Infer meaning | Easy |
| Define phrase\|Interpret phrase | Moderate |
| Interpret phrase\|Define words\|Explain meaning | Moderate |
| Analyze phrases\|Make comparisons\|Explain meaning | Hard |
| Define words and phrases\|analyse sentence\|Draw conclusion | Hard |
| Identify purpose of sentence\|Evaluate sentence\|Interpet sentence | Hard |
| Analyze sentence\|Interpret meaning\|Read critically | Moderate |
| Evaluate passage\|Define terms\|Make descriptions\| | Moderate |
| Compare and contrast\|Identify meaning of theme\|Define words\| | Moderate |

Table 19

Subject Matter Experts Task Descriptive Statements for Calculus 2012 -2013

| TASK DESCRIPTION | PERCEIVED DIFFICULTY |
|---|---|
| Recognize chain rule \| Apply chain rule appropriately | Easy |
| Calculate derivative \| Solve operation \| Interpret meaning of operation \| Sustitute operations | Moderate |
| Interpret area \| Recognize operations \| Calculate area using geometry | Easy |
| Calculate integral \| Apply formula | Easy |
| Recognize geometric series \| Apply formula \| Simplify equation | Moderate |
| Rewrite integral \| Rewrite limits \| Rewrite to balance | Easy |
| Recognize equation \| Rewrite formula \| Apply chain rule \| Simplify | Easy |
| Apply formula \| Interpret tabular data \| Use initial value | Moderate |
| Apply ratio test \| Apply comparison test \| Recognize tests for convergence | Moderate |
| Apply rules for integration \| Simplify equation | Easy |
| Apply definitions of continuity and differentiation \| Sketch graph | Easy |
| Calculate function \| interpret function \| Apply rule | Moderate |
| Apply ratio test for convergence \| Solve expression | Hard |
| Recognize appropriate logistic model \| Categorize choices as not logistic | Hard |
| Apply FTOC \| Use, interpret relationships between functions | Moderate |
| Apply formula \| Simplify equations | Easy |
| Identify series \| Manipulate series | Moderate |
| Apply FTOC \| Calculate area \| Apply basic rules of integration | Moderate |
| Recognize slope of cone \| Apply rule \| Simplify algebra \| Solve equation | Moderate |
| Recognize partial fractions \| Apply Techniques \| Apply rules for integration \| Calculate logs \| Apply log properties | Moderate |
| Interpret horizontal asymptote \| Recall formula \| Determine merit of options | Hard |
| Recognize power series \| Apply rules of convergence \| Determine merit of options | Moderate |
| Identify rate of change \| Determine change | Moderate |
| Recognize integration by parts \| Apply rules \| Connect relationships | Moderate |

| | |
|---|---|
| Apply integral techniques of limits \| Use U-substitution \| Apply limits at rules properly | Moderate |
| Apply rule for derivative \| Recall functions \| Substitute function \| Evaluate trig functions | Easy |
| Recognize series \| Apply rules for series and root test | Moderate |
| Apply rule \| Recognize derivatives \| Interpret used of function | Easy |
| Examine graph \| Interpret graph | Easy |
| Compare rates of change \| Examine graph | Moderate |
| Apply rules for integrals \| Calculate area \| Simplify equation | Moderate |
| Construct polynomial \| Simplify equation | Moderate |
| Determine value of each statement \| Interpret graph | Moderate |
| Recall points of inflection \| Interpret function \| Analyze or interpret tabular data | Moderate |
| Apply formula \| Use key function | Easy |
| Compare and contrast answers \| Apply conditions for continuity | Easy |
| Interpret rules for concave function \| Evaluate from calculator \| Graph function | Moderate |
| Interpret range of speed \| Apply chain rule \| Use calculator to evaluate change | Moderate |
| Interpret relationship between functions \| Recognize function \| Determine displays of correct response | Moderate |
| Recall formula for volume \| Evaluate answer on calculator | Easy |

# APPENDIX C

## TASK MODEL CODING SCHEMA AND SCORING INDICES

Table 20

Task Model Coding Schema and Scoring Indices for English Language 2012-2103

| **Required Task Actions** | **Difficulty/Complexity** |
|---|---|
| Identify_purpose_meaning.simple | 1 |
| Explain/Infer_phrase.simple | 1 |
| Identify_words.simple | 1 |
| Infer_meaning_words.simple | 1 |
| Explain_meaning.simple | 1 |
| Identify_phrases.simple | 1 |
| Compare_ideas.simple | 1 |
| Explain meaning_phrases.simple | 1 |
| Define_word.simple | 1 |
| Compare_sentence_endnote.simple | 1 |
| Identify_sources.simple | 1 |
| Explain_phrase.simple | 1 |
| Describe_words.simple | 1 |
| Define_phrases.simple | 1 |
| Analyze_relationship.moderate | 2 |
| Summarize_meaning.moderate | 2 |
| Analyze_paragraph.moderate | 2 |
| Infer_sentences.complex | 2 |
| Compare_paragraphs.simple | 2 |
| Interpret_phrases.moderate | 2 |
| Analyze_meaning.moderate | 2 |
| Identify_relationship.moderate | 2 |
| Analyze_comparison.moderate | 2 |
| Evaluate_phrases.moderate | 2 |
| Interpret_sentence.moderate | 2 |
| Infer.passage.complex | 2 |
| Infer-meaning.moderate | 2 |
| Differentiate/Compare_endnotes.simple | 2 |
| Evaluate_text/passage.moderate | 2 |

| | |
|---|---|
| Analyze_paragraph | 2 |
| Analyze_comparison.moderate | 2 |
| Interpret_strategy.moderate | 2 |
| Interpret_sentence purpose.moderate | 2 |
| Infer_phrase.complex | 3 |
| Evaluate_passage.complex | 3 |
| Analyze_sentences.complex | 3 |
| Evaluate_paragraph.complex | 3 |
| Summarize-passage_rhetorical strategy.moderate | 3 |
| Evaluate_passage_purpose.complex | 3 |
| Evaluate_sentences.complex | 3 |
| Analyze_theme_passage.complex | 3 |

| **Information Density** | **Difficulty/Complexity** |
|---|---|
| Select.simple | 1 |
| Describe.simple | 1 |
| Conclude.simple | 1 |
| Compare.simple | 1 |
| Critique.simple | 1 |
| Locate.simple | 1 |
| Meaning.simple | 1 |
| Similarity.simple | 1 |
| Locate.simple | 1 |
| Explain.simple | 1 |
| Word_meaning.simple | 1 |
| Clarify_meaning.simple | 1 |
| Apply.simple | 1 |
| Locate.simple | 1 |
| Identify.simple | 1 |
| Integrate_combine.simple | 1 |
| Referencing.simple | 1 |
| Endnoting.simple | 1 |
| Difference.simple | 1 |
| Meaning.simple | 1 |
| Interpret.simple | 1 |
| Effect_repetition.simple | 1 |
| Analogy.simple | 1 |
| Argument.simple | 1 |
| Relate_associate.moderate | 2 |

| | |
|---|---|
| Analyze_moderate | 2 |
| Compare.moderate | 2 |
| Suggest.moderate | 2 |
| Argument.moderate | 2 |
| Storytelling_values.moderate | 2 |
| Description.moderate | 2 |
| Explain.moderate | 2 |
| Inference.moderate | 2 |
| Sentences.moderate | 2 |
| Illustrate.moderate | 2 |
| Assess.moderate | 2 |
| Conclude.moderate | 2 |
| Purpose.moderate | 2 |
| Summarize.complex | 3 |
| Purpose.complex | 3 |
| Summarize.complex | 3 |
| Description.complex | 3 |
| Evaluate.complex | 3 |

| **Contexts** | **Difficulty/Complexity** |
|---|---|
| Context.simple. | 1 |
| Context.moderate | 2 |
| Context.complex | 3 |

Table 21

TASK MODEL CODING SCHEMA AND SCORING INDICES – CALCULUS 2012-2013

| Required Task Actions | Difficulty/Complexity |
|---|---|
| Apply_rule.simple | 1 |
| Calculate_derivative.simple | 1 |
| Solve_operation.simple | 1 |
| Recall_formula.simple | 1 |
| Interpret_operations.simple | 1 |
| Calculate_area.simple | 1 |
| Apply_formula.simple | 1 |
| Calculate_integral.simple | 1 |
| Sketch_graph.simple | 1 |
| Rewrite_operation.simple | 1 |
| Rewrite_formula.simple | 1 |
| Simplify_equation.simple | 1 |
| Interpret_graph.easy | 1 |
| Substitute_function.moderate | 2 |
| Simplify_equation.moderate | 2 |
| Apply_formula.moderate | 2 |
| Interpret_data.moderate | 2 |
| Apply_ratiotest.moderate | 2 |
| Apply_comparisontest.moderate | 2 |
| Calculate_function.moderate | 2 |
| Interpret_function | 2 |
| Apply_FTOC.Moderate | 2 |
| Identify_series.moderate | 2 |
| Manipulate_series.moderate | 2 |
| Calculate_area.moderate | 2 |
| Apply_rules.moderate | 2 |
| Simplify_algebra.moderate | 2 |
| Solve_equation.moderate | 2 |
| Identify_rate.moderate | 2 |
| Determine_change.moderate | 2 |
| Use_substitution.Moderate | 2 |
| Compare_rates.moderate | 2 |
| Construct_polynomial.moderate | 2 |

| | |
|---|---|
| Interpret_graph.moderate | 2 |
| Analyze_data.Moderate | 2 |
| Interpret_relationships.moderate | 2 |
| Categproze_logistics | 3 |
| Interpret_asymtote.hard | 3 |
| Evaluate_function.hard | 3 |

| **Information Density** | **Difficulty/Complexity** |
|---|---|
| Simplify_equations.simple | 1 |
| Calculate.simple | 1 |
| Use_rules.simple | 1 |
| Recall_rules | 1 |
| Values.simple | 1 |
| Ratios_simple | 1 |
| Functional_relationships.moderate | 2 |
| Calculate.moderate | 2 |
| Exponents.moderate | 2 |
| Geometric_sequence | 2 |
| Information_interpret | 2 |
| Simplify.complex | 3 |
| Convergence tests.complex | 3 |
| Calculate.complex | 3 |
| Functions.complex | 3 |
| Values.complex | 3 |

| **Contexts** | **Difficulty/Complexity** |
|---|---|
| Context.simple. | 1 |
| Context.moderate | 2 |
| Context.complex | 3 |

# APPENDIX D

## COGNITIVE COMPLEXITY SCORES

Table 22

Scoring of the Calculus items for Cognitive Task Complexity, Information Density and Context Complexity

| Items | Average Cognitive Task Complexity | Count of Actions | Information Density | Context Complexity | Complexity Index | Complexity Score |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 3 | 2 | 2 | 6 | 24 |
| 3 | 2 | 2 | 2 | 1 | 4 | 8 |
| 4 | 2 | 2 | 1 | 1 | 4 | 4 |
| 5 | 2 | 2 | 3 | 3 | 4 | 36 |
| 6 | 1 | 2 | 1 | 1 | 2 | 2 |
| 7 | 2 | 3 | 3 | 1 | 6 | 18 |
| 8 | 2 | 3 | 2 | 3 | 6 | 36 |
| 9 | 2 | 3 | 2 | 1 | 6 | 12 |
| 10 | 1 | 2 | 1 | 1 | 2 | 2 |
| 11 | 1 | 2 | 2 | 1 | 2 | 4 |
| 12 | 2 | 3 | 3 | 2 | 6 | 36 |
| 13 | 3 | 2 | 2 | 1 | 6 | 12 |
| 14 | 3 | 2 | 2 | 1 | 6 | 12 |
| 15 | 2 | 2 | 3 | 1 | 4 | 12 |
| 16 | 2 | 2 | 2 | 3 | 4 | 24 |
| 17 | 2 | 2 | 1 | 1 | 4 | 4 |
| 18 | 2 | 3 | 3 | 2 | 6 | 36 |
| 19 | 2 | 4 | 2 | 2 | 8 | 16 |
| 20 | 2 | 4 | 3 | 1 | 8 | 24 |
| 21 | 2 | 1 | 2 | 1 | 2 | 4 |
| 22 | 2 | 3 | 2 | 1 | 6 | 12 |
| 23 | 1 | 1 | 2 | 2 | 1 | 4 |
| 24 | 2 | 4 | 3 | 2 | 8 | 48 |
| 25 | 2 | 3 | 3 | 1 | 6 | 18 |
| 26 | 2 | 4 | 3 | 1 | 8 | 24 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 27 | 2 | 2 | 2 | 1 | 4 | 8 |
| 28 | 2 | 2 | 2 | 2 | 4 | 16 |
| 29 | 1 | 2 | 2 | 1 | 2 | 4 |
| 30 | 2 | 3 | 2 | 1 | 6 | 12 |
| 31 | 2 | 3 | 2 | 1 | 6 | 12 |
| 32 | 2 | 2 | 1 | 2 | 4 | 8 |
| 33 | 2 | 2 | 2 | 2 | 4 | 16 |
| 34 | 2 | 3 | 2 | 3 | 6 | 36 |
| 35 | 1 | 2 | 1 | 1 | 2 | 2 |
| 36 | 1 | 1 | 2 | 1 | 1 | 8 |
| 37 | 2 | 2 | 2 | 1 | 4 | 8 |
| 38 | 2 | 3 | 2 | 3 | 6 | 36 |
| 39 | 2 | 2 | 2 | 2 | 4 | 16 |
| 40 | 1 | 2 | 2 | 2 | 2 | 8 |
| 41 | 2 | 2 | 2 | 1 | 4 | 8 |
| 42 | 2 | 2 | 1 | 1 | 4 | 4 |
| 43 | 2 | 2 | 2 | 1 | 4 | 8 |
| 44 | 2 | 2 | 2 | 1 | 4 | 8 |
| 45 | 2 | 3 | 3 | 2 | 6 | 36 |

Table 23

Scoring of the English items for Cognitive Task Complexity, Information Density and Context Complexity

| Items | Average Cognitive Task Complexity | Count of Actions | Information Density | Context Complexity | Complexity Index | Complexity Score |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 2 | 1 | 1 | 2 |
| 3 | 2 | 2 | 2 | 2 | 1 | 16 |
| 4 | 2 | 1 | 1 | 1 | 4 | 2 |
| 5 | 1 | 2 | 2 | 2 | 2 | 8 |
| 6 | 2 | 2 | 2 | 1 | 1 | 12 |
| 7 | 1 | 1 | 1 | 1 | 1 | 1 |
| 8 | 2 | 2 | 2 | 2 | 2 | 16 |
| 9 | 3 | 2 | 2 | 2 | 6 | 24 |
| 10 | 1 | 1 | 2 | 2 | 1 | 4 |
| 11 | 2 | 2 | 2 | 2 | 4 | 16 |
| 12 | 1 | 2 | 1 | 1 | 2 | 2 |
| 13 | 1 | 2 | 2 | 2 | 2 | 8 |
| 14 | 1 | 1 | 2 | 2 | 1 | 4 |
| 15 | 1 | 1 | 1 | 2 | 1 | 2 |
| 16 | 2 | 2 | 2 | 2 | 4 | 16 |
| 17 | 1 | 1 | 1 | 2 | 1 | 2 |
| 18 | 1 | 2 | 1 | 2 | 1 | 4 |
| 19 | 1 | 1 | 1 | 2 | 1 | 2 |
| 20 | 2 | 1 | 2 | 3 | 2 | 12 |
| 21 | 1 | 1 | 2 | 2 | 1 | 4 |
| 22 | 2 | 2 | 2 | 2 | 4 | 16 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 23 | 2 | 1 | 3 | 2 | 1 | 12 |
| 24 | 2 | 1 | 2 | 2 | 2 | 16 |
| 25 | 2 | 2 | 1 | 2 | 4 | 8 |
| 26 | 1 | 1 | 1 | 1 | 1 | 2 |
| 27 | 1 | 1 | 2 | 1 | 1 | 2 |
| 28 | 1 | 2 | 2 | 2 | 2 | 8 |
| 29 | 1 | 1 | 2 | 2 | 2 | 4 |
| 30 | 2 | 2 | 1 | 2 | 4 | 8 |
| 31 | 1 | 2 | 3 | 2 | 1 | 12 |
| 32 | 1 | 2 | 2 | 2 | 1 | 8 |
| 33 | 1 | 2 | 2 | 2 | 1 | 8 |
| 34 | 1 | 2 | 2 | 2 | 1 | 8 |
| 35 | 1 | 1 | 1 | 1 | 1 | 1 |
| 36 | 1 | 2 | 2 | 2 | 4 | 8 |
| 37 | 1 | 2 | 2 | 2 | 4 | 8 |
| 38 | 1 | 2 | 2 | 2 | 2 | 12 |
| 39 | 1 | 2 | 2 | 2 | 1 | 12 |
| 40 | 2 | 2 | 2 | 2 | 4 | 16 |
| 41 | 2 | 1 | 1 | 2 | 2 | 4 |
| 42 | 1 | 1 | 2 | 2 | 2 | 16 |
| 43 | 3 | 2 | 2 | 2 | 4 | 24 |
| 44 | 3 | 2 | 3 | 3 | 6 | 36 |
| 45 | 3 | 2 | 2 | 3 | 6 | 36 |
| 46 | 1 | 1 | 2 | 2 | 1 | 4 |
| 47 | 2 | 1 | 3 | 2 | 2 | 12 |
| 48 | 2 | 2 | 2 | 2 | 4 | 16 |
| 49 | 3 | 2 | 2 | 2 | 6 | 24 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 50 | 3 | 2 | 2 | 3 | 9 | 36 |
| 51 | 2 | 2 | 3 | 2 | 4 | 24 |
| 52 | 2 | 2 | 2 | 2 | 4 | 16 |
| 53 | 2 | 2 | 3 | 3 | 4 | 36 |
| 54 | 2 | 2 | 3 | 2 | 4 | 12 |

# APPENDIX E

# RUBRICS

| | **Levels of Mastery for English Language 2012-2013** | | |
|---|---|---|---|
| **Task Description** | **Advanced (3)** | **Average (2)** | **Below Average (1)** |
| **Vocabulary** | Excellent gasp and knowledge of a wide range of vocabulary words and terms related to the passage to respond to questions.<br>Very good vocabulary | Vocabulary appropriate to respond to most questions.<br>Appropriate use/knowledge of vocabulary and terms. | Limited range of vocabulary works to answer questions.<br>Limited word usage and lower levels of understanding of what is required to solve the task. |
| **Semantics** | Fully Comprehends the meaning of words, sentences and phrases as used in the passages.<br>The student has a good overall understanding of what is read. | Adequately comprehends the meaning of words and sentences and phrases as used in the passages.<br>The student understands most of what is read. | Limited knowledge of the meaning of words and phrases. Often guesses meaning of words and or phrases.<br>The student often fails to understand what is being read |
| **Context** | Fully comprehend the context in which words in the passage or story are used.<br>Most phrases are comprehensible<br>Students clearly understand the concepts, constructs and the context in which words are used | Adequately comprehends the context of the passage or story.<br>Some phrases are comprehensible | Limited comprehension of the context of the story.<br>Many phrases are incomprehensible.<br>There are frequent errors<br>Low levels of understanding of the complexity of the information required to solve the task.<br>Low context Complexity |
| **Task Description** | **Advanced (3)** | **Average (2)** | **Below Average (1)** |
| **Skills** | Can manipulate the advanced level of skills<br>Students maximize the uses of appropriate strategies, cues and aids to correctly solve the task.<br>High complexity | Use of few context clues to assist in understanding the complexity of the task.<br>Moderate complexity | The students makes little use of the context clues and aids to help solve the task.<br>Students possess limited or low levels of skills needed to solve the task. |

| | | | |
|---|---|---|---|
| **Task description** | High complexity and Dense verbal load.<br><br>Implicit association between stem and options. | Moderate level of complexity and information density of material<br><br>Moderately implicit/explicit relations among stem and options (Knowledge Objects). | Low levels of complexity<br>Explicit association between stem and options |

| | Levels of Mastery Calculus 2012-1013 | | |
|---|---|---|---|
| **Task Description** | **Advanced** | **Average** | **Below Average** |
| **Reasoning** | Demonstrates thorough understanding of the relationships between functions and key features that must be used in logically reasoning out solution to the problem. Accurately reasons and follows all steps relevant in solving the problem | Demonstrates clear reasoning and understanding of the relationship between functions and key features required to solve the problem. Follows most steps logically to correctly solve the problem. | Shows limited understanding of what is required in the problem. Fails to follow all steps in solving the problem. Unable to reason out the problem logically. |
| **Knowledge & Under-standing** | Shows excellent understanding of the problem and correctly solves them. Uses correct method and mathematical techniques to solve the problem | Shows adequate understanding of the problem. Uses correct method and mathematical technique to solve problem with minor errors | Shows limited or little understanding of the problem. Has little success in solving the problem |
| **Mathemati-cal Concepts** | Demonstrate a thorough analysis and understanding of three or more Calculus concepts or functions use to solve the problem. apply the concepts Students will effectively use concepts of function, limit, continuity, derivative and integral. | Demonstrate a good understanding and analysis of two calculus functions or concepts to solve the problem Some areas of the response are incorrect. | Minimal understanding and use of one calculus functions or concepts required to correctly solve the problem |
| **Mathemati-cal Terminology and Notation** | Student understands and uses correct symbols and terminology to solve the problem | Accurate or correct use of some terminology and notation required to solve the problem. | Limited use and understanding of appropriate terminology and notation to solve problem. Frequent errors |

| | | | |
|---|---|---|---|
| **Skills** | Computational and Calculations where appropriate.<br>Thorough understanding of how to sketch graphs of polynomials, rational and exponential functions<br>Algebraically and graphically demonstrate thorough understanding of key theoretical and geometrical calculus concepts | Use of some calculations and computations appropriately<br>Adequate understanding of how to sketch graphs of polynomials, rational and exponential functions<br>Algebraically and graphically demonstrate clear understanding of key theoretical and geometrical calculus concepts | Little or no calculations or computation. There are inaccuracies.<br>Sketches graphs of polynomial, rational and exponential functions with errors and omissions.<br>Algebraically and graphically demonstrate limited understanding of key theoretical and geometrical calculus concepts |
| **Application** | Accurate use of the correct formulas/equations to answer the question.<br>Correctly determine which and when to apply calculus function and concepts to solve problem<br>Students will apply methods of calculus to optimization, graphing and approximation. | Appropriately use correct formulas/equations to answer the question most of the time<br>Frequently correctly determine when and where to apply Calculus functions and concepts to solve problem | Little use of the correct formulas/equations to answer the question<br>Seldom correctly determine and apply calculus functions and concepts to solve problem |

# APPENDIX F

## WINSTEPS CODE

```
&INST;
TITLE= 'language 2012'
NI=54;
ITEM1= 2;
DATA=lang2012.txt;
IAFILE=TM2012.txt;
CODES= 01;
LCONV=0.0001;
UDECIIM=5
PTBIS=Y
PVALUE=Y
TOTALSCORE=Y
IFILE=TMlang2012.itm;
PFILE=TMlang2012.pfl
&END ;
```