

Technology Enabled Assessments:
An Investigation of Scoring Models for Scaffolded Tasks

By

Copyright 2012
Brooke L. Nash

Submitted to the graduate degree program in the
Department of Psychology and Research in Education
and the Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy.

Chairperson: William P. Skorupski

Co-chairperson: Vicki Peyton

Neal Kingston

Bruce Frey

Sean Smith

Date defended: April 9th, 2012

The Dissertation Committee for Brooke L. Nash certifies
that this is the approved version of the following dissertation:

Technology Enabled Assessments:
An Investigation of Scoring Models for Scaffolded Tasks

Chairperson: William P. Skorupksi

Co-chairperson: Vicki Peyton

Date Approved: _____

Abstract

While significant progress has been made in recent years on technology enabled assessments (TEAs), including assessment systems that incorporate scaffolding into the assessment process, there is a dearth of research regarding psychometric scoring models that can be used to fully capture students' knowledge, skills and abilities as measured by TEAs. This investigation provides a comparison of seven scoring models applied to an operational assessment system that incorporates scaffolding into the assessment process and evaluates student ability estimates derived from those models from a validity perspective.

A sequential procedure for fitting and evaluating increasingly complex models was conducted. Specifically, a baseline model that did not account for any scaffolding features in the assessment system was established and compared to three additional models that each accounted for scaffolding features using a dichotomous, a polytomous and a testlet model approach. Models were compared and evaluated against several criteria including model convergence, the amount of information each model provided and the statistical relationships between scaled scores and a criterion measure of student ability.

Based on these criteria, the dichotomous model that accounted for all of the scaffold items but ignored local dependence was determined to be the optimal scoring model for the assessment system used in this study. However, if the violation against the local independence assumption is deemed unacceptable, it was also concluded that the polytomous model for scoring these assessments is a worthwhile and viable alternative. In any case, the scoring models that accounted for the scaffolding features in the assessment system were determined to be better overall models than the baseline model that did not account for these

features. It was also determined that the testlet model approach was not a practical or useful scoring option for this assessment system.

Given the purpose of the assessment system used in this study, which is a formative tool that also provides instructional opportunities to students during the assessment process, the advantages of applying any of these scoring models from a measurement perspective may not justify the practical disadvantages. For instance, a basic percent correct score may be completely dependent on the specific items that a student took but it is relatively simple to understand and compute. On the other hand, scaled scores from these scoring models are independent of the items from which they were calibrated from, but ability estimates are more complex to understand and derive. As the assessment system used in this study is a low stakes environment that is mostly geared towards learning, the benefits of the scoring models presented in this study need to be weighed against the practical constraints within an operational context with respect to time, cost and resources.

Acknowledgements

This dissertation is the result of the contributions of several individuals. First I would like to thank the members of my committee, Dr. Sean Smith, Dr. Bruce Frey, Dr. Neal Kingston, Dr. Vicki Peyton and Dr. William Skorupski. Dr. Kingston has provided me continued support throughout my graduate education through advice, opportunity, and knowledge. His energy and passion for the field of measurement has and continues to be an inspiration to me. I would also like to especially thank the co-chairs of this dissertation; Dr. Peyton has provided me endless encouragement, friendship and guidance since I first started the program as a Masters student. Her continuous time and support was fundamental to my success and desire to pursue greater goals. Finally, I would like to thank Dr. Skorupski whose feedback, suggestions, and intellect undoubtedly inspired this project in many ways. Even in times when error messages were occurring more frequently than not, he provided much needed optimism and encouragement. I offer my sincere appreciation for the patience and mentoring he offered throughout this process.

This dissertation would not have been possible without the contribution of the Assistments data. For that I would like to thank Dr. Neil Heffernan who graciously provided me with the data used in this study and for helping me understand the Assistments system.

I could not have begun nor completed this journey without the love and support of my family. To my parents who have always believed in me and supported me in every decision I have made. I would also like to thank my grandmother for her strength and generosity which has always, and will always inspire me. To my brother and sister for making me laugh; that in itself has provided me encouragement even though they may not have realized it! Finally,

all my love and gratitude go to my husband and daughter. Rich has always been there for me throughout it all! His unconditional love and benevolence made this voyage not only possible but also enjoyable. As my unofficial technical support person, he also saved the day when my computer was ready to throw in the towel! To my adorable daughter who makes us laugh everyday! Thank you to you both for taking this journey with me.

Table of Contents

Abstract.....	iii
Acknowledgements.....	v
List of Tables.....	xi
List of Figures.....	xii
List of Appendices.....	xv
Chapter One – Introduction.....	1
Context of Study.....	1
Research Questions.....	3
Significance of Study.....	4
Chapter Two – Literature Review.....	6
Scaffolding.....	6
Features & Characteristics.....	7
Scaffolding Schemas.....	8
Computer-based Scaffolding.....	11
Efficacy of Computer-based Scaffolding.....	13
Features & Characteristics.....	15
Technology Enabled Assessment.....	17
Formative Use of TEAs.....	18
Scaffolded Assessments.....	19
Specific Examples of Scaffolded Assessments.....	20
The Assisments System.....	21
Scoring Scaffolded Assessments.....	31
Purpose of Research.....	32
Item Response Theory & Assumptions.....	33
Dichotomous IRT Models.....	35
The 1PL Model.....	35
The 2PL Model.....	36
The 3PL Model.....	37
Item & Test Information.....	39

Polytomous IRT Models.....	40
Graded Response Model.....	42
Ordinal Response Model.....	42
Item & Test Information.....	43
Bundle Models.....	44
Score-based Approaches.....	46
Item-based Approaches.....	47
Testlet Response Model.....	47
Item & Test Information.....	49
Modeling Assistsments Data.....	49
Model Comparison.....	50
Summary.....	51
Chapter Three – Methods.....	53
Participants.....	53
Assistments Data.....	54
Data Cleaning Procedures.....	54
Missing Data.....	60
Software.....	61
Procedures.....	63
Research Question 1.....	63
The 1PL or 2PL Model?.....	64
Baseline Model.....	69
Comparison Models.....	69
Parameter Estimation.....	70
Overview of Bayesian Inference.....	71
Bayesian Framework in SCORIGHT 3.0.....	72
Specifying Models in SCORIGHT 3.0.....	73
Model Evaluation.....	74
Bayesian Convergence.....	75
Model Fit.....	76
Information.....	78

Research Question 2.....	79
Analyses.....	80
Summary.....	81
Chapter Four – Results.....	82
Research Question 1.....	82
Convergence.....	83
2PL_MainItems Model.....	83
2PL_AllItems Model.....	84
Ordinal Response Model.....	86
Testlet Response Model.....	87
Additional TRM Calibration Procedures.....	94
Summary.....	95
Descriptive Statistics.....	96
2PL_MainItems Model.....	97
2PL_AllItems Model & ORM.....	97
Testlet Response Model.....	102
Summary.....	106
Model Fit.....	108
Information.....	110
Research Question 2.....	113
Relationships between Scoring Models.....	114
Relationships with Criterion.....	115
Summary.....	120
Chapter Five – Discussion.....	121
Research Question 1.....	123
Convergence.....	123
Descriptive Statistics.....	126
Model Fit.....	131
Information.....	132
Research Question 2.....	134
Model Summary & Selection.....	136

Limitations & Future Research.....	140
Conclusions.....	141
References.....	143
Appendices.....	153

List of Tables

Table 1.	Frequency of Number of Items per Bundle	54
Table 2.	Number of Bundles Associated with each Sample Size Category and the Total Number of Bundles if the Category was Removed	55
Table 3.	Frequency and Number of Items by Bundle Size	57
Table 4.	Number and Proportions of Missing Cases for each Main Item	60
Table 5.	Outline of Model Evaluation Procedures	68
Table 6.	Estimation Specifications and PSRFs for the 2PL_MainItems Model	84
Table 7.	Estimation Specifications and PSRFs for each 2PL_AllItems Model	85
Table 8.	Estimation Specifications and PSRFs for each Ordinal Response Model	87
Table 9.	Estimation Specifications and PSRFs for each Testlet Response Model	89
Table 10.	PSRFs for Variances of Gammas for each Testlet Response Model	90
Table 11.	Descriptive Statistics for each Item	99
Table 12.	Summary Statistics for Original Data (not calibrated with IRT)	101
Table 13.	Summary Statistics for the Dichotomous 2PL_MainItems Model	103
Table 14.	Summary Statistics for the Dichotomous 2PL_AllItems Models	103
Table 15.	Summary Statistics for the Polytomous Ordinal Response Models	103
Table 16.	Summary Statistics for the Testlet Response Models	104
Table 17.	Estimated Variances of Gamma (γ) and Standard Errors for each Bundle	105
Table 18.	Deviance Results for each Evaluation Model	109
Table 19.	Correlation Coefficients between Scaled Scores Obtained from each Scoring Model, Percent Correct Scores and State Test Scores	116
Table 20.	Item Fit Statistics for the 1PL and 2PL	153

List of Figures

Figure 1.	Example Assistentment item on congruent triangles	24
Figure 2.	First scaffold question for example congruent triangles Assistentments item	25
Figure 3.	Second scaffold question for example congruent triangles Assistentments item	26
Figure 4.	Third scaffold question for example congruent triangles Assistentments item	27
Figure 5.	Fourth scaffold question for example congruent triangles Assistentments item	28
Figure 6.	Assistentments item example flowchart	30
Figure 7.	ICCs of a dichotomously scored item based on the 1PL	38
Figure 8.	ICC of a dichotomously scored item based on the 3PL	38
Figure 9.	Frequency distribution of the number of main items administered to students	59
Figure 10.	Standardized residuals for each of the 32 main items estimated with the 1PL model	66
Figure 11.	Standardized residuals for each of the 32 main items estimated with the 2PL model	67
Figure 12.	Frequencies of standardized residuals for all 32 items in the 1PL and 2PL models	67
Figure 13.	Model comparison flowchart	75
Figure 14.	Time-series plot for the variance of gamma for Bundle 1 based on 100,000 iterations	92
Figure 15.	Time-series plot for the variance of gamma for Bundle 26 from the TRM (without covariates) model based on 100,000 iterations	92
Figure 16.	Time-series plot for the variance of gamma for Bundle 26 from the TRM + covs model based on 100,000 iterations	93
Figure 17.	A comparison of item discrimination values for each model that did not incorporate covariates in the estimation process	107

Figure 18.	A comparison of item difficulty values for each model that did not incorporate covariates in the estimation process	107
Figure 19.	Total test information for each scoring model	112
Figure 20.	Total bundle information for 2PL_AllItems scoring model (without covariates) which ignore local dependence	112
Figure 21.	Total test information for ORM scoring model (without covariates) which account for local dependence	113
Figure 22.	Scatterplot of percent correct scores on main items only and state test scores	118
Figure 23.	Scatterplot of percent correct scores on all Assisments items and state test scores	118
Figure 24.	Scatterplot of scaled scores from the 2PL_MainItems model and state test scores	119
Figure 25.	Scatterplot of scaled scores from the 2PL_AllItems model and state test scores	119
Figure 26.	Scatterplot of scaled scores from the ORM and state test scores	120
Figure 27.	A rank ordered comparison of models by each evaluation criterion	139
Figure 28.	Information for Bundle 1.	162
Figure 29.	Information for Bundle 2.	162
Figure 30.	Information for Bundle 3.	163
Figure 31.	Information for Bundle 4.	163
Figure 32.	Information for Bundle 5.	164
Figure 33.	Information for Bundle 6.	164
Figure 34.	Information for Bundle 7.	165
Figure 35.	Information for Bundle 8.	165
Figure 36.	Information for Bundle 9.	166
Figure 37.	Information for Bundle 10.	166

Figure 38.	Information for Bundle 11.	167
Figure 39.	Information for Bundle 12.	167
Figure 40.	Information for Bundle 13.	168
Figure 41.	Information for Bundle 14.	168
Figure 42.	Information for Bundle 15.	169
Figure 43.	Information for Bundle 16.	169
Figure 44.	Information for Bundle 17.	170
Figure 45.	Information for Bundle 18.	170
Figure 46.	Information for Bundle 19.	171
Figure 47.	Information for Bundle 20.	171
Figure 48.	Information for Bundle 21.	172
Figure 49.	Information for Bundle 22.	172
Figure 50.	Information for Bundle 23.	173
Figure 51.	Information for Bundle 24.	173
Figure 52.	Information for Bundle 25.	174
Figure 53.	Information for Bundle 26.	174
Figure 54.	Information for Bundle 27.	175
Figure 55.	Information for Bundle 28.	175
Figure 56.	Information for Bundle 29.	176
Figure 57.	Information for Bundle 30.	176
Figure 58.	Information for Bundle 31.	177
Figure 59.	Information for Bundle 32.	177

List of Appendices

Appendix A.	Item Fit Statistics for 1PL and 2PL Preliminary Analyses	153
Appendix B.	DIC Program for Dichotomous Models	155
Appendix C.	DIC Program for Polytomous Models	158
Appendix D.	Information for each Bundle	162

Chapter One – Introduction

Context of Study

In 2001, congress passed the No Child Left Behind (NCLB) Act which requires all students within each state to be tested on their specific state curriculum standards. These statewide mandated tests are intended to measure student proficiency with respect to state standards which is in turn, a reflection of school effectiveness. As a result of this legislation, the need for efficient, precise, and beneficial assessment systems has grown. In other words, educators and policymakers need assessment systems that can provide the biggest bang for their buck. Furthermore, educators and researchers have also claimed that assessment procedures need to be altered in order to not only provide encouragement and motivation but also to ensure that all students will be capable of succeeding (Arter, 2003; Stiggins, 2005). In other words, there is now consensus within the field that assessments need to support and encourage learning rather than just measure it.

In response to this need to support student learning, assessments that provide teachers and students with formative data and feedback that can be used to guide teaching and learning activities have grown in popularity. Formative assessment has been defined as a process used by teachers and students during instruction that is intended to provide feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes (CCSSO, 2006). Since Black & Wiliam's (1998) detailed synthesis of the literature on formative assessment which outlined the positive learning effects of formative procedures, educators have increasingly integrated tools and assessments to be used formatively into the classroom.

Advances in technology have only enhanced researchers' and policymakers' interest in advancing the state of assessment tools including those to be used for formative purposes. States are increasingly incorporating technology into their testing programs in order to provide greater efficiency, to better model effective instructional methods, and to more accurately measure student proficiency (Almond et al, 2010, Bechard et al, 2010). These technological innovations have placed the possibility of integrating instruction with assessment at the forefront of assessment development (Koedinger, McLaughlin, & Heffernan, 2010). Test developers are beginning to experiment with innovative item types that require students to interact and manipulate information on a computer screen while demonstrating deeper levels of knowledge and understanding (Almond et al, 2010). There are several types of interactive assessments and assessment strategies made possible through technology that have the potential to provide feedback to students and teachers during the assessment process (Bechard et al, 2010). One such strategy is to incorporate instruction into assessment using scaffolds.

Instructionally, scaffolding can be used to help students understand content or concepts by providing appropriate supports geared towards their current learning and/or cognitive capabilities (Almond et al, 2010). Scaffolding, if applied appropriately to an assessment environment, allows for more accurate measurement of students' knowledge and skills by providing supports to students that allow them to respond to a task at a level that fits with the students' individual needs and abilities. That is, assessment tasks can be built to provide students with the opportunity and choice to engage in construct-relevant supports when they encounter an item (Almond et al, 2010).

This study utilized data from an existing assessment system, known as the Assistments¹, which currently incorporates scaffolding into the assessment process. This assessment system presents students with published state assessment test items. If students provide a correct response they are given a new one, otherwise they are provided with a small “tutoring” session. The tutoring session breaks down the original item into more manageable skill-based tasks and provides hints to guide the student if he or she has difficulty. By doing this, the system is able to differentiate students who get the original item wrong at first but need different levels of tutoring to get the problem correct eventually (Feng, Heffernan, & Koedinger, 2009).

Research Questions

As the field of educational assessment continues to evolve in conjunction with advancements in technology and assessment systems grow in complexity and value, a need exists for developing ways to score these assessments. This study contributes towards addressing that need by investigating different scoring models that can be applied to scaffolded item types which take into account whether and how a scaffold is used in an item response. Such a scoring model has the potential to provide an efficient measure of student ability which may ultimately be used to gauge student progress towards end of the year assessments.

The purpose of this research is to help advance the development and use of assessment systems that utilize technological innovations and specifically those that incorporate scaffolding into the assessment process. The goal is to make recommendations about optimal

¹ Data provided by *Assistments*. © Worcester Polytechnic Institute. www.assistments.org

scoring models that can be used for scaffolded assessments based on the characteristics of the scaffolds utilized in the example assessment system. Specifically, the research questions in this study are as follows:

- 1) What type of model is the optimal scoring model for the scaffolded data provided by the Assistments system?
 - a. Which scoring model produces the best model fit for the system?
 - b. Which scoring model produces the most precise measures of student ability?
 - c. Do the benefits associated with the better fitting model outweigh any practical concerns due to model complexity?
- 2) Is there a relationship between student ability estimates derived from the scoring models and a criterion measure of student achievement?
 - a. Do any of the scoring models provide student ability estimates that predict a criterion measure of student ability better than a simple percent correct score?
 - b. Do the models that account for the scaffolding features have a stronger relationship with a criterion measure of student achievement than the models that do not account for those features?
 - c. Do the models that account for the local dependence have a stronger relationship with a criterion measure of student achievement than the models that do not account for the dependence?

Significance of Study

In the current age of accountability, there is an increasing need for teachers to assimilate instruction with assessment (Koedinger, McLaughlin, & Heffernan, 2010).

Teachers need frequent and accurate measurements of their students' knowledge, skills and

abilities without consuming valuable instruction time. The use of technology enabled assessments has the potential to address these needs (Bechard et al, 2010). Integrating scaffolds into assessment tasks which emulate what teachers do in the classrooms provides students with individualized instruction while measuring what they know and don't know. In order to provide teachers with the most accurate measures of their students' abilities, one must investigate how these types of scaffolded tasks should be scored. This investigation contributes towards that goal by providing a comparison of scoring models for an operational scaffolded assessment system and evaluating student ability estimates derived from those models from a validity perspective.

Chapter Two - Literature Review

Scaffolding

The first use of the term “scaffolding” in psychological research was proposed by Wood, Bruner & Ross (1976) to describe the process by which a child or novice is able to solve a problem or achieve a goal that would otherwise be beyond the child’s or novice’s ability. This process was described as “...the adult ‘controlling’ those elements of the task that are initially beyond the learner’s capacity, thus permitting him to concentrate upon and complete only those elements that are within his range of competence” (p. 90). While Wood et al. (1976) did not explicitly make the connection in their original research, many have since connected the concept of scaffolding to Lev Vygotsky’s 1930’s concept of the zone of proximal development (Cazden, 1979; Bruner, 1986, Holton & Clarke, 2006; McNiell, Lizotte, Krajcik, & Marx, 2006; Sharpe, 2006; Shepard, 2005; Wood, 1988). Vygotsky (1978) described the zone of proximal development (ZPD) as “the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers. The zone represents the potential for a child’s development when aided by others” (p. 86). From this, it is clear that the concept of scaffolding was implicit within Vygotsky’s envision.

More recently, scaffolding has become commonly used in educational contexts to describe “the precise help that enables a learner to achieve a specific goal that would not be possible without some kind of support” (Sharpe, 2006, p. 212). Thus, scaffolding in this context is the amount of assistance that a learner needs to achieve a goal within the learner’s ZPD. In other words, if a scaffold is to enhance student learning, it needs to reside within a

students' current ZPD (McNeil et al., 2006). A scaffold that provides too much support may result in a less challenging task and decreased motivation while a scaffold that does not provide enough support may result in anxiety and frustration for the learner (McNeil et al., 2006). Thus, scaffolding is the means to which learners can reach their potential development as hypothesized by their zone of proximal development.

Features & Characteristics. While the metaphor used to describe how scaffolding is applied to the context of learning and development varies across researchers (Stone, 1998a; Stone 1998b), there are several key theoretical features and characteristics that are common across successful scaffolding systems. For instance, Puntambekar & Hubscher (2005) delineate four central features that are necessary for successful scaffolding: role of the expert, shared understanding of the goal, ongoing diagnosis, and fading. The most critical of these is the role of the expert (Puntambekar & Hubscher, 2005). The traditional concept of scaffolding assumed that a single, more knowledgeable person, such as a parent or a teacher, helped an individual learner by providing him or her with the appropriate amount of help he or she needed to move forward (Wood, et al, 1976). More specifically, Wood et al. (1976) suggest that there are six key functions or responsibilities of the expert in scaffolded instruction: (1) recruitment, or engaging the learner in a meaningful activity; (2) reduction in degrees of freedom, or simplifying the activity into manageable components; (3) direction maintenance, or keeping the learner on-task; (4) marking critical features, or emphasizing the main elements of the task; (5) frustration control, or attending to the situation so as to reduce the frustration level without creating a dependency issue; and (6) demonstration, or providing a model of the correct method for the learner (p. 98).

Furthermore, scaffolding should incorporate a common understanding of the goal of the task between the expert and the learner (Puntambekar & Hubscher, 2005). Rogoff (1990) referred to this as “intersubjectivity” or shared understanding such that both the expert and the learner take ownership of the task. Another aspect that is vital to successful scaffolding is the role of ongoing diagnosis which provides the expert with information about the learner’s current state of understanding which serves as the basis for a calibrated system of support (Puntambekar & Hubscher, 2005). This leads to the final element of successful scaffolding as delineated by Puntambekar & Hubscher (2005) which is the fading process of the support system. That is, a transfer of accountability from the expert to the learner needs to occur such that the scaffolding can be removed and the learner is capable of independent activity (Puntambekar & Hubscher, 2005). Vygotsky referred to this cognitive process that occurs first on an interpsychological plane and then moves on to an intrapsychological plane, internalization (Vygotsky, 1978). Puntambekar & Kolodner (2005) delineated similar features but also emphasized a dialogic and interactive component between the expert and learner such that interactions can provide ongoing assessment of the learner but also allow the learner to be actively engaged in the scaffolding process.

Scaffolding Schemas. Researchers have defined and described various schemas to try and capture the many facets of scaffolding (Azevedo, 2004; Cagiltay, 2006; Hannafin, 1999; Holton, 2006; Pea, 2004). Drawing on a modified framework originally proposed by Pea (2004), the concept of scaffolding can be broken down into several main components which are described as the *who*, *what*, and *how* of scaffolding. In its traditional form, the *who* of scaffolding would have been efficiently explicated as the tutor, adult expert or more competent peer (Bruner, 1985). More recently however, the *who* of scaffolding has been

extended to incorporate internal scaffolders as well as technological scaffolders (e.g., Cagiltay, 2006; Holton, 2006). Holton (2006) proposes that there are three types of scaffolders: expert scaffolding, reciprocal scaffolding, and self-scaffolding. Expert scaffolding involves a scaffolder with a primary responsibility to help others learn while reciprocal scaffolding involves collaboration with another on a common task such that differing ability levels interact to provide a form of scaffolding (Holton, 2006). Both of these definitions are at least partially inherent in the traditional description of scaffolding proposed by Wood et al (1976). However, self-scaffolding is a relatively modern addition to the notion of scaffolding and involves situations in which new content is being learned and an individual is able to provide scaffolding for him or herself (Holton, 2006). For example, a learner who knows that he or she is primarily a visual learner may draw him or herself a diagram to understand the organization of a new concept. Furthermore, as information technologies become integrated into learning environments scaffolding is now being provided by means of computer software (Cagiltay, 2006; Quintana et al, 2004; Reiser et al, 2001). As discussed later in this review, software-realized scaffolding attempts to embed the concept of scaffolding into a computer-based environment.

The *what* and *how* of scaffolding are related ideas and described as the functions and mechanisms of scaffolding by Hannafin (1999). That is, the functions emphasize the purpose of the scaffold while the mechanisms emphasize the methods through which the scaffolding is provided (Hannafin, 1999). The *what* and *how* of scaffolding are typically divided into four main categories: conceptual, metacognitive, procedural, and strategic (Azevedo, 2004; Cagiltay, 2006; Hannafin, 1999). Conceptual scaffolding helps learners rationalize through complex or commonly misunderstood concepts; it guides learners about what to consider as

they reason through a task (Hannafin, 1999). Conceptual scaffolding is frequently demonstrated through methods such as providing hints and prompts at appropriate times during the learning process, providing outlines or graphical displays of content, and highlighting key concepts (Cagiltay, 2006; Hannafin, 1999). Metacognitive scaffolding supports the learner with how to think when learning; it guides learners on how to manage their own learning processes (Hannafin, 1999). Methods used to exhibit this type of scaffolding include evaluating progress, modeling cognitive strategies, and suggesting self-regulating strategies and milestones for the learner to consider (Hannafin, 1999). Learners are encouraged to reflect on their own learning processes by answering questions posed by the scaffolder and responding to the scaffolder's critiques (Cagiltay, 2006; Hannafin, 1999). Procedural scaffolding supports the learner with how to utilize resources and tools within a particular learning environment (Cagiltay, 2006; Hannafin, 1999). Hannafin et al point out that this type of scaffolding is "frequently provided to clarify how to return to a desired location, how to flag or bookmark locations or resources for subsequent review, or how to deploy given tools" (1999, p. 133). This type of scaffolding can be operationalized through tutoring on given tools, functions, and features. Finally, strategic scaffolding guides learners with how to analyze or approach a learning task or problem (Hannafin, 1999). In other words, it supports necessary skills for solving a problem such as identifying, evaluating and applying relevant information and knowledge and evaluating alternate problem-solving strategies. Methods through which strategic scaffolding can be achieved include providing start-up questions to be considered by the learner, alerting learners to helpful resources, or providing the learner with worked examples or solution paths of peers or experts (Azevedo, 2004; Cagiltay, 2006; Hannafin, 1999).

While other researchers have delineated the *who*, *what* and *how* of scaffolding through different schematic themes than those described here, proposed concepts appear to be encompassed by the definitions described above in one way or another. For example, Holton (2006) describe the *what* of scaffolding with two main scaffolding domains: conceptual and heuristic. Holton (2006) explain that conceptual scaffolding emphasizes the development of conceptual development or content while heuristic scaffolding emphasizes the development of “heuristics for learning or problem solving that transcend specific content” (2006, p. 134). Clearly both domains described by Holton (2006) can be encompassed by the conceptual and strategic categories in the schema presented above. Other researchers have described the *how* of scaffolding in various ways as well. For instance, Pea (2004) specifically describes this function using two groups of assistance: channeling and focusing, and modeling. Channeling and focusing reduces “the degrees of freedom for the task at hand” and focuses the “attention of the learner by marking relevant task features” (Pea, 2004, p. 432). Modeling, on the other hand, generally models more advanced solutions to a problem (Pea, 2004). While these two types of assistance may embody mechanisms beyond those described above, the notions underlying each can be defined by the schema above. That is, the methods used to elicit conceptual as well as reflective scaffolding are similar to the ideas of channeling and focusing; channeling as a way of breaking down a problem into conceptually easier to understand parts and focusing as a way of guiding the learner to reflect on his or her own attention to the task at hand.

Computer-based Scaffolding

With the advancement of technology as well as increased demands for more ambitious learning environments, the idea of scaffolding has been adopted in research on technological

supports for instruction in more recent years (Hannafin & Land, 1997; Puntambekar & Hubscher, 2005; Quintana et al, 2004, Reiser, 2004). Researchers in the educational technology field posit that software can be used to scaffold students by providing support that enables learners to succeed in complex tasks and extend the range of learning experiences (Davis & Linn, 2000; Guzdial, 1994; Guzdial & Kehoe, 1998; Reiser, 2002). In this sense, scaffolding refers to cases in which the tool changes the task such that the learner can achieve a goal that would otherwise be beyond their own abilities (Reiser, 2004). While many contend that human scaffolding is more beneficial to learners than computerized scaffolding due to the human's ability to detect subtle cues from the learner (Holton & Clarke, 2006), others have recognized that it is not always feasible for experts to provide every learner the one-to-one tutoring that may be needed (Cagiltay, 2006; Puntambekar & Hubscher, 2005; Stone, 1998a; Stone 1998b). Furthermore, group work or peer tutoring can also be problematic for several reasons. First of all, peers working together do not necessarily intentionally calibrate their level support based on a diagnosis of their partner's understanding (Puntambekar & Hubscher, 2005). Secondly, while some peers may be more knowledgeable than others, that does not necessarily translate to effective feedback either due to the lack of confidence in that knowledge or the lack of verbal skills needed to express that knowledge (Puntambekar & Hubscher, 2005).

In any case, while human scaffolding undoubtedly has its advantages over computerized scaffolding, the latter may also provide other benefits that are not apparent in the former such as individual tutoring. As Guzdial (1994) points out, the challenge for educational technology researchers is to provide the same scaffolding an effective teacher provides in the classroom environment but in a software environment. In other words, ideally

the designer of the software is defining and creating scaffolding as the teacher but through the mechanism of the software (Guzdial, 1994). Thus, the goals of computerized scaffolding are the same as traditional scaffolding in that it attempts to facilitate student performance and learning (Guzdial, 1994).

Efficacy of Computer-based Scaffolding. Software designers have argued that instructional software tools can support learners by providing needed structure for difficult tasks in the form of scaffolds (Davis & Linn, 2000; Guzdial, 1994; Reiser, 2002). In general, ways in which software tools provide support to learners to help them solve complex tasks include constraining the task itself, providing organizational structure, making processes and strategies more apparent (Puntambekar & Hubscher, 2005), providing feedback and suggestions during the learning process, and eliciting articulation (Guzdial, 1994). While research on the direct effects of many of these supports on student performance and learning is sparse, the findings that are available are positive. For example, Davis & Linn (2000) studied the effects of the Knowledge Integration Environment (KIE) software which incorporates prompts that require students to provide explanations and to reflect on their work at selected points of the project. Their investigations suggest that prompts, tailored to the specific task at hand, can influence student performance by lessening the cognitive load on students and by reminding them how to accomplish the activity (Davis & Linn, 2000). Chang (2001) compared groups of students that received scaffolding to learn science content versus those that did not receive scaffolding within the context of a computer-based concept mapping system. The scaffolding mechanisms in this study were an incomplete framework of an expert concept map as well as specific hints and feedback that describe student performance in reference to a completed expert concept map. Their findings suggest that the feedback

function may not only reduce the student frustration but also further promote students' positive attitudes and participation in the map construction process (Chang, 2001). Thus, it appears that integrating scaffolding into software tools has the potential to reduce the cognitive complexity of tasks and support correct processes which may help reduce learner frustration and help learners develop a positive perspective towards the learning task.

With respect to the effects of student learning, Guzodial et al. (1998) sought to support students to learn and develop computer programming skills, and their analysis of student actions while creating programs suggested that learners using a scaffolded tool produced better programs with less effort. More recently, Koedinger et al. (2010) used quasi-experimental data from a web-based tool that incorporates scaffolding in the form of decomposing the problem into sub-tasks as well as the availability of hints, to analyze the learning outcomes of students who used the tool versus those who did not. Findings indicated that students who used the tool performed better on the year-end exam than those who did not use it; however, due to the lack of random assignment of students in the study, caution was given with regard to implications that the tool caused the difference in performance. Perhaps more compelling is a review of the literature on Cognitive Tutors which are described as interactive software learning environments that provide various kinds of assistance to students as they learn complex cognitive skills (Koedinger & Alevan, 2007). While the assistance provided in these cognitive tutoring systems are not necessarily referred to as scaffolds, their features closely resemble those of traditional scaffolds (e.g., hints and suggestions for correct solution paths, error feedback messages). In summarizing their review, Koedinger & Alevan (2007) concluded that classrooms that use Cognitive Tutors show significant learning advantages over classrooms that do not involve computer tutors. More importantly, they

found that the research provides “suggestive evidence” for positive learning gains for the use of on-demand hints as well as for the use of error feedback messages that are intended to make learning more explicit (Koedinger & Alevan, 2007).

While these investigations suggest that computer-based scaffolding can have a positive impact on student learning, more research is needed to warrant any conclusions. However, it is not surprising that there is a lack of experimental data in this area due to the complex nature of the classroom environment and the role that technology plays in that environment. In fact, as Koedinger & Alevan point out, technology should not be thought of “as a panacea to the achievement problems in education” (2007, p. 491) and that technology alone does not increase learning outcomes. That is, there are many contextual variables that can mediate the effects of a technological tool including implementation procedures, teacher expertise, support and training, and student readiness to use the tool.

Features & Characteristics. Features and characteristics of computer-based scaffolds are theoretically the same as those in traditional scaffolding. In other words, the four scaffolding features described previously can potentially be applied to scaffolds provided in software tools (Puntambekar & Hubscher, 2005). For example, the role of the expert in a computer-based environment (as mentioned previously) ideally is based on what the teacher would provide the learner in the classroom but transmitted through the software tool (Guzdial, 1994). Shared understanding of the goal of the activity can also be achieved in a computer-based environment through preparation or staging activities that are intended to set the “stage” for the main activity which is typically more complex. These staging activities can be used to set expectations and increase learner motivation (Puntambekar & Hubscher, 2005). The scaffolding features of ongoing diagnosis and fading are more problematic in a computer-

based tool due their inherent dynamic and adaptive nature. Puntambekar & Hubscher (2005) argue that while many of the tools that purport to provide scaffolding actually don't due to their lack of adaptability to the student's level of understanding. Tailoring student support has most commonly been addressed in scaffolded cognitive tools by embedding the scaffolds within the structure of the tool as prompts or as representations (Reiser, 2004). Thus, these tools may be adaptable in the sense that they are in the control of the learners who can opt to utilize or ignore them. Pea (2004) explains that there is also concern that many software features may function as scaffolds-for-performance such that desired performances are only continuously achieved when learners utilize the scaffolds; that is, they do not function as scaffolds-with-fading. However, he further explains that as society increasingly relies on technology, the issue of scaffold-fading in software tools may become obsolete.

Educators, policymakers, and learners need to weigh the perceived risks affiliated with the loss of such support with the value of the incremental effort of learning how to do the task or activity unaided should such tools and supports ever become inaccessible, and the answer has to do with the social and technological assumptions humans make. As we approach a world in the coming years with pervasive computing with always-on Internet access, reliable quality of service networks, and sufficient levels of technological fluency, the context assumptions that help shape cultural values for distributed intelligence versus scaffolding with fading are changing (Pea, 2004, p. 442).

Even with, and perhaps due to the apparent complexities involved in operationalizing some of the key features of scaffolds within a software system, advances in technology and design has necessitated a theoretical framework for developing and evaluating scaffolding approaches in software tools (Quintana et al., 2004). In their proposal for such a theoretical framework, Quintana et al. (2004) describe seven guidelines that define ways in which tools modify the task to help learners succeed: (1) use representations and language that connect with learners' prior conceptions; (2) organize tools and interactions with tools around the

specific semantics of the discipline; (3) provide multiple representations of the information for the learner to explore and manipulate; (4) provide task structure so that learners can visualize next steps; (5) embed access to expert guidance; (6) automate nonsalient tasks to reduce cognitive load; and (7) facilitate articulation and reflection. While not necessarily a comprehensive list, these guidelines can serve as a basis for understanding the potential a software system has to instantiate scaffolding.

Technology Enabled Assessment

In response to federally mandated state accountability testing issued by NCLB, state departments of education are increasingly pressured to develop more effective and efficient strategies for measuring student performance. However, defining and developing these strategies has been a relatively slow process such that current testing methods do not serve the educational community as well as they should (Tucker, 2009). Tucker (2009) further discusses the direction of technology and education testing as envisioned by a research scientist by the name of Randy Bennett. Bennett envisioned in the late 1990s while at Educational Testing Service (ETS), that educational testing would reinvent itself in three stages (Tucker, 2009). The first stage would emphasize the use of technology to automate existing testing formats and processes. The second stage would involve using technology to develop more sophisticated test items, formats and scoring procedures to more accurately measure students' skills and abilities. The third stage envisioned was one in which assessment and teaching merged for the purposes of differentiating instruction and increasing learning outcomes (Tucker, 2009). While many of Bennett's envisions have not been fully enacted (Tucker, 2009), researchers and state departments of education are increasingly seeking and

adopting technological innovations to improve their assessment systems (Almond et al, 2010; Bechard et al, 2010; Koedinger, McLaughlin, & Heffernan, 2010; Tucker, 2009).

Technology enabled assessments (TEAs) are assessments that utilize technology to perform some function of the assessment process such as in the administration, scoring, or reporting of results (Bechard et al. 2010; Quellmaz & Pellegrino, 2009). TEAs have the capability of using interactive stimulus environments, innovative item formats, a greater range of response formats, and can more efficiently score, archive and report assessment results (Bechard et al., 2010; Quellmalz & Haertel, 2005). As such, TEAs are purported to increase testing efficiency, model effective teaching practices and provide more accurate measurements of student proficiency (Almond, et al; Bechard et al. 2010; Quellmaz & Pellegrino, 2009; Tucker, 2009). As Tucker (2009) points out, technology can not only dramatically improve assessment practices but more importantly, it can improve teaching and learning as well.

Aside from increased efficiency and innovation in design, TEAs have the potential to reveal cognitive skills and processes that may otherwise be undetected (Quellmalz, 2004). For instance, process indicators of performance may be documented that lead up to the final answer which could capture how a student arrived at his or her answer (Bennett, Persky, Weiss, & Jenkins, 2007). Similarly, complex cognitive skills such as scientific inquiry skills including identifying and evaluating relevant information, planning and conducting experiments, and interpreting results, may be more readily accessible through the use of technology (Puntambekar & Hubscher (2005).

Formative Use of TEAs. Formative assessments, which are the activities and processes undertaken to provide teachers and students feedback intended to differentiate

instruction and guide learning activities, can have a positive impact on student outcomes (Black & Wiliam, 1998; Kingston & Nash, 2010). Researchers contend that there are promising uses of TEAs for formative purposes (Almond et al, 2010; Koedinger et al, 2010; Quellmalz & Pellegrino, 2009; Tucker, 2009) such that technology is now capable of supporting the data collection and analysis as well as the individualized feedback and scaffolding needed in the formative use of assessment (Brown, Hinze & Pellegrino, 2008). Thus, it appears that potential exists to improve student learning outcomes by utilizing technological innovations within the formative assessment process.

As a tool used in the formative process, TEAs can also help promote the integration of assessment with instruction. As schools and teachers struggle to ameliorate the tensions associated with high-stakes testing (e.g., loss of instructional time and possible “teaching to the test”), these tensions beg the question as to how best to achieve accountability while maintaining optimal instructional practices (Koedinger, et al., 2010). As a vehicle for differentiated instruction, TEAs used formatively can provide students and teachers with direct and specific feedback they need to adjust teaching and learning activities while collecting assessment data in preparation for the summative test. In a sense, technology used in the formative assessment process is intended to extend and even emulate good teaching practices, not transform or replace them. That is, while the cognitive theory underlying such technologies may be intended to transform or enhance instructional practices, the technology itself is only meant to facilitate the instantiation of cognitive principles (Quellmalz, 2004; Tucker, 2009).

Scaffolded Assessments. One strategy used to integrate instruction with assessment as well as advance the state of TEAs is to incorporate scaffolding directly into the assessment

making the test itself a learning experience (Almond et al., 2010; Bechard et al., 2010; Camacho, 2009; Koedinger et al., 2010; Thissen-Roe, Hunt & Minstrell, 2004). Scaffolds in this context, “allow students who would otherwise get the item wrong to demonstrate what they do know about the item/task content” (Almond et al., 2010, p. 27). These scaffolds can be viewed as construct relevant supports that can assist students to respond more completely to an assessment item. While expectations for student performance remains the same, the opportunity for student responses across the ability continuum (i.e., including low-level ability students) is broadened (Almond et al., 2010).

Arguably one of the most appealing benefits of incorporating scaffolding into an assessment system is that the need for student performance data is addressed while simultaneously providing instructional assistance to students, thereby preventing the loss of instructional time that usually occurs during the assessment (Koedinger et al., 2010). Furthermore, Bechard et al., explain how “current psychometrics and test designs consistently yield relatively low levels of precision or high levels of measurement error for students at the “extremes” of performance” (2010, p. 23), including students with disabilities. This lack of precision can lead to invalid interpretations of test scores and a misrepresentation of students’ knowledge, skills and abilities which can ultimately lead to teachers making misinformed instructional decisions (Bechard et al., 2010). These researchers further explain that TEAs, such as those that incorporate scaffolding, have the potential to extend and adapt student performance particularly at the extremes, which can increase test score variability and thereby increase the reliability and validity of test score interpretations (Bechard et al., 2010).

Specific Examples of Scaffolded Assessments. There are at least two illustrative research projects that *explicitly* focus on incorporating instructional assistance in the form of

scaffolding into an assessment system: Assistments (Feng, Heffernan, & Koedinger, 2006a, 2006b; Koedinger, McLaughlin, & Heffernan, 2010) and Children’s Progress Academic Assessment (CPAA, Camacho, 2009). While other assessment or software programs may exist that incorporate scaffolding into the assessment process, these were the only two that were found in the literature by this researcher that were overtly characterized as assessments with scaffolding. For instance, an innovative program known as the DIAGNOSER (Thissen-Roe, Hunt & Minstrell, 2004) is a web-based adaptive instructional tool that is used as a formative assessment tool to provide continuous feedback to students and teachers. Although this system achieves the goal of merging instruction with assessment, it does so by emphasizing student and teacher feedback intended to illustrate student misconceptions about “facets” of knowledge (Thissen-Roe, Hunt & Minstrell, 2004) rather than through scaffolding. Conversely, there are numerous instructional software tools that incorporate various types of scaffolding features to assist learners to achieve a specific learning goal such as to design a computer program or scientific experiment (e.g., the Biology Guided Interactive Learning Environment, Reiser et al., 2001; Learning by Design™, Kolodner et al., 2003; Knowledge Integration Environment, Linn, 1995; and Model-It, Jackson, Krajcik, et al., 1998, Jackson, Stratford et al., 1998); however, none of these tools are specifically intended for the purposes of gathering student assessment data. The focus of the present study is on the Assistments system.

The Assistments System. The Assistments system is a web-based mathematics cognitive tutor developed for middle school students for the purposes of addressing the need for assessment while simultaneously providing instruction to students (Koedinger et al., 2010; Heffernan & Heffernan, 2008). As such, the self-described name “Assistment” was coined by

co-founder of the program, Ken Koedinger to express the *assistance* that students receive during the *assessment* (Koedinger et al., 2010). The ultimate goal of the Assistments project is to help students increase learning and achieve proficiency on their state accountability test.

Assistments function as an assessment tool by collecting data on a variety of metrics including the typical correct/incorrect responses to all questions (including scaffolded questions described below) but also measures of the amount of assistance needed by a student to complete an item in the form of number of hints requested, response time, and number of opportunities to practice (Koedinger et al., 2010). As students complete an Assistment, the system gathers information to determine strengths and weaknesses of the individual student as well as of the whole class. This information can be used formatively to guide subsequent teaching and learning activities (Koedinger et al., 2010).

Assistments also function as an instructional tool, first by breaking down items into requisite skills and knowledge components, and second by providing hints to assist the learner throughout the test that are made available upon the learner's request (Koedinger et al., 2010). These broken down knowledge components, or scaffolded questions, are intended to more precisely determine where a student's misconception lies if he or she provides an incorrect response to an item. For instance, a geometry question that involves understanding the concept of congruency may also require measurement skills (e.g., to understand the concept of perimeter) as well as skills in patterns, relations and algebra (e.g., to solve equations). While the original item might address congruency, the scaffolded questions would address each of these requisite skills needed to answer the original item correctly. Hints, on the other hand, are described as "suggestions on how to proceed and often appear as a definition or question similar to what a human tutor might ask or say" (Koedinger, 2010, p. 494). A

student can ask for a hint at any time during the assessment when he or she is confused or does not know how to proceed.

Figures 1 – 5 display an example Assistentment item (adapted from Heffernan & Heffernan, 2008) with accompanying scaffold questions and hints. The example item is based on the concept of congruency of triangles which is broken down into several knowledge components that students need to know to be successful on this item. Specifically, students need to know geometry to understand the meaning of congruent triangles; measurement to understand what and how to apply the concept of perimeter, as well as patterns, relations and algebra to understand how to solve an equation and expressions (Heffernan & Heffernan, 2008).

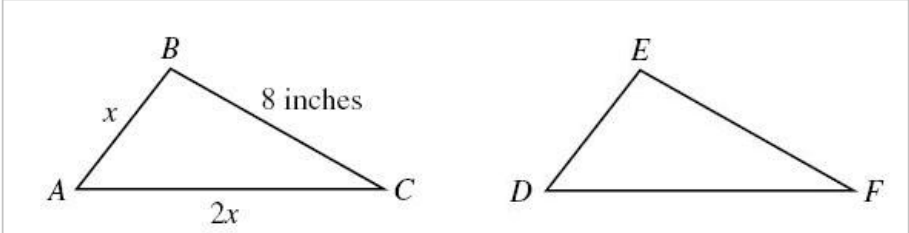
http://www.assistment.org/ - Assistment - Previewing Content - Windows Internet Explorer

Assistment

You are previewing content. Item 19 G-2003 (Congruent triangles) (#4468)

Triangles ABC and DEF are congruent. The perimeter of triangle ABC is 23 inches.

What is the length of side DF in triangle DEF?



[Comment on this question](#)

[Break this problem into steps](#)

Type your answer below (mathematical expression):

[Submit Answer](#)

Figure 1. Example Assistments main item on congruent triangles. Adapted from Heffernan, N. & Heffernan, C. (2008). Assistments: Teacher's Manual. Retrieved from <http://teacherwiki.assistment.org/wiki/images/8/8b/Teachermanualsinglesided.pdf>.

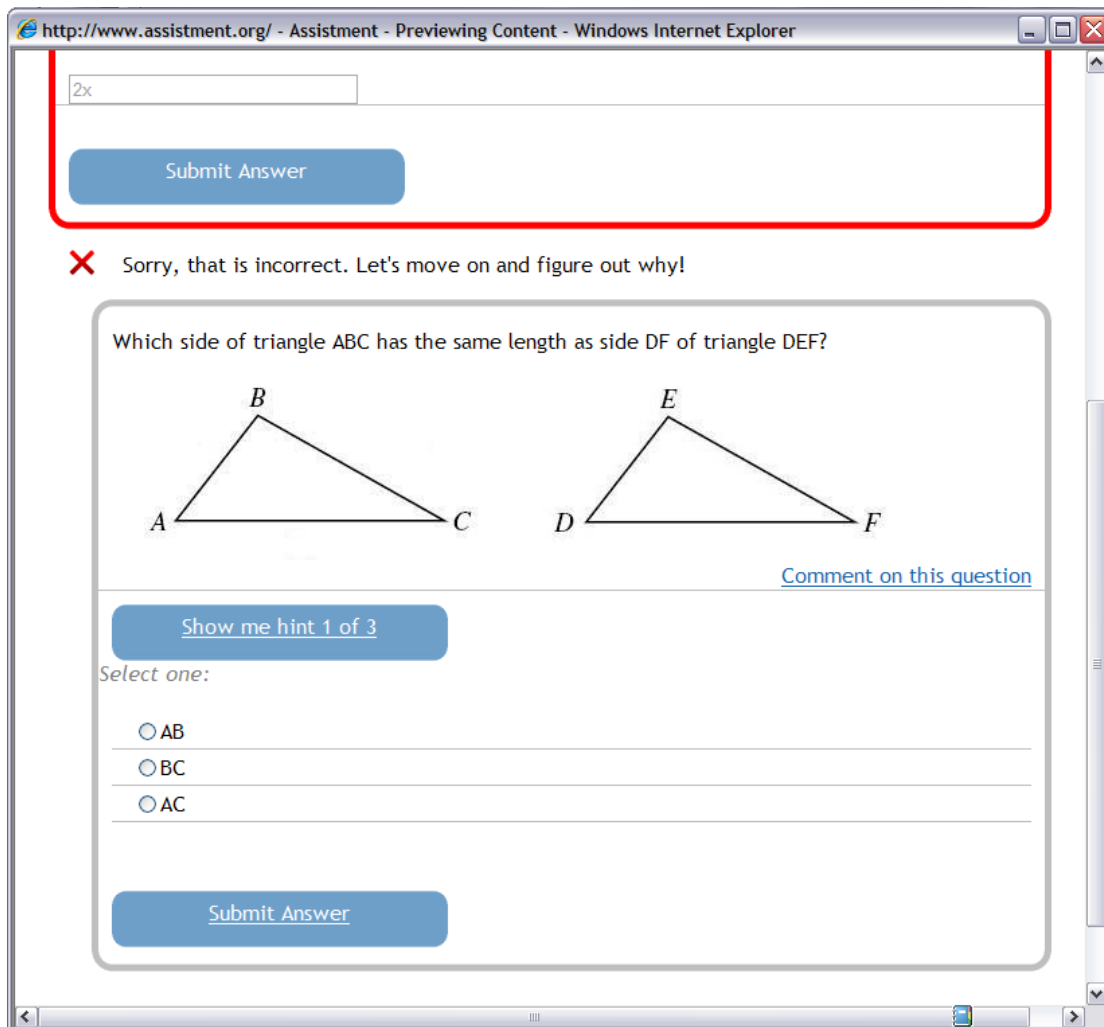


Figure 2. First scaffold question for the example Assistent's item on congruent triangles. Adapted from Heffernan, N. & Heffernan, C. (2008). Assistent's: Teacher's Manual. Retrieved from http://teacherwiki.Assistent.org/wiki/images/8/8b/Teachermanualsingle_sided.pdf.

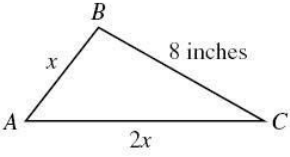
This student answered the original item on congruent triangles incorrectly. He or she was then directed towards the first scaffold question which addresses the congruence skill (geometry) apart from the other skills required in the original question. This student answered the first scaffold question correctly without the use of any hints.

http://www.assistment.org/ - Assistment - Previewing Content - Windows Internet Explorer

Submit Answer

✓ Correct!

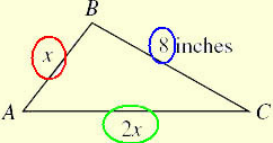
Which expression represents the perimeter of triangle ABC?



[Comment on this question](#)

Perimeter is defined as the sum of all sides of a figure. [Comment on this hint](#)

The perimeter of triangle ABC is the sum of all its sides. [Comment on this hint](#)



The perimeter is equal to $2x + x + 8$.

Select $2x + x + 8$ [Comment on this hint](#)

Select one:

$2x + 8$

$2x + x + 8$

$\frac{1}{2} * 8x$

$\frac{1}{2} * x(2x)$

Submit Answer

No. You might be thinking that the area is $\frac{1}{2}$ base times height, but you are looking for the perimeter.

Figure 3. Second scaffold question for the example Assistments item on congruent triangles. Adapted from Heffernan, N. & Heffernan, C. (2008). Assistments: Teacher's Manual. Retrieved from <http://teacherwiki.Assistment.org/wiki/images/8/8b/Teachermanualsingle-sided.pdf>.

This student did not know how to answer the second scaffold question (i.e., a measurement skill) and requested all three hints available for this question. The student did not select the correct answer and was provided a buggy message that responded to the specific error the student made. The last hint always shows the correct answer so that the student is able to move on to the next scaffold question.

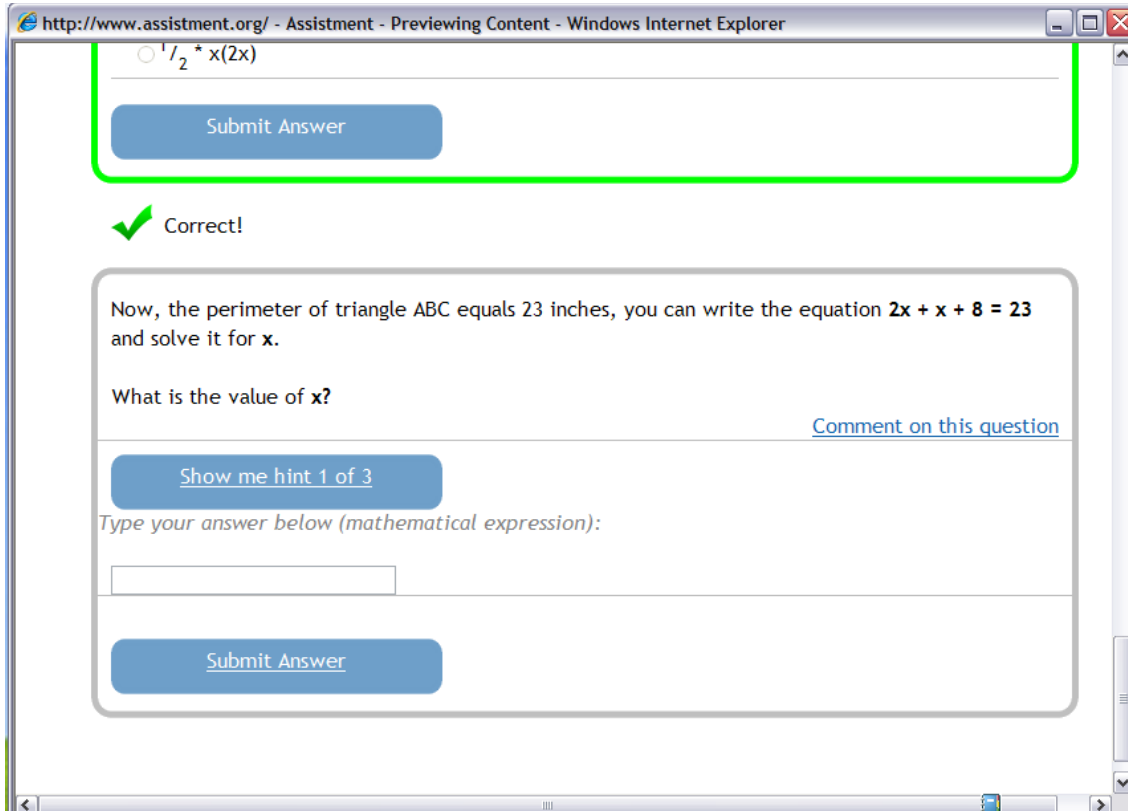


Figure 4. Third scaffold question for the example Assistments item on congruent triangles. Adapted from Heffernan, N. & Heffernan, C. (2008). Assistments: Teacher's Manual. Retrieved from <http://teacherwiki.Assistment.org/wiki/images/8/8b/Teachermanualsingle-sided.pdf>.

The third scaffold question deals with patterns, relations and algebra. This student answered the question correctly without requesting any hints.

http://www.assistment.org/ - Assistment - Previewing Content - Windows Internet Explorer

5

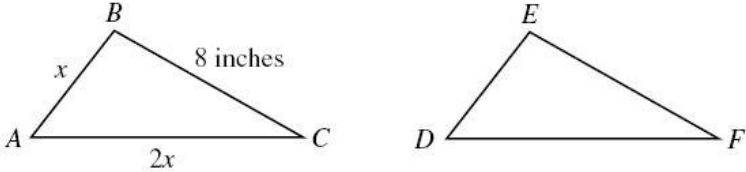
Submit Answer

✓ Correct!

Excellent. Now, try solving the original problem again.

Triangles ABC and DEF are congruent. The perimeter of triangle ABC is 23 inches.

What is the length of side DF in triangle DEF?



[Comment on this question](#)

Show me hint 1 of 4

Type your answer below (mathematical expression):

Submit Answer

Figure 5. Fourth scaffold question for the example Assistments item on congruent triangles. Adapted from Heffernan, N. & Heffernan, C. (2008). Assistments: Teacher's Manual. Retrieved from <http://teacherwiki.Assistment.org/wiki/images/8/8b/Teachermanualsingle-sided.pdf>.

The last scaffold question returns to the original item and asks the student to try it again, now with the knowledge and understanding of the individual steps needed to answer it correctly. If the student solved the previous scaffold questions correctly, the last step needed is a basic multiplication problem (i.e., 5×2).

The original Assistment items were based on previously published Massachusetts Comprehensive Assessment System (MCAS) test items and are both multiple-choice format

and open-ended, fill-in-the-blank questions. As displayed in Figure 6 below, the Assistentment process can be described as follows: (Koedinger et al., 2010; Feng et al., 2006a; Feng et al., 2006b; Heffernan & Heffernan, 2008):

- An item is presented.
- Student provides correct response: the next item is presented.
- Student provides incorrect response: a “tutoring” session is provided. The tutoring session involves presenting the student a series of scaffolded questions that break the original item down into knowledge components or steps. The number of scaffolded questions associated with each item depends on the number of independent skills needed to complete the question.
- Student does not provide a response. That is, the student is confused and does not know how to proceed. The student has the option to go directly to the scaffolded questions to help him know what to do next.
 - The first scaffolded question is presented and the student has the option of accessing a number of hints to help him or her determine the correct answer to the scaffolded question. The last hint essentially gives the correct answer to the question so that the student does not get become frustrated if he or she does not know the correct answer.
 - Once the student has answered all the scaffolded questions, he or she is presented with a form of the original item again and given the opportunity to respond.

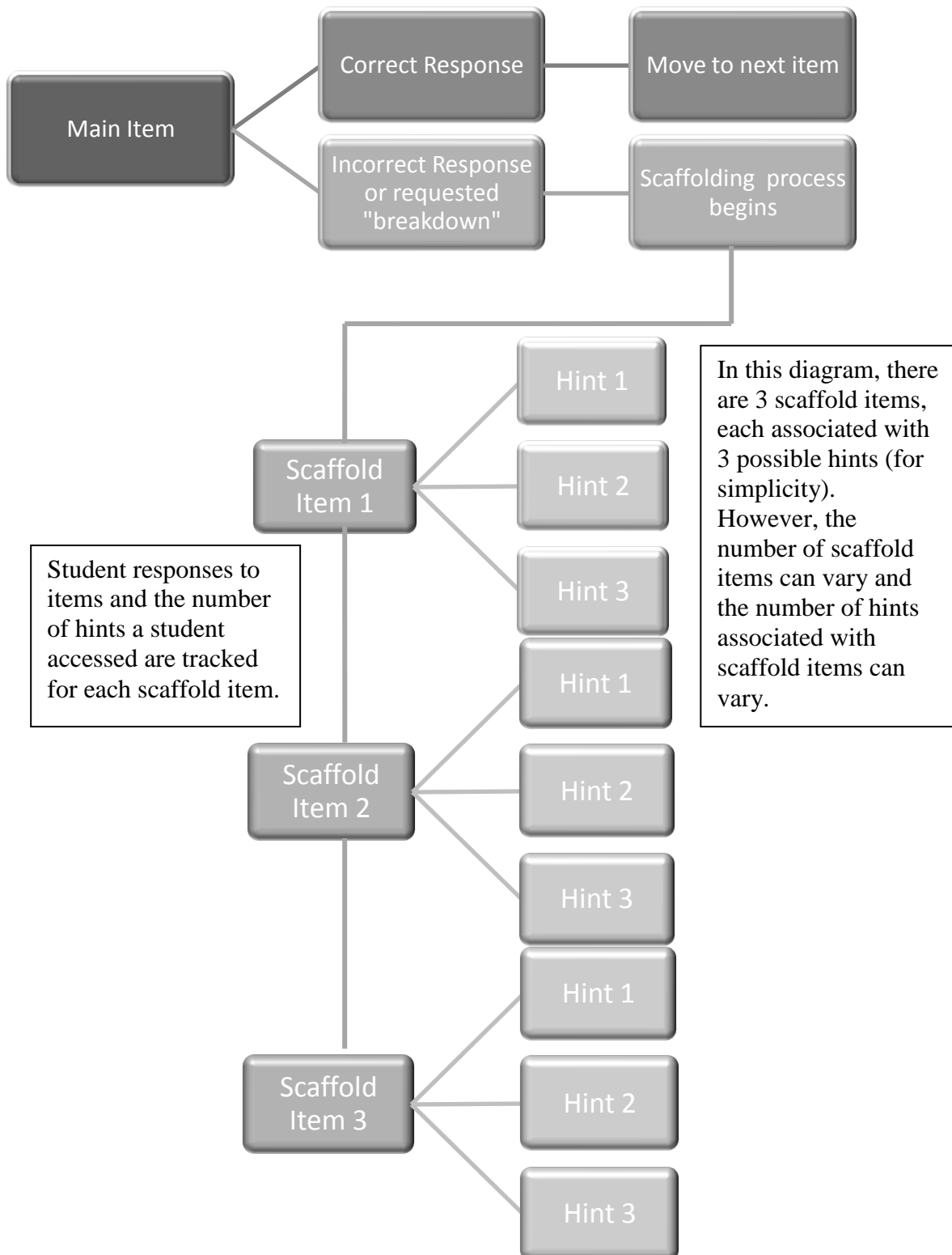


Figure 6. Assisments item example flowchart.

Currently, students receive scores on Assistments based on whether the student answered the original item correctly or incorrectly upon the first presentation of that item. If students choose to go directly to the scaffold questions without providing a response, the item is scored as incorrect. Thus, any time a student receives assistance, the student does not receive any credit on that item regardless of performance on the scaffolded questions. Keep in mind that the teacher also has access to these student performance measures such as the number of hints requested and the number of correctly answered scaffold items which can effectively be used for formative purposes. However, the percent correct score for the main test items does not account for the amount of assistance needed; if assistance is needed, it is reported as incorrect. In other words, partial credit for scaffolded questions is not given in the percent correct score.

Scoring Scaffolded Assessments. It is apparent that significant progress has been made in recent years in the area of TEAs, particularly illustrated by the innovative assessment systems that incorporate scaffolding into the assessment process. However, the area of psychometrics has yet to venture directly into these advancements in technology to determine how statistical methods and procedures can be used to fully capture students' knowledge, skills and abilities as measured by TEAs (Almond et al., 2010; Bechard et al, 2010; Bennett & Gitomer, 2009). As Bennett & Gitomer note, the state of technology and assessment relies not only on advances in learning theory, cognitive science and technology but it also depends on advancing psychometric approaches that characterize how the student interacts with the assessment. Almond et al. (2010) explicitly state this topic as an area of research that is needed to advance the state of TEAs. That is, what types of scoring models can be used, that

are currently used in the field or those used in other fields, to provide valid inferences about students' performance on scaffolded assessments (Almond et al., 2010)?

While the field of psychometrics has been relatively slow to progress in this area, research conducted specifically on the Assistments system has been more advanced. There is a continuous body of research that focuses specifically on predicting state assessment scores with the various metrics obtained during the Assistment process. Methods for predicting state exam scores have included using monthly aggregates of Assistments metrics (Anozie & Junker, 2006), students' skills sets from a Bayes nets approach (Pardos, Heffernan, Anderson & Heffernan, 2006), linear growth curve models for student performance (Feng, Heffernan & Koedinger, 2006a), a linear logistic test model to account for skill type (Ayers & Junker, 2006), and the Rasch model for dichotomous responses (Ayers & Junker, 2008). While each of these methods has demonstrated various degrees of success (or non-success) in predicting state assessment performance, these researchers continue to seek models that can reduce prediction errors and account for the unique instructional features of the system (Ayers & Junker, 2006; Feng, Heffernan & Koedinger, 2006).

Purpose of Research

The purpose of this research is to help advance the development and use of TEAs, specifically those that incorporate scaffolding into the assessment process by comparing several different scoring models for an example assessment system and evaluating criterion-related validity evidence for the scoring models. Specifically, the research questions in this study are as follows:

- 1) What type of model is the optimal scoring model for the scaffolded data provided by the Assistment system?

- a. Which scoring model produces the best model fit for the Assistentment system?
 - b. Which scoring model produces the most precise measures of student ability?
 - c. Do the benefits associated with the better fitting model outweigh any practical concerns due to model complexity?
- 2) Is there a relationship between student ability estimates derived from the scoring models and a criterion measure of student achievement?
- a. Do any of the scoring models provide student ability estimates that predict a criterion measure of student ability better than a simple percent correct score?
 - b. Do the models that account for the scaffolding features have a stronger relationship with a criterion measure of student achievement than the models that do not account for those features?
 - c. Do the models that account for the local dependence have a stronger relationship with a criterion measure of student achievement than the models that do not account for the dependence?

As this research is exploratory in nature, formal hypotheses are not presented. In general, as models progress in complexity and account for more specific features of the data, it is reasonable to believe that model fit will improve and ability estimates will become more precise. This does not necessarily warrant adoption of a more complex model, rather the simpler model may be judged to provide “accurate enough” estimates more efficiently and at a more reasonable cost. In any case, information will be presented for all models and a discussion will follow outlining the costs and benefits associated with each.

Item Response Theory & Assumptions

Item response theory (IRT) is now widely used to model data from educational and psychological tests, instruments and inventories. IRT is a statistical theory linking a trait (i.e., what the test purports to measure) and responses of examinees to assessment items through mathematical models. That is, it models the probability of success on an item given examinee traits or ability levels and item characteristics (Hambleton, Swaminathan & Rogers, 1991). IRT is based on a monotonically increasing logistic curve known as the item characteristic curve (ICC).

IRT operates under a set of common assumptions and properties. While the assumptions of unidimensionality and local independence are often discussed as two assumptions, they are, for all intents and purposes, the same. The unidimensionality assumption states that only one trait is measured by the items that make up the test (i.e., the test measures only one construct). Items on a test are considered to be unidimensional when a single factor or trait accounts for a substantial portion of the test score variance (Hambleton, Swaminathan & Rogers, 1991). In this sense, all the items are “tapping” into a common construct. The local independence assumption states that item responses are independent of each other, given ability. In other words, the correlation between item responses should equal zero when examinee ability is partialled out. Thus, the abilities that are specified in the model are the only factors that influence examinee responses and if the unidimensionality assumption holds, then there is only one factor that accounts for the entire latent ability space (Hambleton, Swaminathan & Rogers, 1991). These assumptions are commonly violated one in many operational contexts and they are discussed in further detail in subsequent sections of this paper.

IRT models further maintain two desirable properties when the model fit the data: the nature of the ICC and parameter invariance. The nature of the ICC models the probability of success based on a monotonically increasing function such that higher trait or ability levels results in a higher probability of success on a given item. The property of invariance states that item parameters are invariant over samples of examinees and ability parameters are invariant over samples of items from, within the linear transformation that accounts for the arbitrariness of the scale. That is, the parameters that characterize an item do not depend on the ability distribution of the examinees and the parameters that characterize an examinee do not depend on the set of test items (Hambleton, Swaminathan & Rogers, 1991).

IRT models permit one or more traits to be included as well as the use of dichotomous or polytomous data. Thus, there are many different types of item response models that may differ in the mathematical form of the item characteristic function or in the number of parameters specified in the model (Hambleton, Swaminathan & Rogers, 1991). However, all models contain at least one item parameter and at least one examinee parameter. Several item response models are described in the following sections which focus first on basic dichotomous models followed by unidimensional polytomous models, and finally a description of a type of multidimensional model known as the testlet model.

Dichotomous IRT Models. Dichotomous IRT models describe the nonlinear relationship between examinee trait level and the probability of correctly responding to an item when the item has only two scoring options (i.e., correct or incorrect; $x = 1$ or $x = 0$), such as a multiple-choice item with only one correct response. The three most commonly used dichotomous IRT models are: 1) the one-parameter logistic model (1PL; Rasch, 1960); 2) the two-parameter logistic model (2PL; Birnbaum, 1968); and 3) the three-parameter logistic

model (3PL; Birnbaum, 1968). These models describe the relationship between examinee ability, θ , and the probability of a correct response with up to three parameters that characterize the item: the level of item difficulty, b , discrimination, a , and examinee guessing behavior or the lower asymptote, c .

The 1PL Model. The 1PL, or Rasch model (Rasch, 1960) allows for one item parameter which describes the level of item difficulty (b) or the location of the position of the ICC in relation to the ability scale. Specifically, it is the theta level (θ) that corresponds to the point of inflection of the ICC where the probability of answering the item correctly is 0.5. In other words, the difficulty parameter is the value on the ability scale where the ICC slope is the steepest. The more difficult the item, the more the ICC shifts farther to the right. The ratio between examinee ability level and item difficulty are assumed to be constant in this model. Hence, as the b parameter increases, more ability is needed for an examinee to have a 50% chance of getting the item correct. The 1PL defines the probability of success ($x = 1$) for a person j with a given ability level (θ_j) on item i as:

$$P_j(\theta) = \frac{e^{(\theta-b_i)}}{1 + e^{(\theta-b_i)}}, \quad (1)$$

where b is the difficulty parameter for item i .

The 2PL Model. The 2PL model was proposed by Birnbaum (1968) and allows for an additional item parameter which describes the degree to which the item discriminates between low ability and high ability examinees. The discrimination parameter (a) is proportional to the slope of the ICC at point b , or the point of inflection on the ability scale (Hambleton, Swaminathan & Rogers, 1991). Thus, the steeper the slope of the ICC, the more useful the

item is for distinguishing between high and low ability examinees. The 2PL model defines the probability of success as:

$$P_j(\theta) = \frac{e^{Da_i(\theta-b_i)}}{1 + e^{Da_i(\theta-b_i)}}, \quad (2)$$

where a is the discrimination parameter for item i . An additional element was added to this model which is described as a scaling factor, D . It was shown that when $D = 1.7$, the logistic function that this model is based on more closely resembled the normal ogive function (Birnbaum, 1968) which was the basis for the original 2PL model proposed by Lord (1952).

The 3PL Model. Finally, the 3PL model, also proposed by Birnbaum (1968), extended the previous model by further accounting for an item parameter that characterized examinee pseudo-guessing behavior (c). The c parameter effects the lower asymptote of the ICC and reflects the probability of low ability examinees correctly guessing the answer to an item. For instance, on a four option multiple choice item, an examinee will have a 25% chance of answering the item correctly simply by randomly choosing one of the options; thus, the lower asymptote is adjusted to 0.25 to account for the probability of guessing on this item. The 3PL model is displayed in Equation 3 below:

$$P_j(\theta) = c_i + 1 - c_i \frac{e^{Da_i(\theta-b_i)}}{1 + e^{Da_i(\theta-b_i)}}. \quad (3)$$

Overall, as parameters are added to the model, the more information is needed to estimate an examinee's probability of success on an item. A sample dichotomously scored item is displayed in Figure 7 below to illustrate how the three parameters impact the ICC. The 1PL model has a difficulty parameter equal to 1.0; the 2PL includes a discrimination parameter equal to 1.5; and the 3PL provides the additional pseudo-guessing parameter set at 0.2. In the example 1PL model, an examinee with an average ability level ($\theta = 0.0$) has a 50%

chance of answering the item correctly. By adding the discrimination parameter in the 2PL model, the slope at that inflexion point becomes steeper; however. Finally, by accounting for potential guessing behavior, the 3PL shifts the lower asymptote upward which means that an examinee with average ability actually has a 60% probability of success on the item. Figure 8 below displays the same item based on the 3PL model but details the each parameter value.

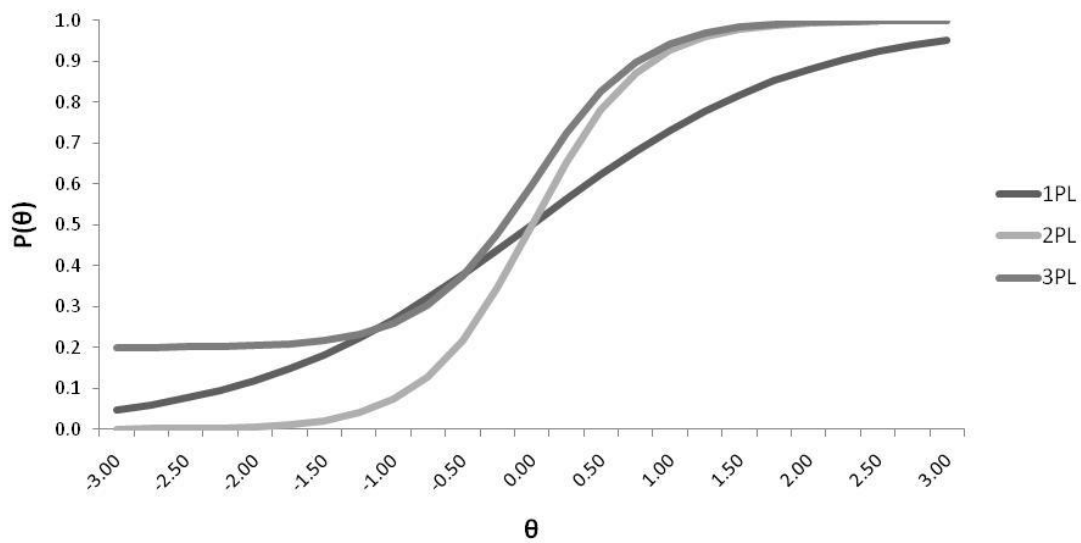


Figure 7. ICCs of a dichotomously scored item based on the 1PL ($b = 0.0$), 2PL ($a = 1.5$) and 3PL ($c = 2.0$)

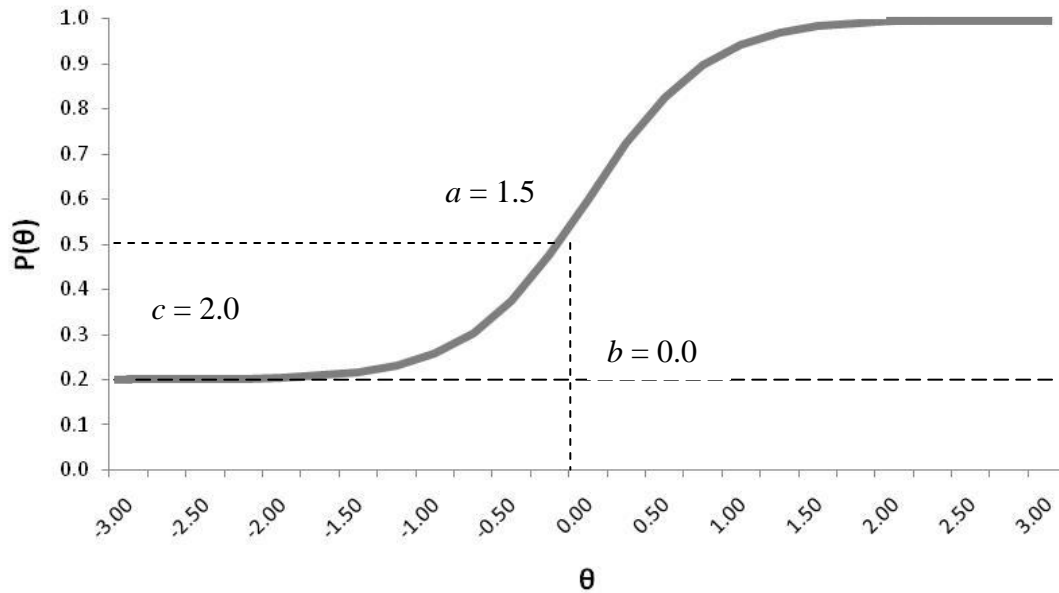


Figure 8. ICC of a dichotomously scored item based on the 3PL ($b = 0.0$, $a = 1.5$, $c = 2.0$)

Item & Test Information. In IRT the quality of an item is evaluated on the degree of measurement precision that it provides at a given ability level. This precision of measurement is known as the item information function (IIF) which indicates how useful an item is at differentiating examinees for any given ability level (Reise, Ainsworth & Haviland, 2005). In other words, information functions indicate how useful an item is at distinguishing examinees of lower ability levels from those with higher ability levels; the more informative (or useful) an item is, the more precise it is at making these distinctions. Information is a function of examinee ability (θ); a particular item could be very informative at some ability levels and uninformative at others. For a dichotomous IRT model, the item information function, $I_i(\theta)$, is expressed as:

$$I_i(\theta) = \frac{2.89a^2(1-c_i)}{\left[c_i + e^{Da_i(\theta-b_i)} \right] \left[1 + e^{Da_i(\theta-b_i)} \right]} \quad (4)$$

In general, relatively easy items are more informative for discriminating among examinees low on the trait of interest whereas more difficult items are more informative for discriminating among examinees high on the trait (Reise, Ainsworth & Haviland, 2005). For every item, as discrimination (a) increases, information increases; as the probability of guessing (c) the right answer increases, information decreases; and as difficulty (b) approaches ability, information increases (Hambleton, Swaminathan & Rogers, 1991).

IIFs can be summed across an entire scale or test to create a test information function (TIF). The information function for a test becomes:

$$I(\theta) = \sum_{i=1}^n I_i(\theta), \quad (5)$$

where $I(\theta)$ is simply the sum of all IIFs at θ . Since the contribution of each item to the TIF is independent of other items' contribution, items can be evaluated independently. All things equal, adding more items to a test provides increased measurement precision. In this sense, IIFs can be used to evaluate the usefulness of individual items in the context of developing a new test or reconstructing an old test. The ability to add IIFs to create an information function for a test is the cornerstone of scale construction in IRT (Reise, Ainsworth & Haviland, 2005).

The amount of information provided by a test at a given ability level, θ , is inversely related to the standard error of estimation, or the precision of ability estimation at that point on the ability scale. The standard error of estimation is defined as:

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}. \quad (6)$$

Polytomous IRT Models. Up until the late 1960's, IRT models used to handle dichotomous data were sufficient for handling most measurement situations. When polytomous data presented itself, researchers simply dichotomized the data prior to analyses and typically used one of the previously described models. It wasn't until Samejima (1969) proposed the graded response model which could account for polytomous data such as responses from items on a Likert scale, that research in this area started to emerge, at first slowly and then with more frequency in the 1980s (van der Linden & Hambleton, 1997). Since then many different types and variations of models have been developed to represent polytomous data.

Polytomous IRT models are needed when items are scored according to multiple response categories (i.e., not scored simply as right or wrong). For example, polytomous data may be obtained from essay type items scored on a rubric, Likert scale items, items with possible partial credit such as multiple-step math problems, performance tasks, or portfolio assessments. Several benefits may be acquired from administering assessments that use polytomous items including: greater efficiency in that fewer polytomous items are typically needed to achieve the same degree of reliability as would be obtained with more dichotomous items, and certain traits are more easily measured and/or more accurately measured on rating scales (van der Ark, 2001).

Thissen & Steinberg (1986) proposed that polytomous IRT (PIRT) models could be classified into two main categories: difference models and divide-by-total models. Difference models include Samejima's (1969) graded response model while divide-by-total models are commonly represented by the partial credit model (PCM; Masters, 1982) and the generalized partial credit model (GPCM; Muraki, 1992). While difference models define the probability

of a response in category k as $P^*(k) - P^*(k + 1)$, divide-by-total models define the probability of a response in category k or category $k - 1$ which results in an exponential that is divided by a sum of the total of all the exponentials (Thissen & Steinberg, 1986). For the purposes of this paper, two related types of difference models are presented.

Graded Response Model. The graded response model (GRM; Samejima, 1969) is appropriate to use with items that have two or more ordered response categories such as letter grading, performance evaluations, or partial credit given on a problem. The GRM preserves the order of the score category thresholds, unlike the PCM or GPCM. The GRM models the cumulative probability of an examinee obtaining a score in a given category (x) or higher with a given ability (θ). Thus for any item, i , there are $x_i + 1$ scoring categories where x_i is the highest possible score, 0 is the lowest, and there are x_i boundaries between categories. For any given category, the model is the same as the dichotomous 2PL model and is described as,

$$P_{ix}^*(\theta) = \frac{e^{Da_i(\theta - b_{ix})}}{1 + e^{Da_i(\theta - b_{ix})}}, \quad (7)$$

where $P_{ix}^*(\theta)$ is the probability of scoring in category x or higher, i represents 1 to n number of items, and x represents the category boundaries for item i from 0 to the highest possible score for item i . Moreover, a_i is the discrimination parameter for item i and b_{ix} is the category boundary for category x of item i (Samejima, 1969) and D is a scaling constant equal to 1.7. Thus, the probability associated with each scoring category is derived by subtracting the cumulative probabilities for adjacent categories and can be defined as,

$$P_{ix}(\theta) = P_{ix}^*(\theta) - P_{ix+1}^*(\theta). \quad (8)$$

Ordinal Response Model. The ordinal response model (ORM) is analogous to the GRM (Wang, Bradlow & Wainer, 2005); however, it defines the cumulative probability

slightly differently than the conventional form that is commonly used for the GRM. Specifically, the ORM models the cumulative probability of an examinee obtaining a score in a given category (x) or *lower* with a given ability (θ). While the GRM predefines the probability of responding in or above the lowest category as equal to 1.0 and the probability of responding above the highest category as equal to 0.0, the ORM defines these two categories in reverse order. That is, for the ORM the probability of responding at or below the highest category is equal to 1.0 and the probability of responding below the lowest category is by definition equal to 0.0. According to Wang, Bradlow & Wainer (2005), the ORM defines the cumulative probability of scoring in a given category (x) or lower conditioned on θ as,

$$P_{ix}^*(\theta) = \Phi(k_x - a_i(\theta - b_i)) \quad (9)$$

where Φ is the normal cumulative density function and k_x , are the latent cutoffs. The normal cumulative density function is approximately equal to the logistic function when θ is multiplied by the factor $D = 1.7$ (Hambleton, Swaminathan & Rogers, 1991). Thus, for the ORM, the probability of scoring in a given category (x) is equal to,

$$P_{ix}^*(\theta) = \frac{1}{1 + e^{(-k_x - (Da_i(\theta - b_i)))}} - \frac{1}{1 + e^{(-k_{x-1} - (Da_i(\theta - b_i)))}} \quad (10)$$

Item & Test Information. Since polytomous IRT models encompass multiple parameters to calculate the probability of response categories, producing the IIFs and TIFs is a more complex process than the process used for dichotomous IRT models. The IIF for a polytomous item can be defined as (Samejima, 1969),

$$I_i(\theta) = \sum_{x=0}^{m_i} \frac{[P_{ix}'(\theta)]^2}{P_{ix}(\theta)} \quad (11)$$

where $P_{ix}(\theta)$ is the probability associated with obtaining a category score of x on item i , given ability, θ , and $P'_{ix}(\theta)$ is the first derivative of $P_{ix}(\theta)$. Formula 11 has been shown to be equal to,

$$I_i(\theta) = a_i^2 \left[\sum_{k=1}^m k^2 P_{ik}(\theta) - \left(\sum_{k=1}^m k P_{ik}(\theta) \right)^2 \right] \quad (12)$$

where m is the number of categories and $P_{ik}(\theta)$ is the probability that person with a given ability θ will be in score category k of item i (Dodd, De Ayala, and Koch, 1995; Veldkamp, 2003). Similar to the dichotomous TIF, the polytomous TIF is the sum of the IIFs and it is related to the standard error of measurement by inverting the TIF.

Bundle Models. The models discussed thus far are based on the aforementioned assumption of local independence. That is, item responses are independent of each other, given ability, such that the probability of obtaining a set of item responses is equal to the product of individual item probabilities. Fitting standard item response models to groups of interdependent items may result in (1) bias in item difficulty estimates, (2) inflated item discrimination estimates, (3) overestimation of the precision of ability estimates, and (4) overestimation of test reliability and test information (Wang & Wilson, 2005a; Weng, Cheng & Wilson, 2005; Zhang, Shen & Cannady, 2000). These biased and overstated estimates can lead to inaccurate inferences about the parameters (Sireci, Thissen, & Wainer, 1991; Wainer, 1995; Wainer & Thissen, 1996; Wang & Wilson, 2005a). In any case, many operational testing situations utilize item groupings that have built in dependencies and do not support the assumption of local independence. Thus, models have been developed to explain item dependencies that are unaccounted for by the latent trait by treating a group of interdependent

items as a unit where the units are assumed to be locally independent (Hoskens & De Beock, 1997; Wainer & Kiely, 1987; Wilson & Adams, 1995).

Units composed of interdependent items have been referred to as subtests (Andrich, 1985), testlets (Wainer & Kiely, 1987), and item bundles (Rosenbaum, 1988; Wilson & Adams, 1995). Wainer & Kiely (1987) originally denoted the term “testlet” to describe locally dependent items in a computerized adaptive testing context that allows for different pathways of administered items. On the other hand, Rosenbaum (1988) coined the term “item bundle” to more generally describe items on a test that share a common stimulus or item stem, or have similar content or structure. As Rosenbaum’s description more closely aligns with the type of item grouping that occurs in the Assisments (i.e., original and scaffold items are based on common content), the term “item bundle” is used throughout this paper to refer to the groups of items that are created by the scaffolding process.

Rosenbaum (1988) proposed that to account for local item dependence, the local independence assumption could be reformulated to describe independence between item bundles rather than items themselves (Rosenbaum, 1988). Using the dichotomous Rasch model, item bundle models can be described, where x_{ci} is a response to item i in bundle c and the vector of responses to items in bundle c is $x_c = (x_{c1}, x_{c2}, \dots, x_{cIc})'$ (Wilson and Adams, 1995). Vector responses can be accumulated into a test response vector denoted $x = (x'_1, x'_2, \dots, x'_I)'$. The probability of the item response becomes

$$P(x_{ci}; \delta_{ci} | \theta) = \frac{e^{x_{ci}(\theta - \delta_{ci})}}{1 + e^{x_{ci}(\theta - \delta_{ci})}}, \quad (13)$$

where δ_{ci} is the item parameter for item i in bundle c . Thus, a vector of item parameters belonging to each bundle c can be written as $\delta = (\delta'_{c1}, \delta'_{c2}, \dots, \delta'_{cC})'$. Based on Rosenbaum (1988), bundle independence is then defined as

$$\Pr(x; \delta | \theta) = \prod_{c=1}^C \Pr(x_c; \xi | \theta), \quad (14)$$

where ξ is the vector of item parameters and x_c is the probability of a bundle response.

Following Wilson & Adams (1995), methods for analyzing item bundles typically adhere to two basic approaches which are characterized as either score-based or item-based. Score-based approaches use the summed scores on the items within a bundle as a starting point for modeling; these summed scores are then used to correspond to the response categories of an artificial polytomous item. Item-based approaches, on the other hand, use item responses to single items as the starting point rather than summed scores. Response patterns of an item bundle are treated as a unit rather than responses to single items (Wang et al., 2005; Wilson & Adams, 1995). Both approaches rely on Rosenbaum's theorem for local bundle independence described above in that the likelihood of a response vector is the product of the probabilities of responses to the bundles rather than items (Wang et al., 2005; Zhang, et al., 2010). Each of these methods is described in more detail in this section.

Score-based Approaches. The score-based approach typically involves directly applying a polytomous IRT model to sets of items so that response patterns of item units are treated as categories in one polytomous "super-item" (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wainer, 1995; Wainer & Kiely, 1987). Essentially, item scores are summed within each bundle such that when total scores are identical, they are assigned to the same category (Wang et al., 2005; Zhang, et al., 2010). The bundle is then

scored polytomously and bundle item responses are calibrated using any of the various polytomous IRT models such as the GRM, the PCM, or the Rating Scale Model (Andrich, 1978). This approach maintains local independence across item bundles while eliminating dependencies within bundles which addresses issues related to the overestimation of test information that exist when dependencies are ignored. However, one major shortcoming of this method is that information is lost in that the exact response patterns within the bundle (Wang et al, 2005; Wang & Wilson, 2005a; Wilson & Adams, 1995). Thus, it has been suggested that the polytomous approach to scoring item bundles might be appropriate when the local dependence between items within a bundle is moderate and the test contains a large proportion of independent items (Wainer, 1995).

Item-based Approaches. Item-based approaches aren't as straightforward and vary in nature more so than the score-based approaches. In general, these approaches account for local dependency within an item bundle by adding an additional component to the model that is either a random effect (Bradlow, Wainer & Wang, 1999; Wainer, Bradlow & Du, 2000; Wang, Bradlow, & Wainer, 2002) or a fixed effect (Hoskens & deBoeck, 1997). The most common of these approaches in the educational measurement field is the testlet model approach (Rijmen, 2010) which adds a random effect parameter to model the local dependence among items within the same bundle (Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000; Wang, Bradlow, & Wainer, 2002). The random effect approach essentially views local item independence as a person characteristic rather than an item characteristic as is the case in the fixed effects approach (Wang & Wilson, 2005b).

Testlet Response Model. The Testlet Response Model (TRM) adds a random effect to the logit of either the 2PL (Bradlow, Wainer, & Wang, 1999) or the 3PL model (Wainer,

Bradlow, & Du, 2000; Wainer & Wang, 2000) to account for the interaction of person j with testlet $d(i)$, the testlet that contains item i . The testlet model approach is essentially a special case of the multidimensional IRT model since an additional latent trait is added to the model for every random effect component (Zhang, 2010). The testlet response model based on the 3PL is written as

$$P_{ji} = c_i + (1 - c_i) \frac{e^{[a_i(\theta_j - b_i + \gamma_{d(i)j})]}}{1 + e^{[a_i(\theta_j - b_i + \gamma_{d(i)j})]}}. \quad (15)$$

The testlet parameter $\gamma_{d(i)j}$ is a random effect and its sum over examinees within any testlet, is equal to zero (Wainer & Wang, 2000). The random effect in this model is driven by its variance such that if its variance is zero there is no excess local dependence within that testlet meaning that the items in the testlet are conditionally independent. As the variance of the effect increases so too does the amount of local dependence (Wainer & Wang, 2000). The variances of the random effects in the 2PL are assumed to be constant across testlets. Thus, while the testlet model based on the 3PL may account for additional variance across testlets produced by guessing behavior, “parameter estimation of this model requires a more computationally intensive procedure that samples over the full grid of parameter values” (Wainer & Wang, 2000, p. 206).

A simplified testlet model was developed by Wang and Wilson (2005a) known as the Rasch Testlet Model which can be written as

$$P_{ji} = \frac{e^{(\theta_j - b_i + \gamma_{d(i)j})}}{1 + e^{(\theta_j - b_i + \gamma_{d(i)j})}}, \quad (16)$$

where P_{ji} is the probability that examinee j correctly responds to item i and $\gamma_{d(i)j}$ is a random effect for the interaction between person j and testlet $d(i)$ (Wilson & Wang, 2005a). If there

are no testlet effects ($\gamma_{d(i)j} = 0$), equation (16) reduces to the dichotomous Rasch model (Wilson & Adams, 2005).

Compared to the polytomous model approach discussed earlier, the testlet model approach has at least one major advantage in that the units of analysis are items rather than testlets; thus, the information in the response patterns is not lost (Wang & Wilson, 2005a). However, as testlet models require the addition of more parameters (one ability parameter for each bundle as well as an ability parameter for the test as a whole), these models can become quite complex which can increase time and efficiency in the estimation process (Zhang et al, 2010). Therefore, potential benefits of using testlet models should be weighed against the added complexity in data analysis (Zhang et al, 2010).

Item & Test Information. According to Wainer, Bradlow & Du (2000), IIFs in the context of the testlet response model can be defined as,

$$I(\theta_i) = a_j^2 \left(\frac{e^{t_{ij}}}{1 + e^{t_{ij}}} \right)^2 \left(\frac{1 - c_j}{c + e^{t_{ij}}} \right), \quad (17)$$

where $t_{ij} = a_j(\theta_i - b_j - \gamma_{id(j)})$ for the testlet $d(j)$ that includes item j . Thus, the addition of the testlet parameter, γ , relocates the mode of the IIF (Wainer & Wang, 2000). Repeating this for all items within a testlet would produce the testlet information function and for all items within a test for the TIF.

Modeling Assistsments Data

The Assistsments data that are the focus of this project consist of dichotomous responses to main items, dichotomous responses to scaffold items which assess the sub-skills that are required for correctly responding to the main item, dichotomous responses to the repeated presentation of the main item (this is always the last item presented in the scaffolding

process and simply restates the original item that was presented prior to the scaffolding), and a count of the number of hints that the student requested for each item. Due to the common content as well as the nature of the presentation of the items that is created from the scaffolding process, local dependency naturally exists in this dataset. Following Wainer & Kiely's (1987) outline for analyses in the presence of possible item dependency, data should first be fit to any basic dichotomous IRT model, assuming conditional independence and ignoring data structure. If the model does not fit the data, create polytomously scored "super-items" and apply an ordered response model. To address potential shortcomings of this latter model, (i.e., the loss of information in the response patterns) fit the data to a testlet response model to account for the interaction between person j and testlet $d(i)$, where i is an item in testlet d (Wainer & Wang, 2000; Wang and Wilson, 2005a).

Model Comparison. While all models are wrong, some are useful (Box, 1979) and selecting a model can be viewed as approximating reality rather than identifying it (Burnham & Anderson, 2003). The purpose of checking and comparing measurement models is to determine which model is *most* appropriate to use for a given dataset. For example, a particular performance-based assessment that measures students' math ability across two content sub-domains may be adequately modeled by a simple unidimensional polytomous IRT model as well as a more complicated 2-dimensional polytomous model. However, in order to know if the more simplistic unidimensional model is good enough or if a multidimensional model is necessary, model comparison techniques need to be employed. All things equal, more complex measurement models (i.e., models that incorporate more parameters) are intended to account for more of the variation in observed responses and thus are intended to provide a more accurate representation of the trait of interest. On the other

hand, adding parameters to model means less data or information for each parameter which also increases computation time. Thus, model complexity is not always desired. In any case, there are several factors such as model simplicity, accessibility and cost-effectiveness that need to be considered in the decision to either retain or reject a particular model. Thus, the goal of model selection is to identify the most parsimonious model that remains consistent with the purpose of the study and adequately accounts for the essential features of the dataset (Pitt, Kim & Myung, 2003).

In the present study, evaluations of models are based on several criteria; however, these practical constraints are also considered in the discussion section of this paper. Methods for model comparison are discussed in the following chapter. Also keep in mind that these models are evaluated in the context of a specific assessment system that employs a specific method for scaffolding. There are other scaffolding mechanisms that could be applied to an assessment system (e.g., hints only or adaptive content) that may fit a different measurement model better.

Summary

The concept of scaffolding has been applied to many different types of educational contexts with the goal of assisting students achieve their learning goals. One relatively recent application of scaffolding has been within the framework of formative assessment. Through the use of technology, formative assessments have the potential to accomplish the dual goals of collecting information with respect to students' knowledge, skills and abilities while providing students with instructional scaffolding on concepts that they struggle with. Allowing students the opportunity to demonstrate what they know (and don't know) after receiving additional scaffolded assistance, has the benefit of providing teachers with a more

detailed, fine-grained analysis of students' abilities. This information can then be used to help guide subsequent teaching and learning activities that are geared towards specific areas of need for an individual student or a group of students.

While the field of technology has made significant advancements towards realizing the potential of TEAs, there has not been as much recent research in the psychometrics field to accompany these advancements. One such area that has not been explored is the analysis of potential scoring paradigms that can be used to provide valid inferences about students' performance on scaffolded assessments. This research focuses on one type of scaffolding that is available within a formative assessment framework. Several IRT based scoring models are presented and evaluated including the dichotomous 2PL model, the polytomous ordinal response model, and a testlet response model.

Chapter Three – Methods

The purpose of this research is to help advance the development and use of TEAs, specifically those that incorporate scaffolding into the assessment process by comparing measurement models for an assessment that directly account for the scaffolding process. In an effort to identify the optimal scoring model for the scaffolded data provided by the Assistment system, this research evaluates four types of scoring models: one that is considered the baseline model, and three additional comparison models. These models are evaluated against each other with respect to several different measures of model adequacy. Criterion related validity evidence is also presented for the various models to evaluate the relationships between model estimates and an external measure of student ability. This chapter describes the participants, instruments, software, model fit and parameter estimation procedures, and statistical analyses.

Participants

The participants are a sample of 7th and 8th grade students in mathematics courses from an east coast state that were administered Assistments during the 2005-06 school year. The Assistments data provided for this research were for 5,910 students that were administered at least one Assistment item. While demographic information was not provided for this original dataset, an additional 778 student profiles were provided that had end-of-year state assessment scores of which also had demographic information. It was presumed that these additional students were representative of the original dataset and their demographic characteristics are provided. Of these 778 students, 51.0% were female, 59.4% received a free or reduced lunch, and 12.3% received special education services. With respect to students' race or ethnicity group, 53.0% were white, 27.1% were Hispanic, 12.3% were

African American, 7.1% were Asian or Pacific Islander, and 0.6% were Native American.

The data contained no identifiable information and this research was approved by the appropriate human subjects committee.

Assistments Data

Student responses on Assistments items were gathered by student profiles or assignments such that not every student was administered the same set of items. In other words, students completed assignments that comprised a given number of Assistment items which differed across assignments. Thus, while many items were administered to a large sample of students other items were administered to a relatively small number of students. Student profiles were associated with anywhere between two and 789 main and scaffold items; however, the distribution was right skewed with a median of 23 items and the first and third quartiles of 11 and 72 items, respectively. The mean number of items was 58. While concern is warranted for the student profiles that were associated with numbers of items greater than say, 250, the sample size criteria discussed in a subsequent section indirectly addressed this issue. The series of data cleaning procedures are discussed next and the treatment of missing data is presented in the following section.

Data Cleaning Procedures. The data were restructured and cleaned for the purposes of subsequent analyses using Fortran (Silverfrost Ltd, 2007) programming. Item bundles (main items and corresponding scaffold items) were initially analyzed for small sample sizes and bundles for which the main item was administered to less than 200 students were removed. Items were also removed if they did not clearly belong to a set of items based on the item identification information; i.e., an original item without any associated scaffold items or vice versa. The number of items provided in the original dataset was 2,914 which was

pared down to 1,122 items (261 main items and 861 scaffold items). The number of scaffold items associated with a particular main item ranged from one to 15 items; however, as shown in Table 1 below, most item bundles contained between three and six items. Student profiles that only contained data for the items that were deleted were also removed resulting in 5,083 profiles.

Table 1

Frequency of Number of Items per Bundle

Bundle Size	Frequency
2	11
3	72
4	85
5	56
6	20
7	8
8	4
9	3
13	1
16	1
Total	261

A more stringent sample size criteria was set to address potential estimation issues associated with small sample sizes in the context of complex models as well as issues associated with the number of bundles (and thus dimensions) estimated in the testlet models. The majority of research that has been conducted in the area of testlet models utilized samples sizes that were greater than 500 examinees and more than half of the studies identified in this area of research had sample sizes greater than 1,000 (Zhang et al, 2010). Moreover, Reise and Yu (1990) recommended sample sizes of at least 500 to achieve adequate calibration of parameters when using polytomous models such as the GRM. Since two of the models evaluated in this study require relatively large samples sizes to achieve adequate calibration,

bundles in the original dataset that were administered to less than 450 examinees were removed for a total of 36 bundles which is equal to a total of 159 items. Table 2 below displays the number of bundles associated with various sample size categories. It should also be noted that while the research on the stability of parameter estimation using polytomous models focuses mostly on sample size, the number of items calibrated is often less than 30 (see for example, Resise & Yu, 1990; Ankenman & Stone, 1992). Therefore, reducing the number of bundles (i.e., super-polytomous items in the polytomous comparison model) to 36 should not negatively affect calibration of a polytomous model.

Table 2

Number of Bundles Associated with each Sample Size Category and the Total Number of Bundles if the Category was Removed.

Sample Size	Number of Bundles Affected	Number of Items Affected	Total Number of Bundles if Removed	Total Number of Items if Removed
200-300	2	7	259	1115
301-350	84	364	175	751
351-400	118	502	57	249
401-450	21	90	36	159
451-500	6	30	30	129
501-1000	9	33	21	96
1001-1500	14	67	7	29
1501-2000	7	29	0	0
Total	261	1122		

Reducing the total number of bundles to a more manageable size is also necessary for practical limitations associated with estimating complex models in virtually any software program. While some programs theoretically have the potential to estimate large complex models (e.g., WinBUGS, Spiegelhalter, Thomas, Best, & Lunn, 2003), the time it would take to do so may not be of any practical value. Other programs that can be used to estimate the

testlet models often have a limit to the number of dimensions that a model can have (e.g., ConQuest, Wu, Adams & Wilson, 1998). Therefore, the number of bundles in the present study needed to be significantly reduced and setting the sample size criteria to a sufficiently large number addressed both estimation issues.

Furthermore, based on the characteristics of the entire dataset, the average number of bundles (i.e., the average number of main items with scaffolding) that a student was administered was 23.38; the median was 14 and the mode number of bundles was 20. Therefore, within an operational context, the majority of students' scores would be based on approximately 23 item bundles. Thus, reducing the dataset to 36 bundles still represents the majority of students.

As a preliminary step in the model analyses procedures, (described in a subsequent section), data were calibrated using IRT software. This analysis indicated that several items could not be calibrated due to lack of variance. In total, four bundles were affected by having at least one item with zero variance (responses to the items were either invariably correct or incorrect). These items (and the bundles they were associated with) were removed for the final model fit analyses for a final total of 140 items; 32 main items and 108 scaffold items. Table 3 below displays the frequencies of bundle sizes for the set of items.

Table 3

Frequency and Number of Items by Bundle Size.

Bundle Size	Frequency of Bundle Size	Total Number of Items
2	1	2
3	6	18
4	13	52
5	8	40
6	2	12
7	1	7
9	1	9
Total	32	140

As the matrix of items for students was reduced, the numbers of items per student were also reduced. Student profiles that had response data for fewer than five main items were also removed for a total of 2,745 profiles. The number of total items administered to students in the final dataset (140 items; 32 bundles) ranged from 15 to 127 items; however, the number of main items administered to students ranged from five to 29. The mean number of total items administered to students was 49 and the median was 41 items; the mean number of main items was 11.2 and the median was 10. Figure 9 below displays the frequency distribution of the number of main items administered to students. This distribution is clearly bimodal with most students taking either five or 18 main items (or bundles of items). The majority of the remaining students were administered a number of main items somewhere between these two modes.

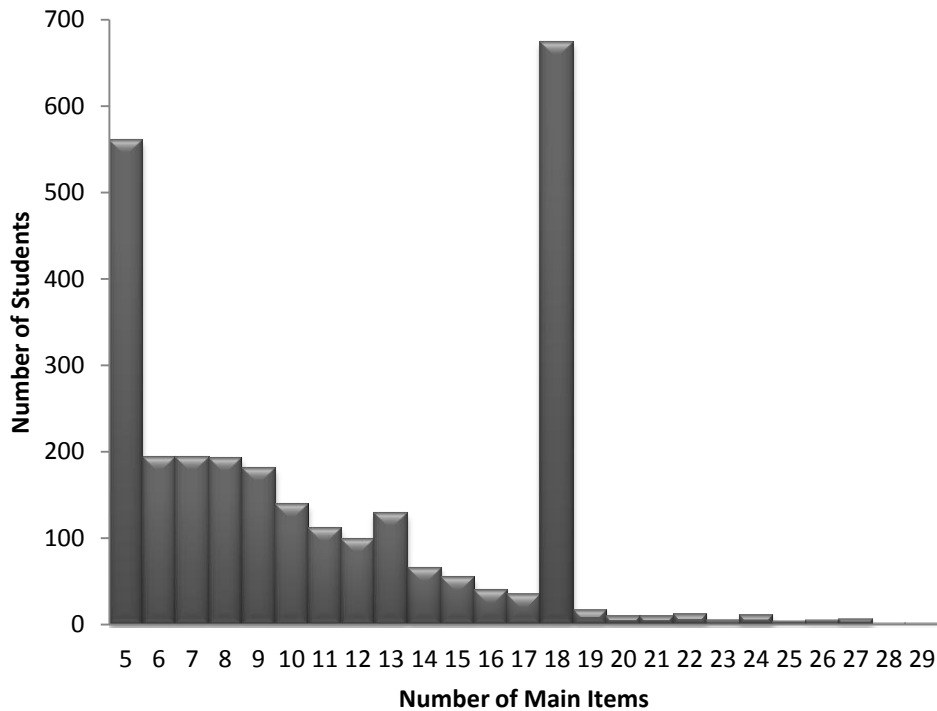


Figure 9. Frequency distribution of the number of main items administered to students.

Again, as the number of students was reduced, the sample sizes associated with each item were also reduced. In an effort to retain data, no additional items were deleted. The smallest sample size associated with a main item in the final dataset was 301. While this was lower than the original criterion of 450, deleting additional items to retain this criterion would have further contributed to fewer items per student and the matrix of data would have continued to shrink. Thus, a sample size criterion of 300 or greater was deemed sufficient for the purposes of this study.

In summary, as part of the data cleaning process, criteria were originally set for the removal of items and cases that were associated with an inadequate amount of data in accordance with previous research and preliminary item analyses. Based on these criteria a dataset of 140 items (32 main items and 108 scaffold items) was derived. From there, student

profiles were evaluated and cases in which students were administered fewer than five main items were removed. Thus, the final dataset was based on 32 bundles of items and 2,745 student profiles.

Missing Data. Since none of the students were administered all of the items and scaffold items were only presented when the main item was answered correctly, the dataset contained a large portion of missing data. Table 4 below displays the proportions of missing data for each of the 32 main items. There did not appear to be any main items that were administered significantly more or less than the others to the sample of students in this study.

Table 4.

Number and Proportions of Missing Cases for each Main Item.

Main Item	Number of Missing Cases	Proportion of Missing Cases	Main Item	Number of Missing Cases	Proportion of Missing Cases
1	2551	0.69	17	1998	0.54
2	2453	0.66	18	2360	0.64
3	2662	0.72	19	2464	0.67
4	2572	0.70	20	2501	0.68
5	2600	0.70	21	2504	0.68
6	3239	0.88	22	2768	0.75
7	2574	0.70	23	3064	0.83
8	2624	0.71	24	2509	0.68
9	2659	0.72	25	3233	0.87
10	3227	0.87	26	3081	0.83
11	1997	0.54	27	2718	0.73
12	2020	0.55	28	3214	0.87
13	2640	0.71	29	2929	0.79
14	2018	0.55	30	2932	0.79
15	2026	0.55	31	2933	0.79
16	3232	0.87	32	2945	0.80

Ayers and Junker (2008) conducted IRT modeling on a similar Assistments dataset and treated the missing data as completely at random (MCAR) due to the fact that “...problems were assigned to students randomly by the Assistment software from a ‘curriculum’ of possible questions designed for all students by their teachers in collaboration with project investigators” (Ayers & Junker, 2008, p. 976). However, since the missing values on the scaffold items was a systematic function of the response to the main item (i.e., students that answered the main item correctly were not presented with any of the scaffold items), it was assumed that correct answers would have been provided to the scaffold items had they been presented to those who responded correctly to the main item. This is a reasonable assumption given that the scaffold items simply represent the sub-skills needed to understand the main item. Thus, missing values on scaffold items were assigned a 1 (correct response) if the response to the corresponding main item was correct.

Software

The software programs used in this study included SPSS (IBM: Version 19.0), Fortran (Silverfrost Ltd, 2007), BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), SCORIGHT 3.0 (Wang, Bradlow & Wainer, 2005), and WinBUGS, (Spiegelhalter, Thomas, Best, & Lunn, 2003). The first program was used to conduct the statistical analyses in the second research question; the second program was used to write programs for data cleaning purposes (as previously mentioned) and for model fit calculations; and the last three programs were used to fit the data to the various scoring models. The main model fitting analyses were performed using SCORIGHT 3.0; however, a brief description of BILOG-MG is also provided.

BILOG-MG is used to fit unidimensional dichotomous data only and is typically employed to estimate the 1-, 2-, and 3PL models. BILOG-MG estimates item parameters using marginal maximum likelihood (MML) estimation. In short, this method estimates item parameters while integrating out the ability parameters. Once the item parameters are known, the ability parameters are estimated with either Newton-Raphson Maximum likelihood (ML) techniques, expected a posteriori (EAP) techniques, or maximum a posteriori (MAP) (Hambleton, Swaminathan & Rogers, 1991). MML estimation procedures uses two methods for solving the marginal likelihood equations: expectation maximization (EM) and Newton-Gauss iterations (Zimowski, Muraki, Mislevy, & Bock, 1996).

SCORIGHT 3.0 software is designed to facilitate analysis of item response data that may contain testlets (Wang, Bradlow & Wainer, 2005). The program is capable of handling both dichotomous and polytomous data that are either independent or nested within bundles. The model used for binary data is the 3PL model which can be adjusted to the 2PL as well. The model used for polytomous data is the ORM. SCORIGHT 3.0 employs Bayesian estimation techniques to estimate model parameters. In short, Bayesian methods involve modifying the likelihood function to incorporate any prior information that is known about model parameters. In SCORIGHT 3.0, inferences for unknown parameters are obtained by drawing samples from their posterior distributions using Markov Chain Monte Carlo (MCMC) techniques (Wang, Bradlow & Wainer, 2005). Further details on estimation procedures are provided in a following section. There are not any limitations explicitly mentioned in the SCORIGHT 3.0 user's manual on the number of dimensions (therefore, bundles) that can be incorporated into the model (Wang, Bradlow & Wainer, 2005). WinBUGS is similar to SCORIGHT 3.0 in that it can estimate statistical models, including

IRT models, using Bayesian analyses but it has the added flexibility of altering existing code to fit variations of models and using different prior information (Curtis, 2010).

Procedures

Research Question 1: What type of model is the optimal scoring model for the scaffolded data in the Assistment system? The first research question is related to identifying the most appropriate model for the Assistments data. To address this question, a sequential procedure for fitting and evaluating increasingly complex models is outlined. A baseline model was established and compared to three additional comparison models such that the former did not account for any of the scaffolding features or complexities in the dataset whereas the latter group of models did, each in a different way. Specifically, the baseline model only accounted for the independent dichotomous responses to the main items and does not account for responses to scaffold items, the grouping effect created by the scaffolding process, or the number of hints accessed by the student. This model served as the baseline for comparison purposes with subsequent models that accounted for these scaffolding features.

The comparison group of models accounted for the scaffolding features first by simply including responses to the scaffold items in the model and then by evaluating the items as bundles and accounting for the dependency that exists within each of these bundles of items. Two different methods that account for local dependence between items were examined. Within all of the comparison models, the number of hints a student accessed was also evaluated to determine if doing so would improve model convergence, model fit and/or model estimates. This was accomplished by running each of the models twice; once with the covariates and once without. Model evaluations were based on several factors which

included: model convergence measures such as statistical detection of convergence and the time it takes a model to converge, model estimates, a model fit statistic, and test information provided by each model. Measurement procedures for each of these factors are described in detail in later sections. However, prior to establishing any of the evaluation models (i.e., the baseline model and the comparison models), the number of appropriate parameters to be incorporated into all models needed to be determined. Specifically, the 1PL and the 2PL models were fit to the data to determine the number of parameters to include in the evaluation models.

The 1PL or the 2PL Model?

To make sensible comparisons between the various measurement models (e.g., dichotomous, polytomous and testlet models) any parameters incorporated in the baseline model (e.g., a discrimination parameter) also need to be added to each of the comparison models. In other words, for comparison purposes, the models needed to incorporate the same number of parameters. Thus, the first step that was needed, prior to establishing the baseline model or comparison models, was to determine the number of parameters that would be included in each of the evaluation models. That is, would all models be based on the 1PL model or the 2PL model?

In order to determine the number of parameters to be estimated in each of the evaluation models, responses were calibrated in BILOG-MG using each model. ML estimation procedures were employed with the theta (θ) scale set at 0,1 (default settings). The number of cycles for the EM algorithm was set at 10 and the number of Newton steps was set at 2 (default settings). To facilitate estimation for both models, item parameter estimates obtained from the initial 1PL were provided as starting values for estimating item parameters

in the second 1PL model as well as the 2PL model. These two models were compared with respect to relative fit to the data and the model that fits the data better is used as the basis for subsequent models. That is to say, if the 2PL model which accounts for the additional discrimination parameter was found to fit the data better, then the discrimination parameter will be estimated in all subsequent models. On the other hand, if the 1PL model fit the data better without the additional discrimination parameter, then it will *not* be estimated in subsequent models. Because the results of these model fit analyses determine the methods to be used for all subsequent analyses (i.e., types of models, software and model fit indices), it is imperative to evaluate these results prior to explicating further procedures.

The following model fit procedures were based on the sample size criteria of 450 or greater which was equivalent to 159 items total (36 main items and 123 scaffold items). However, as mentioned previously, initial analyses indicated that several items could not be calibrated due to lack of variance. These items (and the bundles they were associated with) were removed for the final model fit analyses.

Estimates obtained from the 1PL and 2PL models were evaluated with respect to overall model fit as measured by the change in log likelihood estimates, item-by-item fit statistics and item residual information. In general, the 2PL model appeared to fit the data better than the 1PL model. Allowing the slopes to vary in the 2PL model produced a statistically significant decrease in the overall misfit as indicated by the -2 log likelihood difference for the models ($\Delta\chi^2 = 4548.803$, $df = 139$, $p < 0.00001$). Based on item fit statistics provided from each model, the 2PL model fit the data better for 87 out of the 140 items as shown in Appendix A. Finally, standardized residuals were calculated from the raw residuals for the 32 main items to assess the accuracy of model predictions against the actual data. The

standardized residual is the difference between observed proportions correct on a given item and the predicted probability estimated by the model. This difference is then divided by the standard error of the expected proportion correct. Overall, standardized residuals were much smaller for the 2PL model than the 1PL model as shown in Figures 10 and 11 below. Figure 12 displays the frequencies of standardized residuals for all items in both models.

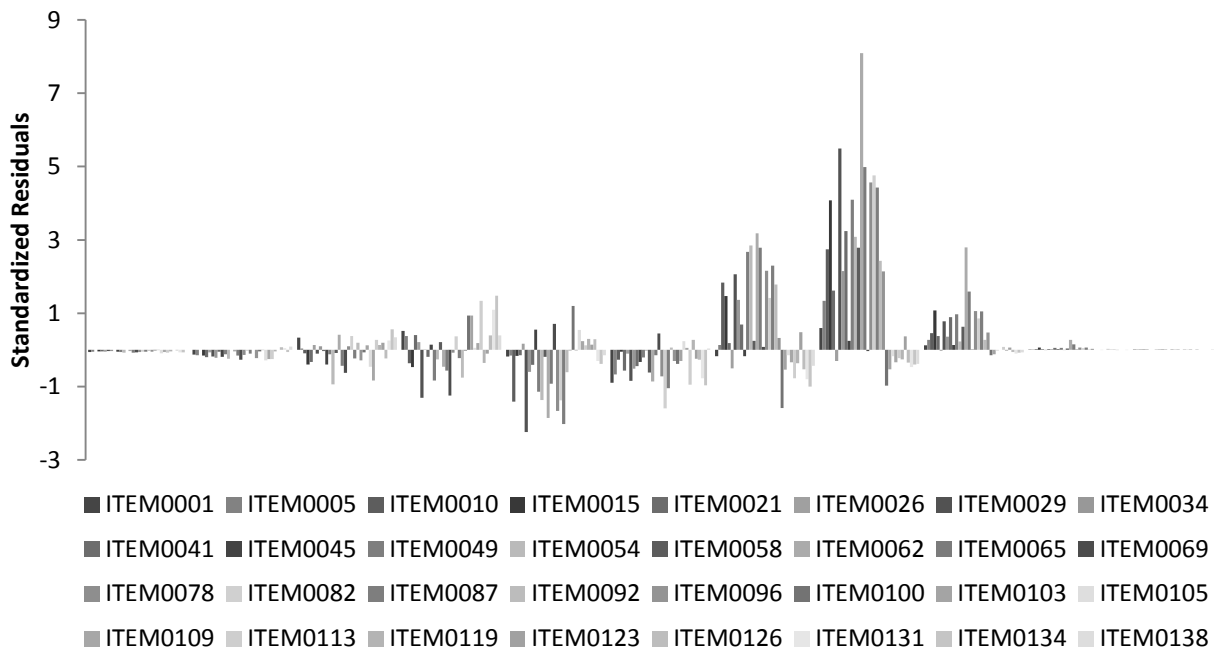


Figure 10. Standardized residuals for each of the 32 main items estimated with the 1PL model

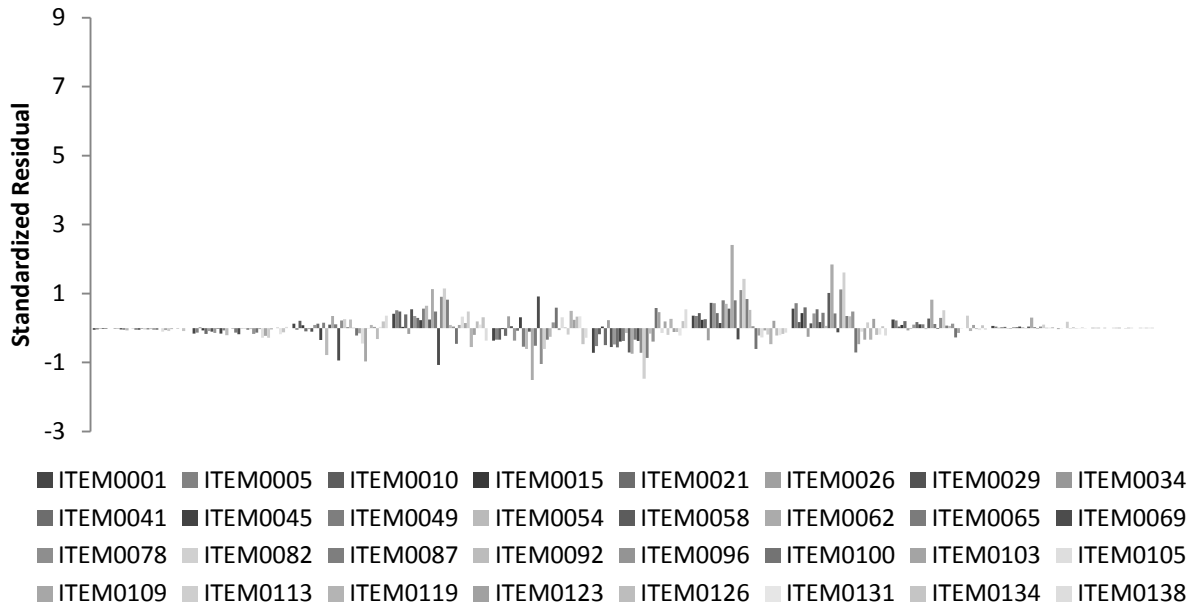


Figure 11. Standardized residuals for each of the 32 main items estimated with the 2PL model

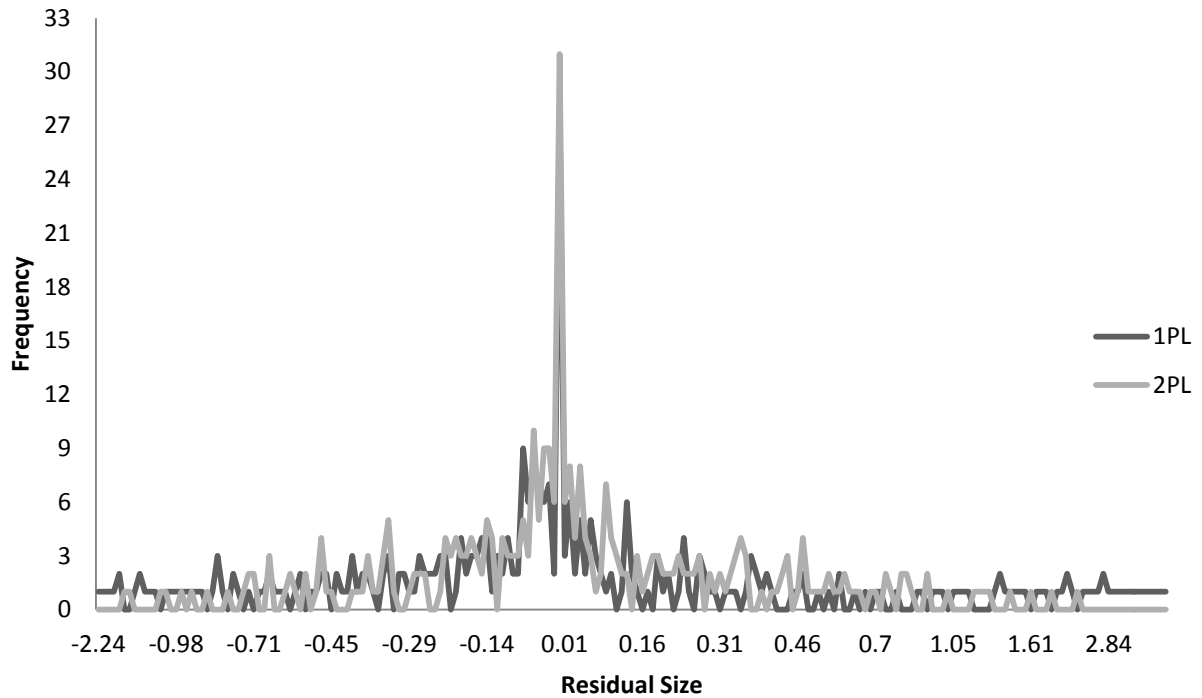


Figure 12. Frequencies of standardized residuals for all 32 items in the 1PL and 2PL models

Based on these results, the discrimination parameter was included in all of the following model fit analyses as it appears to more adequately fit the data than the more restrictive 1PL model. This decision is also consistent with the current scoring methods of the MCAS (which are the basis for most Assistments items). That is, the MCAS multiple-choice questions are scaled with the 3PL model whereas the short-answer (fill in the blank) questions are scaled with the 3PL model whereas the short-answer (fill in the blank) questions are scaled using the 2PL model (Massachusetts Department of Education, 2004). The 3PL model was not considered in the present study as the guessing parameter for this assessment is presumably very low due to the constructed response nature of many of the test items and the scaffolding features which guide students to the correct answer. Therefore, the 2PL model was the basis for all evaluation models. A detailed discussion of each of these evaluation models is provided in the following sections and an outline of the model comparison steps for this study is displayed in Table 5 below.

Table 5.

Outline of Model Evaluation Procedures

Type	Description	Purpose	Model(s)
<u>Baseline Model:</u> Dichotomous Model for Main Items	Main items only; ignores all scaffolding features and data	Used as baseline for comparison purposes	2PL
<u>Comparison Models:</u> Dichotomous Model for All Items	Main and scaffold items included but ignores bundle structure and hints; hints used as a covariate	Most simplistic model that accounts for all item responses	2PL with & without covariates
Polytomous Model	Bundle treated as a super-item polytomously scored; hints used as a covariate	Account for dependencies using a fairly complex model but loses some information	ORM with & without covariates
Testlet Response Model	Random effect added for person/bundle interaction; hints used as a covariate	Account for dependencies using a complex model but retains information	TRM with & without covariates

Baseline Model. The baseline, referred to as the 2PL_MainItems model hereafter, signifies a simplistic representation of the Assistments data in that it models the responses to the main items only without accounting for the scaffold items, the bundle structure that is inherent within the scaffolding process or the number of hints a student accessed for a given item. The response data for the main items only, was calibrated to the unidimensional dichotomous 2PL model. Estimation procedures for this model are discussed in a later section.

Comparison Models. The comparison group of models extends the 2PL_MainItems model to assess the scaffolding features of the data first from a local independence assumption and second from a local dependence assumption. The first model, referred to hereafter as the 2PL_AllItems model, simply extends the 2PL_MainItems model to account for the additional scaffold items. This model assumes local independence between items and ignores the item grouping that occurs as a result of the scaffolding process.

The next two comparison models address the issue of local dependence using two different methods; both assume local independence between bundles but account for local dependence within bundles. Specifically, the second comparison model, denoted as the ORM, accounted for local item dependence by treating the response patterns of item bundles as categories of a polytomous item. Summed scores were obtained for each item bundle in the Assistments data which were then treated as single super-items that were scored polytomously using the ORM. The ORM is the model used in the SCORIGHT 3.0 software program. The third model, known as the Testlet Response Model or TRM, also accounted for item dependency within bundles but did so by adding a random effect component to explain the interaction between the person and the bundle; i.e., a bundle ability parameter.

All three of the comparison models (i.e., the 2PL_AllItems models, the ORMs and the TRMs) were evaluated twice; once with covariates and once without. The average number of hints accessed was used as a covariate for both person and item parameters. That is, an average number of hints for a given student (relative to the number of items the student was administered) was used as a covariate for estimating student ability. An average number of hints accessed for a given item (relative to the number of students it was administered to) was used as a covariate for estimating item difficulty. The comparison models were calibrated both with and without these covariates in an effort to evaluate the value of this data in the estimation process.

Another approach for incorporating the number of hints into the scoring models could have been to employ an item bundle model that allows for both dichotomous and polytomous items and assigning partial credit to items based on the number of hints needed to answer an item correctly (see for example, Wang & Wilson, 2005a). However, as the current Assistentment scoring system automatically assigns a 0 to any scaffold item in which hints were accessed, it was not possible to know when a student actually responded correctly to an item after receiving hints versus when a student received hints but still responded incorrectly. Thus, assigning partial credit based on the number of hints needed to correctly respond to an item was not possible for the current project.

Parameter Estimation. In order to obtain IRT item parameters for all of the evaluation models, the response data from the items was calibrated using SCORIGHT 3.0 which employs Bayesian estimation techniques. Since SCORIGHT 3.0 is a general program that can facilitate data that is composed of dichotomous or polytomously scored items that are

independent or nested within testlets (Wang, Bradlow & Wainer, 2005), it was used to estimate all of the evaluation models in this study.

Overview of Bayesian Inference. In order to combine information across examinees, items and any potential testlets, SCORIGHT 3.0 embeds a hierarchical Bayesian framework into the model which allows for more precise estimates (Bradlow et al., 1998; Wang, Bradlow & Wainer, 2005). While a full synopsis is beyond the scope of this paper, in general, Bayesian inference rests on Bayes' theorem which states that a representation of the conditional probability of one event given another in terms of the opposite conditional probability (Kim & Bolt, 2007).

In IRT, information about item and person parameters is reflected in the relative likelihoods of these particular parameter values given the observed item response data. The type of IRT model employed (e.g., the 2PL) provides a basis for describing the opposite conditional probability (i.e., the probability of observing the item response data given the model parameters) (Kim & Bolt, 2007). In IRT applications, Bayes' theorem can be written in the form of continuous probability density functions (e.g., the normal density function) which represent the relative likelihood of each outcome. What is referred to as the joint posterior density is used to derive estimates of the model parameters. In order to evaluate the joint posterior density, the particular item response model is needed (e.g., the 2PL) as well as knowledge about the prior density of the parameters which represents information about the relative likelihoods of parameter values prior to data collection (Kim & Bolt, 2007). However, even with this information, the exact density of the posterior density is typically unknown and difficult to determine; therefore, MCMC sampling procedures are used to theoretically reproduce the density by sampling observations with respect to it (for a detailed

description of these procedures, see Spiegelhalter, Thomas, Best & Gilks, 1995) (Kim & Bolt, 2007). Based on many samples, characteristics of the density can be determined and used as the basis for model parameter estimates. While there are several different sampling procedures within MCMC (e.g., Gibbs sampler, Metropolis Hastings), all require specification of priors as the prior densities are needed to define the posterior densities (Kim & Bolt, 2007). Choosing a prior depends on several factors including the type of distribution of the posterior density and the type of model chosen, as well as the desired strength or influence of the priors on the posterior density.

Bayesian Framework in SCORIGHT 3.0. Within the SCORIGHT 3.0 Bayesian framework, prior distributions are asserted for each of the corresponding parameters. These parameters are assumed to be normally distributed as follows,

$$\theta_i \sim N(0,1) \tag{18}$$

$$h_j \sim N(\mu_a, \sigma^2_a) \tag{19}$$

$$b_j \sim N(\mu_b, \sigma^2_b) \tag{20}$$

$$\gamma_{id(j)} \sim N(0, \sigma^2_\gamma) \tag{21}$$

where h_j is equal to the $\log(a_j)$ (Bradlow et al., 1998; Wang, Bradlow & Wainer, 2005).

Among these four random effects distributions, two of the means are set to zero and one of the variance components is set to one in order to identify the model (Bradlow et al., 1998).

Furthermore, in SCORIGHT 3.0 covariates can be incorporated into the model via the mean of the prior distribution of the item parameters and the ability parameters. For computation of the posterior density function, SCORIGHT 3.0 utilizes MCMC techniques; specifically, it employs a combination of the data augmented Gibbs sampler (Tanner & Wong, 1987) and a

Metropolis-Hastings step (Hastings, 1970). For a detailed discussion of posterior computation procedures in this software program see Wang, Bradlow & Wainer (2002).

Specifying Models in SCORIGHT 3.0. As mentioned, all evaluation models were estimated using SCORIGHT 3.0. The baseline model and the dichotomous comparison model were both based on the dichotomous 2PL model; the polytomous comparison model was based on the ORM; and the testlet comparison model was based on the 2PL TRM. Each of these models was specified and convergence was assessed in the SCORIGHT 3.0 program. For each model, the following SCORIGHT 3.0 specifications and procedures were employed:

1. The number of Markov chains to be run was set to three in order to facilitate detection of any potential convergence issues.
2. The number of iterations was set at 10,000 which is a value that has previously recommended for MCMC estimation procedures (Sinharay, 2004). However, if a model did not converge with 10,000 iterations, this number was increased until convergence was reached.
3. The number of initial draws to be discarded was set to 5,000 in order to decrease the likelihood that parameter estimates would be based on draws that were sampled prior to model convergence (Wang, Bradlow & Wainer, 2005). Again, this number changed as the number of iterations increased.
4. The number of times the posterior draws were recorded was set to 10 to decrease the likelihood that the posterior draws kept will be autocorrelated (Wang, Bradlow & Wainer, 2005).
5. SCORIGHT 3.0 was instructed to automatically select starting values for the initial parameter estimates. However, based on convergence results from

initial model runs, parameter estimates from previous models that successfully converged may be used as starting values for subsequent model calibrations if needed.

6. *The average number of hints a student accessed were incorporated as a covariate for estimating person parameters, θ .
7. *The average number of hints accessed for a given item were incorporated as a covariate for estimating item difficulty, b_i . In the polytomous model, the average number of hints for each item within the bundle were averaged across all items within the bundle.

*Covariates were not included in the 2PL_MainItems model and each of the comparison models was calibrated both with and without the covariates.

Model Evaluation. A sequential procedure for assessing each model was conducted. Each of the comparison models was compared to the baseline model to determine if accounting for the scaffolding features in the Assistments system provided a better model. Comparison models were also evaluated against each other to determine if the additional complexities provided better models for the data. Each comparison model was also evaluated with respect to the utility of the average number of hints as a covariate. Models were mainly assessed according to model convergence and test information. Comparisons were also made between models that used covariates and those that did not using a model fit statistic. Model fit statistics and parameter estimates are provided for all models as descriptive measures. Figure 13 provides a visual display of all model comparisons. This process ensured that all evaluation models were assessed relative to all other potential models evaluated in this study.

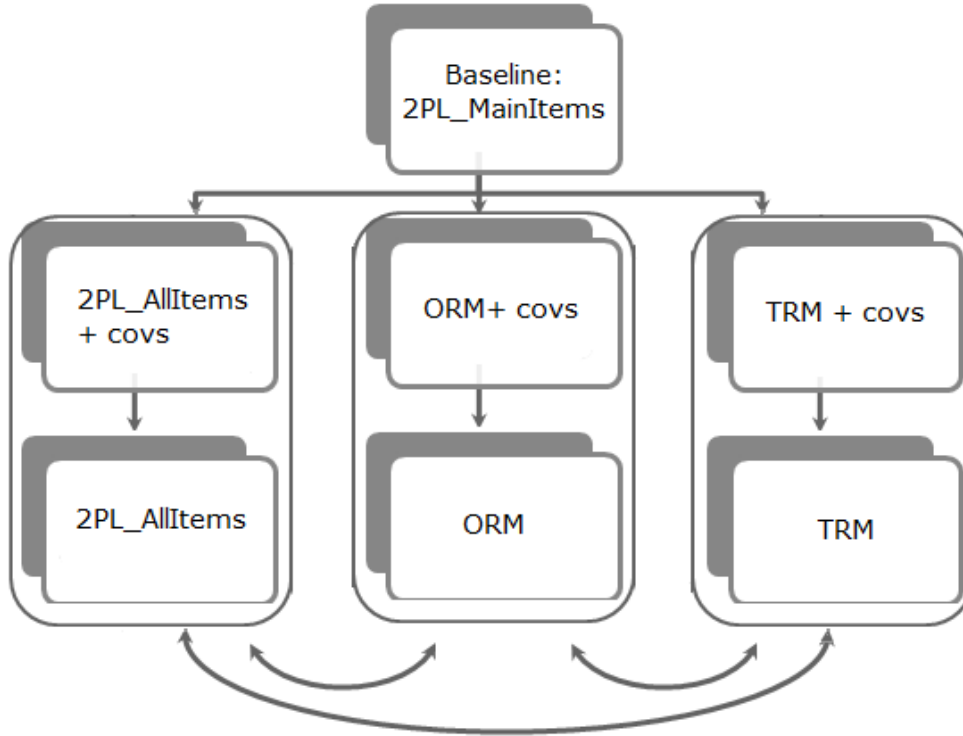


Figure 13. Model comparison flowchart

Bayesian Convergence. As suggested by Wang et al. (2005), convergence was assessed by evaluating the similarity of output across the chains for each model; if convergence was achieved estimates should be approximately the same for each chain. To statistically assess the similarity of output across chains, the F -test convergence criterion of Gelman and Rubin (1993) was calculated for every parameter within each model. When multiple parallel chains are specified, SCORIGHT 3.0 automatically provides this index for each parameter in the model. The Gelman and Rubin (1993) diagnostic, often referred to as the potential scale reduction factor (PSRF), is based on the last n iterations in each of m parallel chains that each ran for $2n$ iterations. The PSRF is then calculated as,

$$PSRF = \sqrt{\frac{n-1}{n} + \frac{m+1}{mn} \left(\frac{B}{W} \right)}, \quad (22)$$

where B is the between-chain variance and W is the within-chain variance (Gelman & Rubin, 1993). As chains converge to a common distribution, the between-chain variability should become small relative to the within-chain variability and PSRF should be close to 1.0. Gelman and Rubin (1993) suggest that PSRF values less than 1.2 indicate reasonable convergence. Conversely, if PSRF is large, this suggests that either the between-chain or the within-chain estimates of variance can be further decreased by more simulations, or that further simulation will increase the within-chain variance in the case that the simulated sequences have not yet sampled from the entire target distribution (Gelman & Rubin, 1998). SCORIGHT 3.0 provides the PSRF for the 50% and 97.5% quantiles based on the Student t distribution (Wang, Bradlow & Wainer, 2005). It is recommended that the PSRFs at both quantiles be at or below 1.2 (Gelman & Rubin, 1993; Wang, Bradlow & Wainer, 2005).

Other convergence issues were noted and are described in the next chapter as a point of comparison with the other models. For example, it is of value to know which model converged the easiest (i.e., with fewer iterations and/or with fewer Markov chains). Similarly, approximate time for a model to converge was also tracked as a means to evaluate model efficiency. Once convergence was attained for all of the models, the statistical fit of each model was assessed.

Model Fit. When researchers are interested in finding the best model that fit a particular dataset, in the context of several possible models, model comparison techniques can be conducted. There are several different indices that are useful for model comparison purposes such as the Pearson χ^2 test, the likelihood ratio G^2 statistic, Akaike's Information Criterion (AIC; Akaike, 1974), Schwarz's Bayesian Information Criterion (BIC; Schwarz, 1978), Bayes Factors (BF; Kass & Raftery, 1995), and the Deviance Information Criterion

(DIC; Spiegelhalter, Best, Carlin & van der Linde, 2002). Among these, the Pearson χ^2 and the likelihood ratio G^2 statistics are only appropriate for comparing nested models; however, the other four criteria can be used to compare either nested or non-nested models (Zhu, 2009). The AIC and BIC are information-based criteria and are often used when ML estimates of model parameters are obtained (Zhu, 2009); although the BIC can also be employed in the context of MCMC. The DIC and BF are two specifically Bayesian criteria used for model comparisons when MCMC techniques are used to estimate model parameters (Zhu, 2009).

As the models in the present study are not nested and Bayesian estimation procedures are employed within SCORIGHT 3.0, the DIC was chosen for calculating the fit of each model. However, model fit statistics are based on the assumption that models are fit to the same exact dataset. While the data for this study are based on the same students, the inclusion of items differed for each model. The 2PL_AllItems models and TRMs are fit to 140 items whereas the 2PL_MainItems model is fit to 32 dichotomous items and the ORMs are fit to 32 polytomous items. Due to the differences in data structures, sensible comparisons using the DIC can only be made between the 2PL_AllItems models and the TRMs as well as between each model with covariates and without. For the sake of completeness and descriptive information, the DIC was calculated for all of the models that converged.

The DIC is similar to other commonly used fit statistics (e.g., the AIC and BIC) in that it considers the penalty on model complexity in identifying the preferred model (Spiegelhalter, Best, Carlin & van der Linde, 1998). The DIC is composed of two terms which represent model deviance and model complexity (Spiegelhalter et. al, 1998). The DIC is defined as

$$DIC = \bar{D}(\eta) + p_D \quad (23)$$

The first term, $\bar{D}(\eta)$, is a Bayesian measure of fit and is equal to the posterior mean of the deviance between the data and the model which is calculated as

$$\bar{D}(\eta) = E_{\theta|y}[-2 \ln f(y | \theta)] \quad (24)$$

where θ represents model parameters and y represents the data. The second term, p_D , is a measure of model complexity and is equal to the difference between the posterior mean of the deviance and the deviance at the posterior mean of the parameters which is defined as

$$p_D = \bar{D}(\eta) - D(\bar{\theta}) \quad (25)$$

where $D(\bar{\theta})$ is the deviance evaluated at the posterior mean, $\bar{\theta}$, of the parameters (Spiegelhalter et. al, 2002). The smaller the value of the DIC, the better the model fits the data.

As the SCORIGHT 3.0 does not automatically provide the DIC index, two programs were written by the author in Fortran to calculate this statistic from the MCMC output for the 2PL_MainItems model, the 2PL_AllItems models and the ORMs. The code for these programs is provided in the Appendices. While a third program was planned to calculate the DIC statistic for the TRM models, unfortunately, the output from these models did not follow a consistent, pre-defined format and could not be used to calculate the DIC. This issue is discussed in detail in the next chapter.

Information. The models were also evaluated with respect to their information functions across the theta scale. That is IIFs and TIFs were calculated for each of the evaluation models to assess the precision of θ estimates produced by each model.

When local dependencies exist and are not accounted for, it has been found that test reliability and test information tends to be overestimated (Wang & Wilson, 2005; Weng,

Cheng & Wilson, 2005; Zhang, et al, 2010). Therefore, it was expected that the dichotomous comparison model (which models all of the items but ignores dependencies) would have an inflated test reliability estimate. If either of the item bundle models was able to account for the local item dependence and provide an equivalent or better estimate than the inflated estimate, then it could be taken as supportive evidence in favor of such a model.

Research Question 2: Is there a relationship between student ability estimates derived from the scoring models and a criterion measure of student achievement? The second research question evaluates one aspect of the validity of interpreting scores from the models established in the first research question by determining the degree to which these scores relate to subsequent performance on the state's end-of-year accountability assessment. Criterion-related validity evidence, or external validity, refers to the extent to which test scores relate to other measures of the construct being assessed (Messick, 1995). In this sense, evidence for the validity of the construct being measured is supported when scores from the assessment can account for the pattern of relationships in the criterion measure. As one of the intended purposes of the Assistments system is to help students prepare for the end-of-year state assessments (Heffernan & Heffernan, 2008), it is of value to examine the statistical relationships between student performance on the Assistments and their subsequent performance on the state assessment. In this context, student ability estimates that have the strongest relationship with scores from the end-of-year accountability exam may be used as supportive evidence for the particular scoring model from which the estimates were derived.

As such, the scoring models were first evaluated against a simple percent correct score to determine if student ability estimates calibrated from IRT models more strongly relate to a criterion measure of student ability than the percent correct score. Second, statistical

relationships were also compared between the 2PL_MainItems model, which did not account for any scaffolding features and the comparison models, which did account for these features. Finally, statistical relationships were compared between the comparison models to determine if accounting for local dependence in the scoring model produced ability estimates that were more strongly correlated with a criterion measure than estimates derived from a model that ignored the data structure.

Analyses. Scaled scores for examinees from each of the evaluation models established in the first research question were correlated with end-of-year state assessment scores for a subsample of 778 students using SPSS 19.0. First, a percent correct score was calculated for each student in the subsample and this score was also correlated with students' state assessment scores. Next, scaled scores from the 2PL_MainItems model, the 2PL_AllItems models, the ORMs and the TRMs were each correlated with students' state assessment scores. Since student-level metrics were either calibrated directly into student ability estimates (e.g., responses to scaffold items) or were used to facilitate the estimation process (e.g., number of hints used as a covariate), no other variables were included in these models. As each of the models produced scaled scores that encompassed the metrics of interest in this study, deriving regression models with multiple predictors was not sought; rather simple correlation coefficients were calculated for each set of scores to evaluate the overall relationship between model estimates of student ability and student performance on the criterion measure. This made for a matrix of relationships between nine types of scores; a percent correct score, scaled scores from each of the seven scoring models, and students' state assessment scores.

Summary

Overall, seven scoring models were fit to the Assistments data using MCMC estimation techniques employed in SCORIGHT 3.0 in order to determine which model was optimal with respect to model convergence, model fit, and information. Each of the scoring models accounted for the data differently; the 2PL_MainItems model only calibrated data for the main items and did not account for any scaffolding features; the 2PL_AllItems models accounted for all of the items but ignored local dependence that is created by the scaffolding process; the ORMs accounted for local dependence applying a polytomous model to bundle summed scores; and the TRMs also accounted for local dependence but did so by adding a random effect component to account for bundle ability. The average number of hints for each person and item were provided as covariates in estimating person and item difficulty parameter estimates, respectively. Each of the latter three models was calibrated twice; once with covariates and once without covariates. Statistical relationships between scaled scores from each of the scoring models, a percent correct score and a criterion measure of student ability were also evaluated.

Chapter Four – Results

The purpose of this research is to help advance the development and use of assessment systems that utilize technological innovations and specifically those that incorporate scaffolding into the assessment process. The goal is to make recommendations about optimal scoring models that can be used for scaffolded assessments based on the characteristics of the scaffolds utilized in the example assessment system. As such, a number of different models were applied to the Assistments data and relevant parameters were estimated for each model using MCMC estimation techniques. Several indices were calculated from the MCMC output in order to evaluate and compare the models with respect to convergence, model fit, and precision of model estimates. Finally, criterion related validity evidence was evaluated for the person ability parameters from each model using student scores from an external measure of student ability as the criterion.

Research Question 1: What type of model is the optimal scoring model for the scaffolded data in the Assistment system?

The 2PL_MainItems model calibrated parameters for the main items only without accounting for the scaffold items or the number of hints a student accessed for each item. The comparison models calibrated parameters for all 140 items (i.e., main and scaffold items) as well as the number of hints a student accessed; however, each of the comparison models differs with respect to how the scaffold items are accounted for. The 2PL_AllItems model ignores the item grouping that occurs as a result of the scaffolding process while the ORM and the TRM represent two different methods of accounting for the local dependence. Each of the comparison models was employed twice; once with and once without incorporating the average number of hints a student used in the scaffolding process as a covariate. In total, the

Assistments data were calibrated according to seven different models. The results of the model calibration and convergence process are presented next followed by a summary of model estimates, model fit statistics and information functions.

Convergence. Model convergence was evaluated from a number of perspectives. Convergence was statistically assessed using the PSRF convergence criterion of Gelman and Rubin (1993) that is automatically provided by SCORIGHT 3.0 whenever multiple parallel chains are specified. The PSRF was calculated for each estimated parameter at the 50% and 97.5% quantiles based on the Student t distribution (Wang, Bradlow & Wainer, 2005) after discarding the first half of the samples (or the specified number of samples to be discarded). However, if multiple chains could not be simultaneously analyzed (due to insufficient computer memory) the PSRFs were calculated from retained output. In addition to the PSRFs, other convergence issues such as number of required iterations and amount of time needed to converge are discussed with respect to each model in the following sections.

2PL_MainItems Model. The 2PL_MainItems model calibrated response data for the 32 main items only based on the 2PL model and did not account for any scaffolding features. Three parallel chains were run for the 2PL_MainItems model, each of which contained 10,000 draws from the posterior distribution, 5,000 of which were discarded for burn-in. From the last 5,000 iterations, every 10th draw was recorded for a total of 500 retained draws for each parameter. Using these specifications, convergence was attained as indicated by PSRFs equal to 1.00 at both quantiles points for both item difficulty (a) and discrimination parameters (b), as shown in Table 6 below. Given the relative simplicity of the model and the small number of items being calibrated, it was reasonable to believe that convergence may have been achieved with a fewer number of iterations. Therefore, the model was run a second time with

only 3,000 iterations, 1,500 of which were discarded. Of the 1,500 retained draws, every 10th draw was recorded for a total of 150 draws for each parameter. Using this second set of specifications, convergence was again attained.

Table 6.

Estimation Specifications and PSRFs for the 2PL_MainItems Model

Model Specifications	<u>PSRF for parameter <i>b</i></u>		<u>PSRF for parameter <i>a</i></u>		Approx. Run Time
	50%	97.5%	50%	97.5%	
10000/5000/10	1.00	1.00	1.00	1.00	0:25
3000/1500/10	1.00	1.01	1.00	1.00	0:15

Note. Model Specifications = number of total iterations/number of iterations discarded for burn-in/size of gap between posterior draws recorded

There were no notable convergence issues and the amount of time (hr:min) needed to run the model was considerably fast in the context of MCMC estimation. Specifically, the model that iterated 10,000 times took approximately 0:25 minutes to complete; however, as was shown by using the second set of model specifications, convergence was met after only 3,000 iterations. This second model run only took 0:15 minutes to complete. While the goal of this analysis was not to determine the absolute minimum amount of time that is needed to attain convergence, it is sensible to think that 0:15 minutes is the maximum amount of time needed for this particular model to converge and that it could be achieved in even fewer iterations and chains. In any case, the 2PL_MainItems model attained convergence in a relatively short amount of time with no issue.

2PL_AllItems Model. The 2PL_AllItems model calibrated response data for all 140 items but did not account for the local dependence created by the scaffolding process. This model was calibrated both with and without incorporating covariates in the estimation process. Similar to the 2PL_MainItems model, three parallel chains were ran, each of which

contained 10,000 draws from the posterior distribution, 5,000 of which were discarded for burn-in. From the last 5,000 iterations, every 10th draw was recorded for a total of 500 retained draws for each parameter. Using these specifications, convergence was attained for the model that incorporated the covariates in the estimation process. As shown in Table 7 below, the PSRFs for both item difficulty and discrimination parameters met the criterion of less than or equal to 1.2 for the model with covariates; however, the *b* parameter for the model without covariates did not meet this criterion at the 97.5% quantile. Therefore, this model was calibrated a second time with twice as many iterations. Convergence was attained for the model without covariates with 20,000 iterations.

Table 7.

Estimation Specifications and PSRFs for each 2PL_AllItems Model

Model	Model Specifications	<u>PSRF for parameter <i>b</i></u>		<u>PSRF for parameter <i>a</i></u>		Approx. Run Time
		50%	97.5%	50%	97.5%	
2PL_AllItems + cov	10000/5000/10	1.01	1.05	1.00	1.01	1:50
2PL_AllItems	10000/5000/10	1.09	1.34	1.01	1.05	2:00
	20000/10000/20	1.00	1.00	1.00	1.00	4:00

Note. Model Specifications = number of total iterations/number of iterations discarded for burn-in/size of gap between posterior draws recorded

Aside from the additional iterations needed for the 2PL model without covariates to meet the specified convergence criterion, there were no other convergence issues associated with model estimation. The amount of time a model would run was rounded to the nearest quarter of an hour. The amount of time needed to run each 2PL_AllItems model was approximately 2:00 hours. There was not a significant difference in estimation time between the model with covariates and the model without covariates; the model with covariates took

about 10 minutes less to complete than the model with covariates. However, the b parameter for the model without covariates did not sufficiently meet the convergence criterion at the 97.5% quantile. This model was run a second time with 20,000 iterations which allowed for the b parameter to sufficiently converge. The increase in the number of iterations also increased the time needed to run the model to approximately 4:00 hours. Overall, the model that incorporated the covariates in the estimation process took less time to complete and successfully converge.

Ordinal Response Model. The polytomous ORM calibrated response data for the 32 bundles in order to account for the local dependence between items within a bundle. Thus, summed scores were calculated for item bundles and treated as single super-items that were scored polytomously using the ORM. This model was also calibrated with and without incorporating covariates in the estimation process. The same initial model specifications that were applied to the previous two model types were also used for estimating the polytomous models. Convergence was attained for the model that incorporated the covariates in the estimation process. As shown in Table 8 below, the PSRFs for both item difficulty and discrimination parameters met the criterion of less than or equal to 1.2 for the model with covariates; however, the b parameter for the model without covariates did not meet this criterion at the 97.5% quantile. Similar to the previous 2PL model, the ORM without covariates model was calibrated a second time with twice as many iterations. Again, convergence was attained for the model without covariates after 20,000 iterations.

Table 8.

Estimation Specifications and PSRFs for each Ordinal Response Model

Model	Model Specifications	PSRF for parameter <i>b</i>		PSRF for parameter <i>a</i>		Approx. Run Time
		50%	97.5%	50%	97.5%	
ORM + covs	10000/5000/10	1.01	1.03	1.03	1.11	6:00
ORM	10000/5000/10	1.07	1.29	1.01	1.03	4:30
	20000/10000/20	1.00	1.01	1.00	1.01	8:15

Note. Model Specifications = number of total iterations/number of iterations discarded for burn-in/size of gap between posterior draws recorded

Once convergence was achieved with the additional number of iterations for the model without covariates, there were no other convergence issues associated with model estimation. The ORM with covariates took approximately 6:00 hours. The model that did not use covariates in the estimation process appeared to take less time to complete than the model that incorporated covariates. However, the former model did not meet the convergence criterion and when it was ran a second time with 20,000 iterations, the length of time to complete increased to approximately 8:15 hours. Thus, the amount of time for the ORM to complete *and* converge was faster for the model that incorporated the covariates in the estimation process.

Testlet Response Model. The TRM calibrated the response data to all 140 items but additionally accounted for a random effect component to explain the interaction between the person and the bundle; i.e., a bundle ability parameter. This random effect component essentially accounts for item dependency within bundles but unlike the polytomous approach, the TRM does not lose the response patterns to individual items. However, since each bundle ability parameter is treated as an additional dimension in the model, the TRM can become quite complex to estimate. The TRM was calibrated both with and without incorporating

covariates in the estimation process. The same initial model specifications that were applied to the previous three model types were also used for estimating the TRMs.

Convergence was not attained for either the model with covariates or the model without covariates. To assist the estimation process, parameter estimates from the 2PL model (with covariates) were used as starting values for all subsequent model estimations. As displayed in Table 9 below, the number of iterations was increased ultimately to 100,000 iterations to try to achieve convergence. PSRFs were provided for item parameters as well as for the variances of each testlet parameter (γ); however, the convergence indices for the variances of γ s are only provided for the final models that used 100,000 iterations. The models that used 50,000 iterations or more could not be estimated when three chains were specified due to lack of working memory space using a dual processor P8700 at 2.53GHz. Therefore, the model that used 50,000 iterations was estimated with only two chains and model run time was approximated for three chains based on the time needed for two chains. The model that used 100,000 iterations was estimated one chain at a time; as such, SCORIGHT 3.0 could not provide convergence diagnostics as was the case with the other models. For this model, PSRF values were calculated for each set of parameter draws. While item parameters, a and b , finally met the convergence criterion after 100,000 iterations, there were at least three testlet parameters that did not converge (based on the PSRF value at the 50% quantile) for the model that incorporated covariates and one testlet for the model without covariates. As shown in Table 10 below, most testlet parameters did not meet the convergence criterion of less than 1.2 at the 97.5% quantile.

Table 9.

Estimation Specifications and PSRFs for each Testlet Response Model

Model	Model Specifications	PSRF for parameter <i>b</i>		PSRF for parameter <i>a</i>		Approx. Run Time
		50%	97.5%	50%	97.5%	
TRM + covs	10000/5000/10*	1.04	1.12	4.66	8.78	5:30
	10000/5000/10	1.01	1.07	4.60	8.69	5:30
	20000/10000/20	1.10	1.21	4.43	7.34	11:00
	50000/40000/20†	1.00	1.01	1.45	2.98	26:00
	100000/90000/20††	1.00	1.00	1.00	1.00	51:00
TRM	10000/5000/10*	1.17	1.54	6.84	12.28	5:00
	10000/5000/10	1.11	1.52	6.79	12.24	5:00
	20000/10000/20	1.12	1.33	4.46	7.76	10:00
	50000/40000/20†	1.01	1.06	1.82	3.27	25:00
	100000/90000/20††	1.00	1.00	1.00	1.00	50:00

Note. Model Specifications = number of total iterations/number of iterations discarded for burn-in/size of gap between posterior draws recorded; *did not use parameter estimates from 2PL as initial values; all subsequent models were provided these initial values; Approx. Run Time = indicates approximate time in hours:minutes for 3 chains; † Due to large amount of working memory required for 50000 iterations, only 2 chains were run simultaneously; †† Due to large amount of working memory required for 100000 iterations, only 1 chain was run at a time

Table 10.

PSRFs for Variances of Gammas for each Testlet Response Model

Bundle #	<u>TRM + covs</u>		<u>TRM</u>	
	50%	97.5%	50%	97.5%
1	1.00	1.01	1.01	2.46
2	1.00	1.42	1.00	1.00
3	1.00	1.30	1.00	1.05
4	1.00	1.11	1.00	1.25
5	1.00	0.99	1.00	1.00
6	1.03	2.58	1.02	1.22
7	1.02	3.42	1.05	2.89
8	1.00	1.40	1.00	1.52
9	1.00	1.08	1.00	1.48
10	1.00	1.38	1.02	2.99
11	1.00	1.22	1.00	0.97
12	1.00	1.15	1.00	1.67
13	1.00	0.98	1.00	1.09
14	1.00	1.13	1.01	1.68
15	1.00	1.06	1.00	1.18
16	1.02	2.30	1.07	4.65
17	1.00	1.13	1.00	1.07
18	1.01	1.71	1.01	1.00
19	1.00	1.14	1.00	0.97
20	1.00	1.07	1.00	1.01
21	1.00	1.44	1.00	1.11
22	2.47	11.51	1.00	1.12
23	2.04	11.73	1.01	3.22
24	1.01	3.40	1.01	2.03
25	1.02	2.17	1.01	1.59
26	1.00	2.33	2.52	14.47
27	1.02	2.74	1.00	2.00
28	1.03	4.64	1.01	2.96
29	1.00	1.82	1.02	3.33
30	1.04	3.45	1.01	2.12
31	2.70	9.93	1.04	4.30
32	1.11	5.84	1.00	1.59

Note. Highlighted cells indicate testlets that did not converge

Moreover, in evaluating convergence of the testlet effects (variance of gamma, γ), it was discovered that at least one bundle in each of the model calibrations (i.e., with and without covariates) had estimates that approached infinity. That is to say, that as the samples were drawn from the posterior, these draws appeared to become increasingly larger than the last and continued to cycle upwards infinitely. After 100,000 iterations, these estimates had exponents of 63 and greater (e.g., 3.491E+63). As an example, Figure 14 below displays draws from the posterior for the variances of gamma for the first bundle in the dataset. It is easy to see that within the first 5,000 iterations, the algorithm continues to draw larger and larger samples from the posterior but then maintains a relatively consistent distribution thereafter. Conversely, Figure 15 below displays draws for the variances of gamma for Bundle 26 estimated using the TRM (without covariates). This bundle never achieves a stationary distribution; rather, it continues to cycle upwards with no end in sight. Similarly, Bundle 22 calibrated with the TRM + covs model followed the same infinite cycle pattern as shown in Figure 16. Needless to say, these parameters were not within an interpretable range and attempts at rectifying this issue are discussed in the next section.

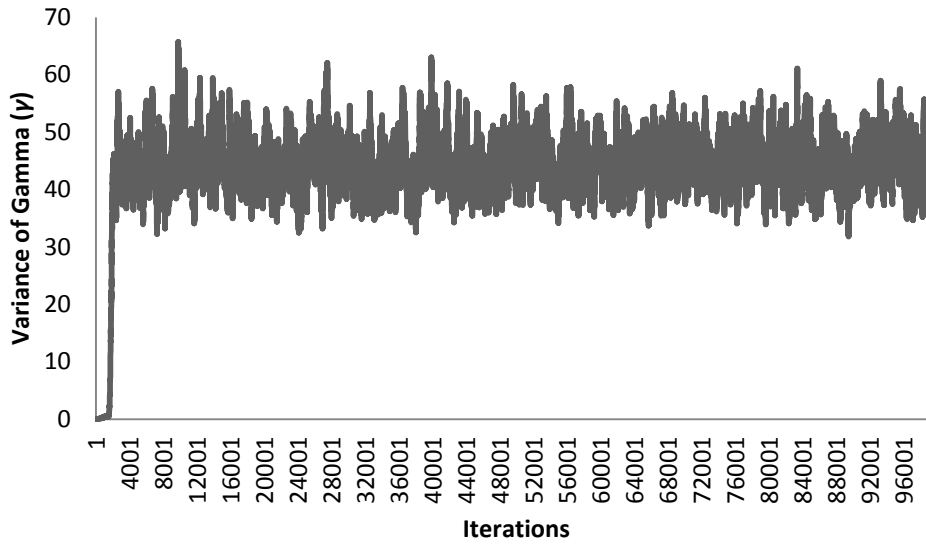


Figure 14. Time-series plot for the variance of gamma for Bundle 1 based on 100,000 iterations.

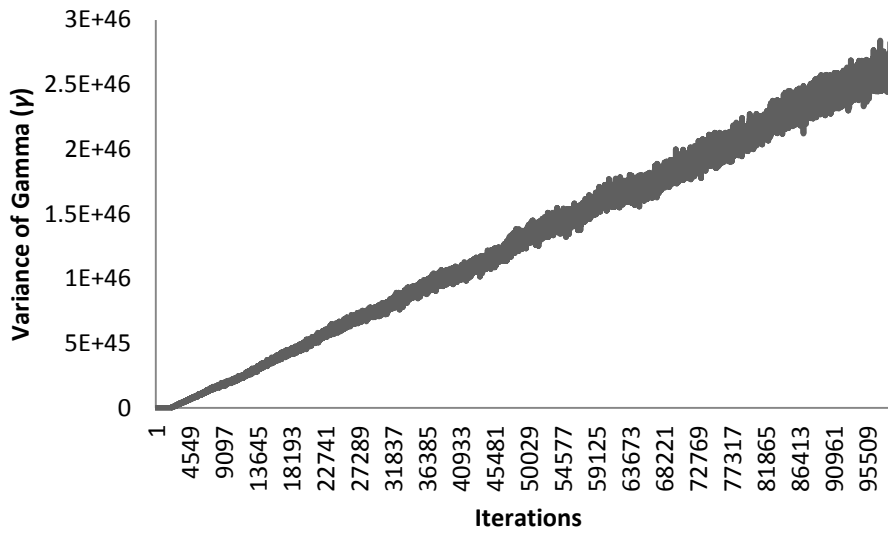


Figure 15. Time-series plot for the variance of gamma for Bundle 26 from the TRM (without covariates) model based on 100,000 iterations.

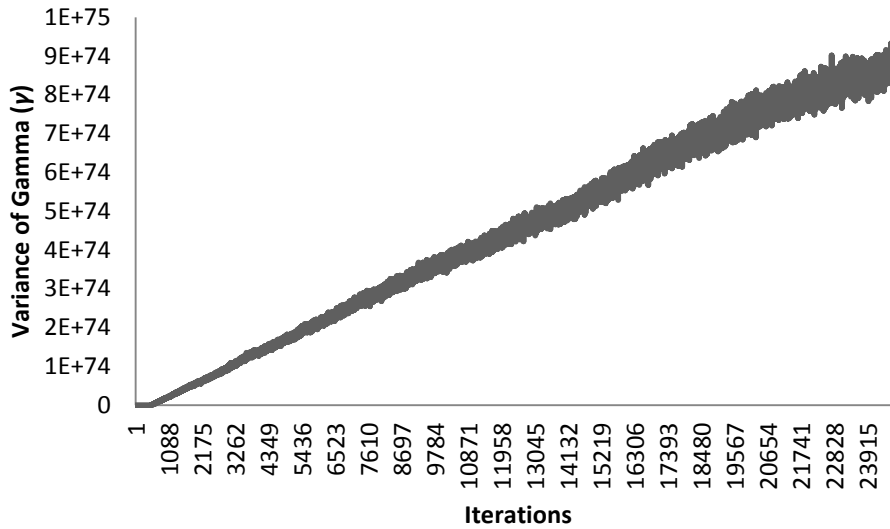


Figure 16. Time-series plot for the variance of gamma for Bundle 26 from the TRM + covs model based on 100,000 iterations.

Clearly there were several estimation issues associated with calibrating the TRMs. The model with the fewest number of iterations (10,000) required more time to complete than any of the other previous models and this model was far from converging. One chain of 10,000 iterations took approximately 1:40 hours to complete whereas one chain of 100,000 iterations ran for roughly 16:30 hours. Moreover, multiple chains could not be estimated in the same run for models that used 100,000 iterations (i.e., each chain had to be run separately due to lack of sufficient working memory space) and therefore, convergence could not be as readily obtained as the other models (i.e., it had to be calculated based on output from each chain). Even after 100,000 iterations, there were a few variances of testlet parameters that did not appear to converge at the 50% quantile. Implications of these results are detailed in the next chapter; however, for the sake of completeness, attempts were made to include this model in the all of the model analysis procedures. Issues associated with this model are

discussed within each of the following sections and procedures taken to resolve estimation difficulties are detailed in the following paragraphs.

Additional TRM Calibration Procedures. In an effort to attain convergence and reasonable estimates for the testlet parameters, several additional steps were taken towards calibrating the data using the TRM. Solutions were sought first from a software or model perspective and second from a data perspective. First, in order to determine if estimation issues may be due to a limitation in the software, the TRM (without covariates) was also run in WinBugs software (Spiegelhalter, Thomas, Best, & Lunn, 2003). Errors occurred when this model was run for any number of iterations greater than 1,500. Furthermore, based on a test run of 1,000 iterations that took almost 4 hours to complete, the amount of time estimated to complete 100,000 iterations was more than two weeks. As demonstrated from the SCORIGHT 3.0 output, the gamma estimates for some of the testlets were unreasonable and may have been causing the errors to occur in the estimation process in WinBugs. Therefore, it was decided that it may not be feasible to estimate the variances of gamma for this particular dataset. As such, these variances were set to equal one and the model was run again in WinBugs. While the model successfully completed after a test run of 5,000 iterations (which took approximately five hours to compile and run), the output for the gamma estimates could not be opened. The program attempted to retrieve the output for approximately four hours and then indicated that errors had occurred in the process.

Several changes to the dataset were also made in an effort to resolve the estimation issues and with each change, the model was re-run in both software programs. First, the data were cleaned for student profiles that had exact same response patterns. While it was not possible to determine with certainty, it appeared that some groups of students worked together

or as a class in responding to the Assistments items and as a result had the same response patterns. This accounted for approximately 10.8% of the data (399 cases). It was reasonable to believe that these response sets may be contributing to an overestimation of the testlet effects. However, results did not improve for either software program after deleting these cases. Some of the gamma estimates in SCORIGHT 3.0 still appeared to approach infinity and estimation procedures in WinBugs still resulted in errors. To potentially help address any estimation issues related to missing data, 15 bundles that were administered to fewer than 500 students were removed from the original dataset (i.e., the dataset with 2,745 student profiles). The models were again, re-run in both software programs and the same problems occurred. Finally, a sequential deletion of potentially problematic bundles was performed. Based on the SCORIGHT 3.0 output, gamma estimates for Bundles 22, 23, 26 and 31 did not converge. These bundles were removed sequentially and after each removal, the model was re-run in both programs. Again, the same estimation difficulties occurred after each bundle was removed.

In the end, an extensive amount of time and effort was undertaken to resolve the estimation issues associated with calibrating the TRM. Unfortunately, none of the steps described resulted in a solution. While a solution for fitting the data to the TRM was not determined, valuable information was obtained through this process and implications of these results are discussed in the next chapter.

Summary. Overall, the 2PL_MainItems model, both 2PL_AllItems models, and both polytomous models achieved convergence for all relevant parameters. The 2PL_MainItems model attained convergence with the fewest number of iterations. The 2PL_AllItems and polytomous models that incorporated covariates in the estimation process both met the

convergence criterion with 10,000 iterations while their counterparts that did not incorporate covariates required additional iterations to achieve convergence. None of these models had any notable convergence issues. On the other hand, the TRM required almost 10 times as many iterations; this was far more difficult for a standard processor to estimate which resulted in running one chain at a time. The TRMs were also the only set of models that were specified initial values to assist in the estimation process; random initial values were sufficient for the calibrating the other models. However, even with the initial values and additional iterations, conclusive evidence for convergence could not be attained for either of the TRMs and some of the testlet parameters obtained from these calibrations were not interpretable.

Descriptive Statistics. Table 11 below displays the number of students that were administered each item as well as the number and percent correct for each item. This descriptive information for the number and percent correct data is summarized in Table 12. Overall, with the exception of items 103 through 140, most items appeared to be relatively easy with most students responding to correctly to both the main and scaffold items. The average percentage correct for all items was 81.9%; the mean percentage correct for only the main items was lower at 70.4%. There were a number of items that had appeared to be much more difficult for students than most of the other items. For example, less than 19% of students answered item 109 correctly. Similarly, there were five other main items (items 100, 103, 113, 119 and 138) that had fewer than 40% correct responses.

The average percentage correct for scaffold items was rather high at 93.4%; however, it is important to keep in mind that missing data were recoded such that if a student responded correctly to a main item, then his or her responses to the scaffold items were coded as correct. Thus, this average percentage correct on scaffold items does not represent only those

students' responses that went through the scaffolding process; rather it also represents assumed student responses for those that did not go through the scaffolding process. If the average were taken from response patterns of students that only responded incorrectly to the main item, it would inevitably be lower than 93.4%.

Item parameter estimates were obtained for each of the evaluation models from the SCORIGHT 3.0 output. Table 13 below displays the average estimates for difficulty, discrimination and ability parameters for the 2PL_MainItems model. Tables 14-16 outline the relevant parameter estimates for each of the six comparison models.

2PL_MainItems Model. The average difficulty (b) parameter for the 2PL_MainItems model was -0.71 which indicates that on average a slightly below average ability level was required to have a 50% chance of getting a main item correct. The average discrimination (a) parameter was 2.81 which signifies that, on average, the main items discriminate between low and high ability students extremely well.

2PL_AllItems Model & ORM. In general, the pattern of average item parameters appeared to be fairly consistent for all of the comparison models except for the TRM. On average, the items calibrated by the 2PL_AllItems models and the polytomous ORMs, require an average to somewhat below average ability level to have a 50% chance of success. Furthermore, the items calibrated by these models also appear to discriminate unusually well. The average b parameters for the 2PL_AllItems models had greater negative values than the 2PL_MainItems model but smaller negative values than the ORMs. The average discrimination parameter was the highest for the 2PL_AllItems and the lowest for the polytomous ORMs. Thus, including the scaffold items in the estimation process increased average item discrimination; however, accounting for local dependence using the polytomous

approach decreases average a values. In any case, all of the discrimination estimates produced by these models were unusually high, particularly given the relatively low difficulty estimates. Possible explanations for these findings are presented in the discussion section. It should also be noted that adding the covariate into the estimation process of the b parameter did not appear to meaningfully impact item parameter estimates.

Table 11.

Descriptive Statistics for each Item

Item	Item Indicator*	n	n Correct	Percent Correct	Item	Item Indicator*	n	n Correct	Percent Correct
1	1	1103	1023	92.75	34	1	1020	911	89.31
2	0	1102	1050	95.28	35	0	1017	971	95.48
3	0	1101	1053	95.64	36	0	1016	934	91.93
4	0	1100	1061	96.45	37	0	1015	939	92.51
5	1	1162	984	84.68	38	0	1015	986	97.14
6	0	1160	1109	95.60	39	0	1015	983	96.85
7	0	1159	1080	93.18	40	0	1015	963	94.88
8	0	1157	1115	96.37	41	1	983	872	88.71
9	0	1154	1030	89.25	42	0	982	934	95.11
10	1	975	817	83.79	43	0	980	937	95.61
11	0	975	881	90.36	44	0	979	955	97.55
12	0	974	868	89.12	45	1	407	256	62.90
13	0	966	882	91.30	46	0	406	312	76.85
14	0	966	881	91.20	47	0	405	329	81.23
15	1	1071	978	91.32	48	0	402	342	85.07
16	0	1070	999	93.36	49	1	1623	1457	89.77
17	0	1069	1046	97.85	50	0	1622	1542	95.07
18	0	1068	1044	97.75	51	0	1622	1525	94.02
19	0	1068	1001	93.73	52	0	1621	1590	98.09
20	0	1068	1043	97.66	53	0	1620	1559	96.23
21	1	1046	953	91.11	54	1	1612	1459	90.51
22	0	1047	1009	96.37	55	0	1611	1492	92.61
23	0	1046	1003	95.89	56	0	1608	1534	95.40
24	0	1045	1028	98.37	57	0	1607	1504	93.59
25	0	1045	1031	98.66	58	1	1003	900	89.73
26	1	377	247	65.52	59	0	1002	953	95.11
27	0	376	344	91.49	60	0	1002	977	97.50
28	0	375	313	83.47	61	0	1002	942	94.01
29	1	1064	834	78.38	62	1	1609	1432	89.00
30	0	1064	890	83.65	63	0	1607	1495	93.03
31	0	1053	923	87.65	64	0	1604	1470	91.65
32	0	1048	966	92.18	65	1	1610	1476	91.68
33	0	1048	921	87.88	66	0	1608	1535	95.46

Table 11 continued.

Descriptive Statistics by Item

Item	Item Indicator*	n	n Correct	Percent Correct	Item	Item Indicator*	n	n Correct	Percent Correct
67	0	1608	1521	94.59	100	1	809	218	26.95
68	0	1608	1518	94.40	101	0	804	541	67.29
69	1	301	176	58.47	102	0	801	454	56.68
70	0	299	227	75.92	103	1	494	196	39.68
71	0	298	248	83.22	104	0	489	266	54.40
72	0	296	194	65.54	105	1	1000	682	68.20
73	0	295	226	76.61	106	0	998	786	78.76
74	0	293	242	82.59	107	0	992	730	73.59
75	0	293	268	91.47	108	0	984	816	82.93
76	0	265	225	84.91	109	1	416	78	18.75
77	0	291	262	90.03	110	0	408	156	38.24
78	1	1652	1462	88.50	111	0	375	215	57.33
79	0	1643	1537	93.55	112	0	387	222	57.36
80	0	1638	1565	95.54	113	1	541	214	39.56
81	0	1635	1509	92.29	114	0	532	293	55.08
82	1	1275	1071	84.00	115	0	518	436	84.17
83	0	1265	1111	87.83	116	0	517	221	42.75
84	0	1262	1147	90.89	117	0	502	357	71.12
85	0	1261	1146	90.88	118	0	496	412	83.06
86	0	1254	1131	90.19	119	1	886	295	33.30
87	1	1172	992	84.64	120	0	882	497	56.35
88	0	1170	1091	93.25	121	0	878	632	71.98
89	0	1171	1154	98.55	122	0	868	547	63.02
90	0	1171	1088	92.91	123	1	415	227	54.70
91	0	1171	1098	93.77	124	0	413	308	74.58
92	1	1146	1040	90.75	125	0	413	291	70.46
93	0	1140	1109	97.28	126	1	725	390	53.79
94	0	1140	1085	95.18	127	0	722	466	64.54
95	0	1138	1095	96.22	128	0	721	582	80.72
96	1	1147	1049	91.46	129	0	720	512	71.11
97	0	1144	1087	95.02	130	0	719	529	73.57
98	0	1142	1067	93.43	131	1	727	450	61.90
99	0	1140	1112	97.54	132	0	726	581	80.03

Table 11 continued.

Descriptive Statistics by Item

Item	Item Indicator*	n	n Correct	Percent Correct
133	0	723	530	73.31
134	1	722	340	47.09
135	0	721	494	68.52
136	0	720	549	76.25
137	0	720	517	71.81
138	1	712	237	33.29
139	0	712	353	49.58
140	0	713	527	73.91

* 1 = main item; 0 = scaffold item

Note. Response data were recoded such that correct responses on main items corresponded to correct responses on scaffold items.

Table 12.

Summary Statistics for Original Data (not calibrated with IRT)

Data		Mean	Std Dev	Min	Max
Bundles	n	954.98	389.58	265	1652
	n Correct	820.52	424.39	78	1590
	Percent Correct	81.87	17.42	18.75	98.66
Main Items Only	n	962.66	394.39	301	1652
	n Correct	741.13	455.18	78	1476
	Percent Correct	70.44	23.14	18.75	92.75
Scaffold Items Only	n	1012.41	175.37	375	1160
	n Correct	948.34	176.16	313	1115
	Percent Correct	93.39	4.05	83.47	98.66
Students	n Items	48.71	25.60	15	127
	n Items Correct	41.85	27.78	0	123
	Percent Correct	82.39	23.17	0.00	100

Note. n = number of students that were administered items; n Items = number of items that were administered to students

Testlet Response Model. Conversely, the TRM produced very different results for this dataset. Please note that the TRMs had several estimation issues and a number of the bundles did not meet the convergence criterion. The summary of item parameter estimates is only presented for the sake of completeness; however, implications associated with interpreting these parameter estimates are discussed in the next chapter. The average difficulty parameter for the TRM was -6.89 suggesting that the items were very easy; that is, based on the TRM an ability level at the extreme low end of the ability scale was needed to successfully respond to an item. The average discrimination parameter estimate was approximately equal to 1.00 which reflects a more typical level of discrimination. The fact that these results do not follow conventional form for IRT models (i.e., typically easier items discriminate less well) warrants even more caution in their interpretation.

The gamma estimates in the TRMs are not, by themselves, that meaningful. It is the variances of gamma that are useful for describing the amount of local dependence that exists in a group of items. The estimates of the variances of gammas, presented in Table 17, were extremely large and for a couple of bundles (bundles 22 and 26) they appeared to approach infinity and were not interpretable. As the testlet effects are relative to the variance of person abilities, a variance of gamma equal to 1.0 is considered a large variance (Wang et al., 2002). Variances of gamma greater than 2.0 are considered very large testlet effects. As the smallest variance of gamma for the TRMs was approximately 26.0 to 27.0, it is clear that a significant amount of local dependence exists in this dataset.

Table 13.

Item Parameter Mean, Average Standard Error and Range for the Dichotomous 2PL_MainItems Model

	Mean	Average Std Error	Min	Max
Difficulty (<i>b</i>)	-0.711	0.077	-1.344	0.553
Discrimination (<i>a</i>)	2.814	0.335	0.758	5.662

Table 14.

Item Parameter Mean, Average Standard Error and Range for the Dichotomous 2PL_AllItems Model

Parameter	<u>2PL_AllItems + covs</u>				<u>2PL_AllItems</u>			
	Mean	Mean Std. Error	Min	Max	Mean	Mean Std Error	Min	Max
Difficulty (<i>b</i>)	-1.121	0.059	-1.715	-0.129	-1.118	0.056	-1.705	-0.135
Discrimination (<i>a</i>)	3.813	0.454	1.585	6.937	3.832	0.456	1.604	6.891

Table 15.

Item Parameter Mean, Average Standard Error and Range for the Dichotomous Polytomous Ordinal Response Models

Parameter	<u>ORM + covs</u>				<u>ORM</u>			
	Mean	Average Std Error	Min	Max	Mean	Average Std Error	Min	Max
Difficulty (<i>b</i>)	-1.829	0.111	-2.592	-0.809	-1.816	0.110	-2.550	-0.814
Discrimination (<i>a</i>)	1.551	0.136	0.733	2.436	1.551	0.138	0.753	2.447
Category Boundary (<i>k</i>)	1.257	0.127	0.361	2.975	1.216	0.114	0.373	2.952

Table 16.

Item Parameter Mean, Average Standard Error and Range for the Dichotomous Testlet Response Models

Parameter	Mean	<u>TRM + covs</u>			Mean	<u>TRM</u>		
		Avg. Std Error	Min	Max		Avg. Std Error	Min	Max
Diff. (<i>b</i>)	-6.89	0.64	-9.95	9.72	-6.62	0.55	-9.96	5.69
Discrim. (<i>a</i>)	0.99	0.91	0.11	13.04	0.77	0.51	0.09	22.41
Variance of gamma (γ)	313.02	132.21	27.62	2983.90	318.24	500.98	25.69	3050.76

Note. Variance of γ calculations based only on estimates that were less than "infinity" (non-highlighted cells in Table X).

Table 17.

Estimated Variances of Gamma (γ) and Standard Errors for each Bundle

Bundle	<u>TRM + covs</u>		<u>TRM</u>	
	Variance of γ	Standard Error	Variance of γ	Standard Error
1	38.377	3.334	45.336	4.090
2	45.779	3.621	50.315	4.061
3	86.961	8.932	100.338	11.576
4	31.411	2.800	30.900	2.722
5	27.617	2.591	25.687	2.284
6	79.764	10.473	82.153	17.750
7	126.337	14.659	187.647	28.991
8	47.121	4.706	41.916	3.751
9	33.744	3.012	31.513	2.946
10	169.154	31.981	355.768	85.081
11	29.226	2.221	28.624	2.058
12	45.391	3.506	49.871	4.162
13	34.917	3.086	36.164	3.396
14	61.849	4.927	67.806	5.350
15	41.198	3.080	45.888	3.212
16	206.553	40.960	349.580	84.710
17	47.996	3.683	47.888	3.363
18	92.655	8.995	111.123	11.284
19	36.255	3.120	34.886	2.915
20	34.621	3.030	32.399	2.963
21	36.540	3.356	39.557	3.753
22	8.434E+74	2.744E+73	343.774	37.344
23	2983.899	5.835E+62	3050.763	645.123
24	462.544	462.544	415.054	68.335
25	472.987	472.987	313.643	57.053
26	479.942	479.942	2.530E+46	9.332E+44
27	520.096	520.096	550.117	5.495E+33
28	628.768	628.768	801.298	801.298
29	874.219	508.018	827.175	827.175
30	738.415	285.956	621.050	621.050
31	802.642	335.697	707.954	5898.309
32	386.517	106.212	439.119	5783.179

Note. Highlighted cells indicate estimates that approached infinity.

Figures 17 and 18 below display the item discrimination parameter and difficulty parameter, respectively, for every item calibrated by each of the comparison models. As incorporating the covariate for the b parameter did not appear to change the estimate, only the models estimated without covariates are presented. The 2PL_AllItems models, which ignore local dependence, had larger discrimination parameters than the other two models that account for local dependence. This was true for models both with and without covariates. Item difficulty estimates for the 2PL_AllItems models and the polytomous ORMs were relatively consistent at around zero on the ability scale. However, as noted earlier, the b parameters for the TRM were vastly lower than the other two models. Again, this was true for models both with and without covariates.

Summary. In review, the percentage correct scores indicated that most items were relatively easy with the exception of a handful of items that had percentages less than 40%. Calibrating the Assistments data with IRT models revealed relatively consistent item parameters for the 2PL_MainItems model, the 2PL_AllItems models and the polytomous ORMs. In general, the difficulty parameters were at or somewhat below zero indicating that an average or somewhat below average ability level was required to have a 50% chance of success on an item. Incorporating the covariate in the estimation process of the b parameter did not appear to change the resulting parameter estimates. While the 2PL_AllItems models which included the scaffold items had higher discrimination parameter estimates than the 2PL_MainItems model or polytomous models, a parameter estimates were relatively high across all three types of models.

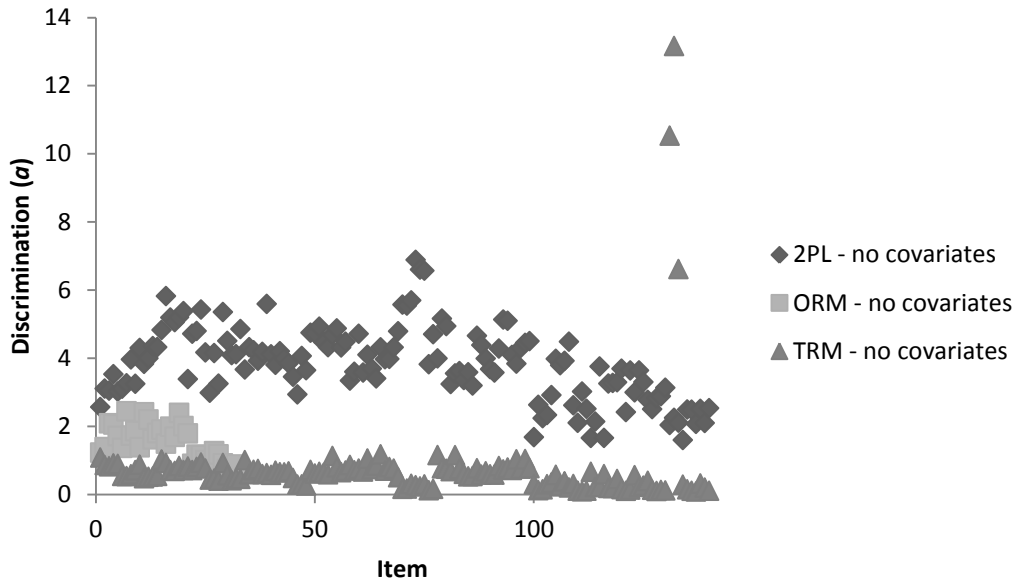


Figure 17. A comparison of item discrimination values for each model that did not incorporate covariates in the estimation process.

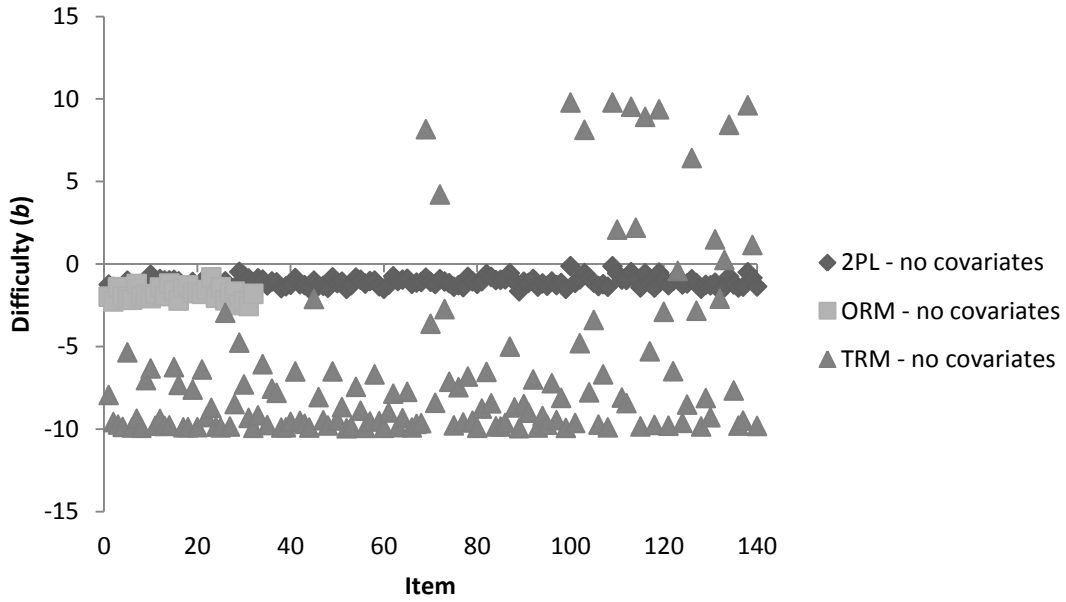


Figure 18. A comparison of item difficulty values for each model that did not incorporate covariates in the estimation process.

The TRM had several estimation issues rendering interpretation of item parameter estimates difficult if not untenable. The estimates obtained were vastly different from those of the other models. In particular, the b parameter estimates were much lower than the other model estimates. The a parameters for the TRM were also consistently lower than the other models. The estimates of the variances of gamma were extremely large, and in many cases, unreasonably large suggesting the presence of significant local dependence. However, given the lack of convergence and interpretable parameter estimates for a number of the bundles, direct comparisons between the TRMs and the other evaluation models are not justified.

Model Fit. The DIC statistic was calculated for the 2PL_MainItems model, the 2PL_AllItems models and the polytomous ORMs. Two programs were written in Fortran: one to calculate the DIC for the dichotomous models and one for the polytomous models which are provided in Appendix B and C, respectively. The DIC for the TRMs could not be obtained from the output provided by the SCORIGHT 3.0 program due to the gamma draws that approached infinity. These estimates, which appeared in both the TRM + covs and the TRM, rendered the output unusable. The amount of space that was allocated for these draws was not large enough to retain reasonable formatting. Values that appeared after those that approached infinity (e.g., 3.491E+63) were not differentiated with a space or tab which made the matrix of posterior draws completely unusable from a programming standpoint. As described previously, every effort was made to calibrate this model in WinBugs as the DIC is automatically calculated in this software program; however, WinBugs was not able to successfully calibrate the model.

Table 18 below summarizes the calculations for the DIC index for the 2PL_MainItems model, the 2PL_AllItems models and the polytomous ORM models. The pD value is

typically used to estimate the ‘effective number of parameters’ and it is equal to the difference between the mean of the deviance (\bar{D}) and the deviance at the posterior expectations ($D\bar{\theta}$). Again, it should be noted that due to the differences in data structures, only direct comparisons of DIC can be made between the + covs models and their counterparts that did not use covariates. Unfortunately, as DICs for the TRMs could not be calculated, comparisons with the 2PL_AllItems models could not be made.

Table 18.

Deviance Results for each Evaluation Model

Model	\bar{D}	$D\bar{\theta}$	pD	DIC
2PL_MainItems	17848.82	15583.66	2265.16	20113.98
2PL_AllItems + covs	55826.34	56550.24	-723.90	55102.44
2PL_AllItems	55819.53	56557.09	-737.55	55081.98
ORM + covs	36298.60	35714.68	583.92	36882.52
ORM	36296.34	35732.91	563.43	36859.77
TRM + covs	---	---	---	---
TRM	---	---	---	---

Note. \bar{D} = posterior mean of the deviance; $D\bar{\theta}$ = deviance at the posterior means; pD = measure of model complexity based on difference between \bar{D} and $D\bar{\theta}$; DIC = deviance information criterion ($\bar{D} + pD$)

For the 2PL_AllItems models and the ORMs, incorporating a covariate into the estimation process appeared to worsen the fit of the model. The difference in DIC values for the 2PL_AllItems models with and without covariates was 20.46, in favor of the model that did not use covariates. Similarly, the difference between the two ORMs was 22.75, again in favor of the model without covariates. While it is difficult to evaluate DIC error (Zhu and Carlin, 2000), the Bugs Project website (<http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/dicpage.shtml>) has suggested that differences of 10 or more would be more than substantial evidence for model selection. As mentioned, while the DIC values were readily interpretable

for evaluating the utility of the covariates, comparisons of the DICs across the various model types are not warranted due to the differing ways in which the data were defined for each model. Explanations and implications of these results are presented in the next chapter.

Information. Item and test information functions were calculated for each model. The TIFs for each model are displayed in Figure 19. The solid line in the graph below represents the 2PL_MainItems model; the dotted lines signify the 2PL_AllItems models which ignore local dependence; and the dashed lines are for the ORMs which account for local dependence. Again, information for the TRM could not be calculated due to the testlet effects that were out of reasonable range. The results clearly showed that more information is provided by the 2PL_AllItems models when theta is between approximately -2.0 and 0.0. The peak of the 2PL_AllItems curve is much higher reaching a maximum height that is more than four times greater than the peak of the of the ORMs. However, when theta is at the low end of the spectrum, i.e., less than -2.50, the ORMs appear to provide more information than the 2PL_AllItems models. When theta is average or above average the models seem to provide relatively the same amount of information. The 2PL_MainItems model, not surprisingly, provided the least amount of information across the ability scale.

The TIF for the 2PL_AllItems + covs model completely overlapped with the 2PL_AllItems model that did not use covariates. The TIFs for the ORMs were very similar; however, the ORM + covs was shifted to the right and had slight spike in information at the very low end of the theta scale. In general, incorporating covariates into the estimation process did not appear to significantly impact the amount of information that the test provides.

Information for each type of scoring model was also calculated for each group of items that formed a bundle. When all items were calibrated using the 2PL_AllItems model,

information was summed across each item in the bundle for a total bundle information function. These bundle information functions were compared to the item information functions for the polytomous model which are sums of score category information functions. Essentially, information is compared across bundles when local dependence is ignored and when it is accounted for. Figures 20 and 21 below display each bundle information function for the 2PL_AllItems model (without covariates) and the ORM (without covariates), respectively. The same comparisons are made between these two scoring models and the IIFs for the 2PL_MainItems model on a bundle-by-bundle basis and are available in Appendix D. In general, the bundle information functions followed the same pattern as the TIFs. That is, for most of the bundles the 2PL_AllItems provided more information when theta was just below average than the ORM but the ORM appeared to provide more information at the lower end of the ability scale. There was one bundle (Bundle 19) that the ORM provided slightly more information than the 2PL_AllItems but the difference was slight. A comparison between item information functions for the 2PL_MainItems and bundle information for the ORM indicated that there were no consistent patterns in the amount of information provided by these two scoring models. That is, for some bundles, accounting only for responses to the main item provided more information, while for other bundles, accounting for responses to the scaffold items and the local dependence within the bundle provided more information across the theta scale. There were also bundles in which the same amount of information was provided by both of these scoring models.

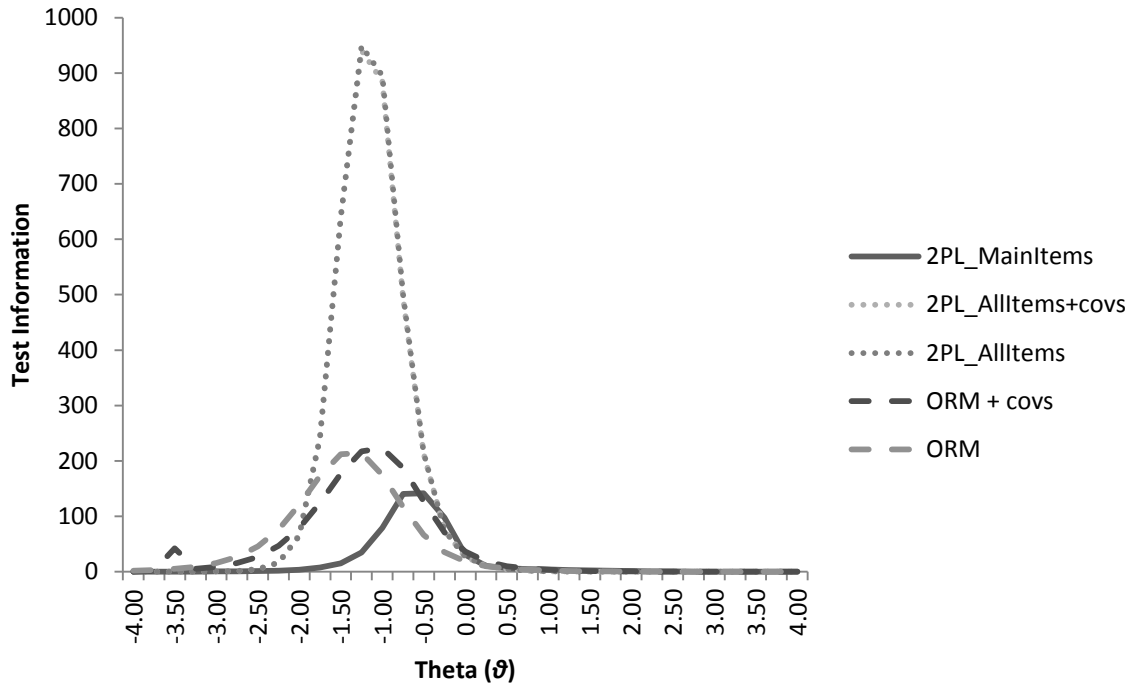


Figure 19. Total test information for each scoring model

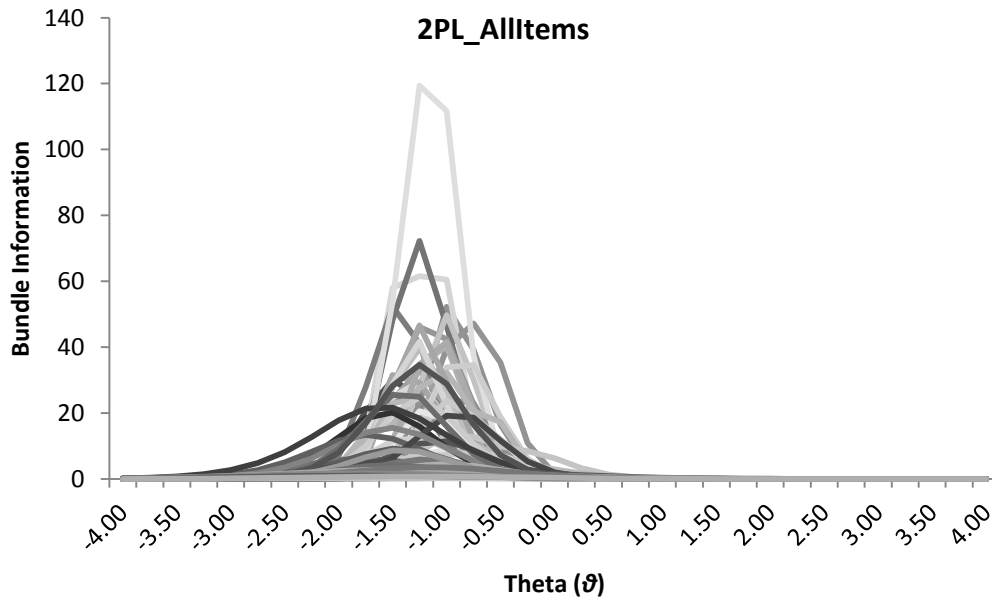


Figure 20. Total bundle information for 2PL_AllItems scoring model (without covariates) which ignore local dependence

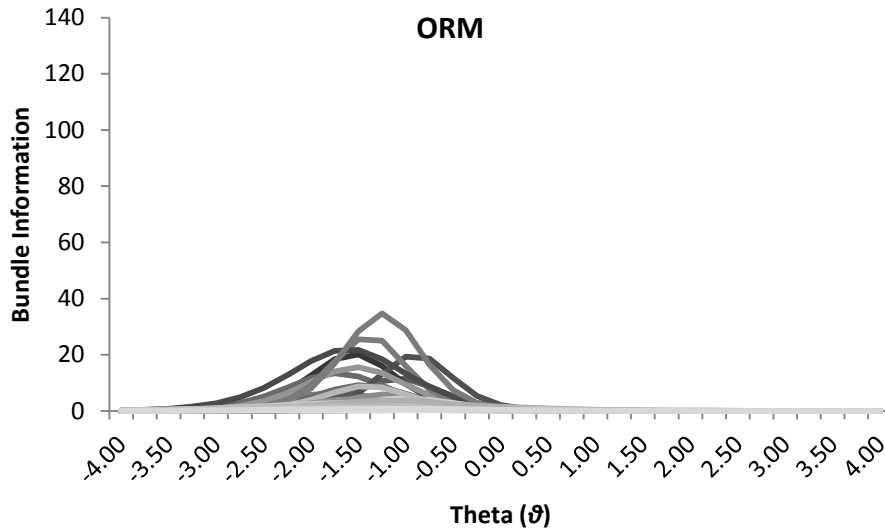


Figure 21. Total test information for ORM scoring model (without covariates) which account for local dependence

Overall, the 2PL_AllItems provided the most information across the majority of the ability scale. Thus, the models that ignored local dependence provided more information than the ORMs that accounted for it. The loss of response patterns that occurs when summed scores are obtained for the polytomous approach resulted in substantial loss of information. However, for very low ability examinees, the ORMs provided more information than the 2PL_AllItems models. Not surprisingly, the 2PL_MainItems model provided the least amount of information across the ability scale. Incorporating covariates into the estimation process did not appear to meaningfully alter model precision. Comparing bundle information functions demonstrated similar patterns for the majority of bundles.

Research Question 2: Is there a relationship between student ability estimates derived from the scoring models and a criterion measure of student achievement?

Ability estimates from each of the scoring models calibrated in the previous section were evaluated to determine the degree to which they related to student performance on a

state accountability assessment. Student ability estimates along with percent correct scores derived from the response data were correlated with a criterion measure of student ability based on the end-of-year state assessment. Statistical relationships were also computed between each of the scoring models and the percent correct score.

The average number of Assistentms items that a student took was approximately 49 and most students responded correctly to about 42 of those items. Thus, on average students appeared to perform well on the items that they were administered. The range of possible scores on the state assessment used in this study is 200 – 280. There were 778 student profiles that had corresponding state test scores. The mean state test score for this subsample of students was 229.13 (sd = 17.11). While the range of test scores was 204 to 280 which mostly covered the range of possible scores; the mean appeared to be below average assuming that the average score would be about mid-range. Information regarding average state test scores dating back to 2005 could not be retrieved. As such, it is difficult to make any implications about the representativeness of this subsample of students with respect to their peers.

Relationships between Scoring Models. The correlation matrix, displayed in Table 19 below, allows for comparisons of the relationships between ability estimates calibrated by each model and with the criterion measure. Statistical relationships between the scoring models indicated statistically significant and moderate to strong relationships between proficiency estimates from all of the scoring models (except the TRMs) and the percent correct metrics. Correlation coefficients between each the types of scores were all greater than 0.88 and all were significant at the $p < .001$ level. Thus, scaled scores obtained from the IRT models were quite consistent across the model variations and these scaled scores were

also strongly related to percent correct metrics. There did not appear to be any noteworthy differences in proficiency estimates between models that ignored local dependence and those that accounted for it. Also, for each type of model, there were no differences between those that incorporated a covariate for theta and those that did not.

The proficiency estimates for the TRMs were included in this analysis but should be interpreted with much caution as these models encountered several estimation issues including lack of convergence. These results are presented here to be used as potential information to help explain issues associated with this model. While the relationships between proficiency estimates from the TRMs and scores from the other models were statistically significant, they were considered trivial or weak at best, with the exception of the ORMs. The correlation coefficients between the TRMs and the ORMs were relatively strong, albeit not as strong as those found between the other types of scoring models. Thus, while scaled scores from the TRMs were mostly unrelated to the other types of scores that ignored local dependence, they *were* related to the scores that accounted for local dependence.

Relationships with Criterion. The correlation coefficients between state assessment scores and seven of the nine different types of scoring metrics ranged from $r = 0.50$ to 0.63 ; scores from the TRMs had no relationship with state test scores. These relationships are moderate to strong and were statistically significant at the $p < .001$ level. The percent correct score for all of the items had the strongest relationship with state assessment scores. However, the proficiency estimates from the 2PL_AllItems models also had relatively strong statistical relationships with the criterion and were only slightly less than that of the percent correct score for all items. The same was true for the ORMs which had coefficients that were marginally less than those found for the 2PL_AllItems models. Interestingly, when the same

sets of items (all items or main items only) are calibrated using a 2PL IRT model, the proficiency estimates obtained from these calibrations correlate slightly less with the criterion than their corresponding percent correct scores.

Table 19.

Correlation Coefficients between Scaled Scores Obtained from each Scoring Model, Percent Correct Scores and State Test Scores

Model Type	State Test Score	Perc Corr_ Main Items	Perc Corr_ All Items	2PL_ Main Items	2PL_ All Items + covs	2PL_ All Items	ORM + covs	ORM	TRM + covs
PercCorr_ MainItems	0.597								
PercCorr_ AllItems	0.625	0.926							
2PL_ MainItems	0.500	0.905	0.835						
2PL_ AllItems + covs	0.606	0.883	0.899	0.908					
2PL_ AllItems	0.606	0.884	0.900	0.908	1.000				
ORM + covs	0.592	0.924	0.895	0.927	0.969	0.969			
ORM	0.591	0.923	0.894	0.927	0.968	0.968	1.000		
TRM + covs	0.017	0.127	0.153	0.108	0.111	0.112	0.611	0.611	
TRM	0.009	0.110	0.132	0.087	0.090	0.090	0.622	0.622	0.863

Note. Light grey cells indicate coefficients significant at $p \leq .05$; darker grey cells indicate coefficients significant at $p \leq .001$

To facilitate the interpretation of these correlation coefficients, scatterplots depicting each of the scoring metrics against the state test scores are provided in Figures 22 – 26 below. While the correlation coefficient between the percent correct scores for all Assistentments items and the criterion was stronger than those found for the scaled scores, this may be a result of a relatively small number of students that had high percent correct scores and also performed

extremely well on the state test. That is to say, this correlation appears to be an average of two groups of students: one group that performed poorly on the state test and had a wide range of percent correct scores (i.e., practically no relationship), and another group whose percent correct scores appeared to be strongly related to state test scores. On the other hand, the relationships between the scaled scores and the criterion appeared to be relatively consistent; there did not appear to be two different groups of students. While there were certainly more students that were below average on the ability scale who also scored low on the state test, there were not many students that were high on the ability scale but scored low on the state test (as was the case for the percent correct scores). Overall, while the coefficients for the scaled scores and the criterion were smaller, there was less dispersion than the relationships with the percent correct scores suggesting that the latter correlations may be driven by a relatively small group of students that performed well on both measures.

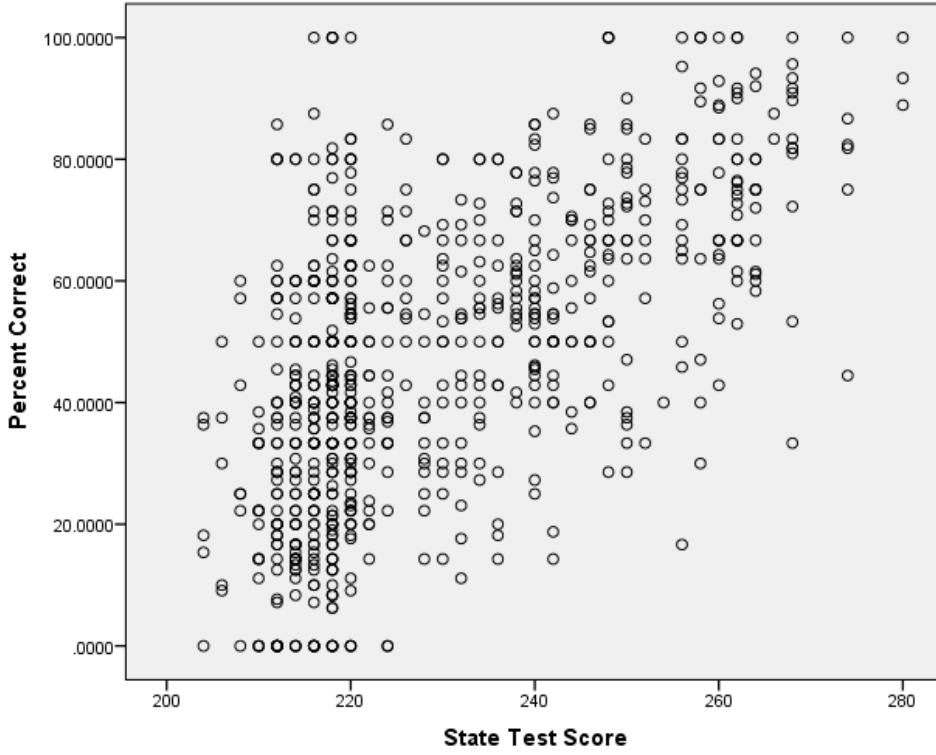


Figure 22. Scatterplot of percent correct scores on main items only and state test scores.

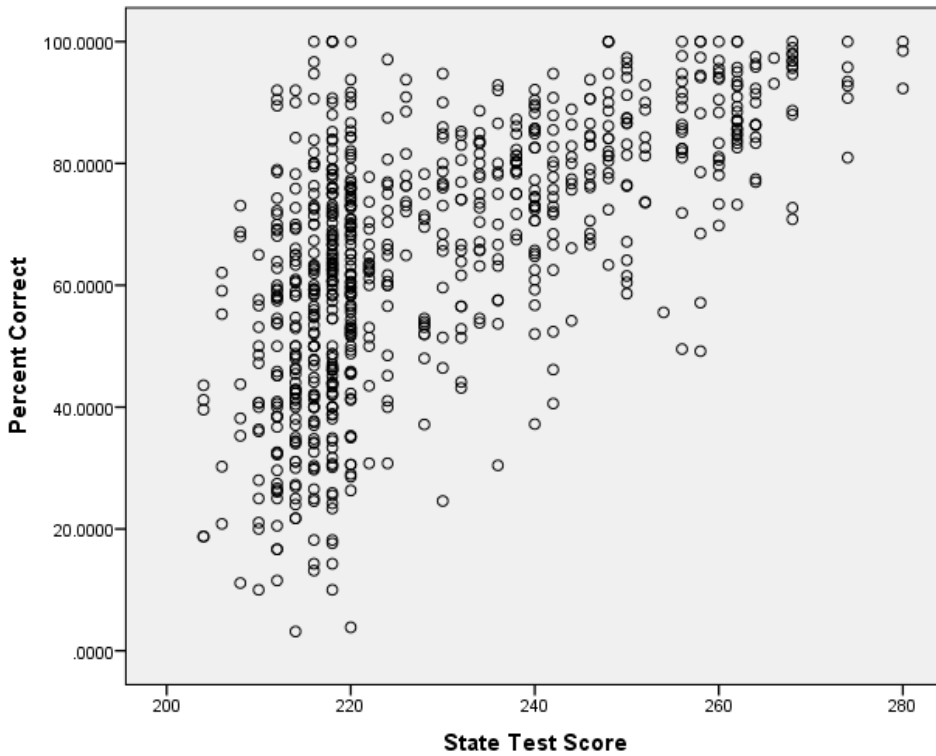


Figure 23. Scatterplot of percent correct scores on all Assistments items and state test scores.

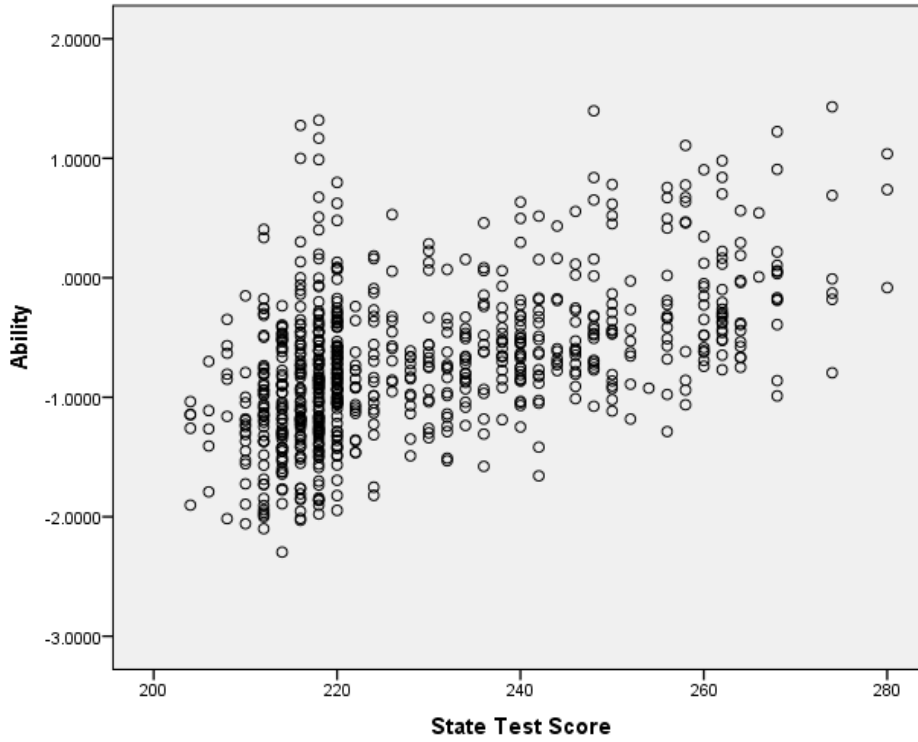


Figure 24. Scatterplot of scaled scores from the 2PL_MainItems model and state test scores.

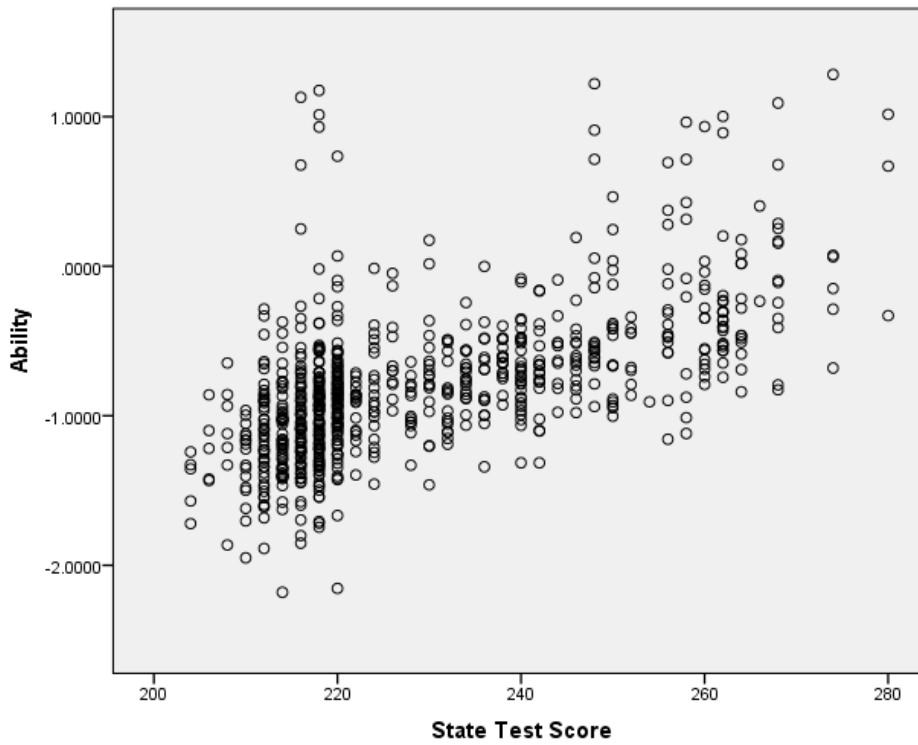


Figure 25. Scatterplot of scaled scores from the 2PL_AllItems model and state test scores.

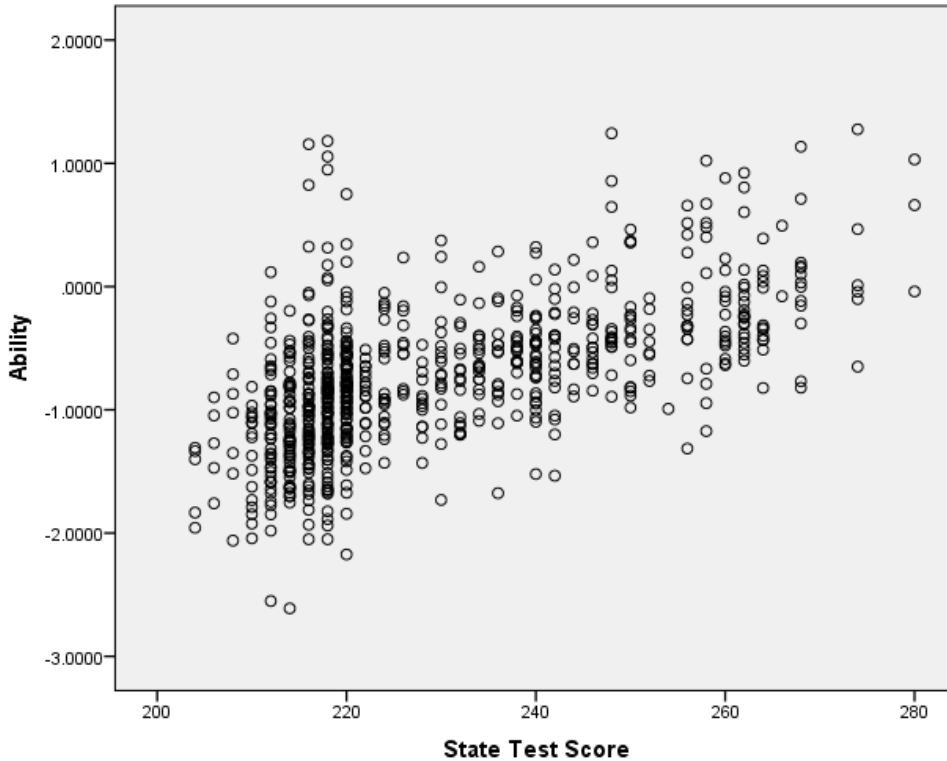


Figure 26. Scatterplot of scaled scores from the ORM and state test scores.

Summary. Overall, except for the TRMs, the scaled scores from each of the scoring models and the percent correct metrics are strongly correlated with each other and moderately correlated with state test scores. The correlation coefficients between the scoring metrics that were related to state test scores were all approximately equal to $r = 0.6$ with the exception of the 2PL_MainItems model which was somewhat lower. While the coefficients for the percent correct scores were stronger, they appear to have been driven mostly by a group of students that performed well on both metrics. While the TRMs were not necessarily expected to correlate well with any of the other metrics due to the model calibration issues outlined in a previous section, it was of interest to note that proficiency estimates from these models only correlated well with the other models that also accounted for local dependence (i.e., the ORMs).

Chapter Five – Discussion

While significant progress has been made in recent years on technology enabled assessments (TEAs), including assessment systems that incorporate scaffolding into the assessment process, the area of psychometrics has yet to venture directly into these advancements in technology to determine how statistical methods and procedures can be used to fully capture students' knowledge, skills and abilities as measured by TEAs (Almond et al., 2010; Bechard et al, 2010; Bennett & Gitomer, 2009). This exploratory investigation has contributed towards this advancement by providing a comparison of scoring models for an operational scaffolded assessment system, the Assisments system, and by evaluating the statistical relationships between scores derived from those models and a criterion measure of student ability.

Two main research questions were addressed in this study. The first research question was aimed at determining which type of scoring model is the optimal model for scoring scaffolded assessment data from the Assisments system. To address this question, a sequential procedure for fitting and evaluating increasingly complex models was conducted. The 2PL_MainItems model was established and compared to three additional comparison models; the 2PL_MainItems model did not account for any of the scaffolding features or complexities in the dataset whereas the comparison group of models did, each in a different way. The 2PL_MainItems model only accounted for the independent dichotomous responses to the main items and was calibrated using the 2PL model. The 2PL_AllItems comparison model additionally calibrated all of the scaffold items but ignored local dependence that is created by the scaffolding process. The polytomous ORM accounted for local item dependence by treating the response patterns of item bundles as categories of a polytomous

item. Finally, the TRM accounted for local dependence within bundles by adding a random effect component to explain the interaction between the person and the bundle. All three of the comparison models were evaluated twice; once with the average number of hints accessed for a student and for an item as covariates for both person and item parameters in the estimation process and once without using any covariates. A total of seven scoring models were calibrated and evaluated with respect to model convergence, model fit, and test information.

The second research question evaluated one aspect of the validity of interpreting scores from these seven models by determining the degree to which scores relate to subsequent performance on an end-of-year accountability assessment. The scoring models were compared to one another to determine which model produced scores that most strongly correlated with students' state test scores. The scoring models were also evaluated against a percent correct score to determine if student ability estimates calibrated from IRT models had stronger relationships to a criterion measure of student ability than the percent correct score.

The results of the analyses for each research question are discussed in the following sections. Model convergence issues and implications of non-convergence are discussed first followed by a summary of model estimates obtained from each model. A comparison of model fit statistics (where appropriate) and item information is also discussed. A summary of the statistical relationships between all of the scoring models and the criterion measure is also discussed. The models are then summarized and an optimal model is recommended based on a comparison of the criteria discussed and practical advantages and disadvantages of implementing a scoring model. Limitations, future research directions and conclusions end this chapter.

Research Question 1: What type of model is the optimal scoring model for the scaffolded data in the Assistments system?

The Assistments data were calibrated according to seven different models each of which accounted for different features of the Assistments system and/or the nature of the data. Each of the seven scoring models calibrated in this study were assessed according to parameter convergence, model fit, and precision of model estimates. In addition to a discussion regarding parameter estimates from each model, issues, results and implications associated with each of the evaluation criteria are discussed next.

Convergence. One of the main challenges in using MCMC algorithms is the inherent difficulty of assessing the degree to which those algorithms have converged (Sinharay, 2004). Convergence is necessary if one is to assume that the sample generated from the MCMC algorithm is representative of the posterior distribution of interest (Sinharay, 2004). In this study, convergence was statistically assessed using the PSRF convergence criterion of Gelman and Rubin (1993). The PSRF was calculated for each estimated item parameter at the 50% and 97.5% quantiles based on the Student t distribution.

For the 2PL_MainItems model, convergence was readily achieved at both quantile points for parameters a and b . While the 2PL_MainItems model required relatively few iterations (3,000 or less), the 2PL_AllItems model (which additionally calibrated the scaffold items) required more than three times as many iterations to achieve convergence when covariates were included in the estimation process and more than six times as many iterations when covariates were not included. This was not surprising given that the scaffold items accounted for 108 additional items; more than four times as many items as the 2PL_MainItems model. However, it is interesting to note that incorporating the covariate for

the b parameter appeared to facilitate the estimation process for the 2PL_AllItems model. The model that did not incorporate these covariates did not meet the convergence criterion after the same number of iterations as the model that did incorporate them; the additional iterations needed for the model without covariates resulted in a less efficient estimation process with respect to time needed for completion. It should be noted that only the b parameter at the 97.5% quantile did not meet the convergence criterion of 1.2 or less and the difference between ~ 1.3 and 1.2 may not be great enough to justify the additional iterations. Using other means for assessing convergence (e.g., graphical displays of stability) may result in less conservative decisions regarding attainment of convergence. Similar results were found for the polytomous ORMs. That is, the model with covariates achieved convergence after 10,000 iterations while the model without covariates required additional iterations to successfully converge on the b parameter.

Of the models discussed thus far, the 2PL_MainItems model was not surprisingly, the most efficient in terms of time needed to achieve convergence. However, clearly this model was the most simplistic and did not account for any of the scaffolding features. Of the models that accounted for the scaffold items, the 2PL_AllItems models, which ignored local dependence, took less time to converge than the ORMs, which accounted for local dependence. For both of these types of models, the models that incorporated covariates into the estimation process appeared to be more efficient than their counterpart models that did not incorporate any covariates.

The TRM calibration process was vastly different from the other models with respect to number of iterations needed and convergence criteria. Firstly, the TRMs were calibrated multiple times, each time with additional iterations up to 100,000 iterations. Each model run

was also conducted using parameter estimates from the 2PL_AllItems model as initial values. After 100,000 iterations, the item parameters, a and b , met the convergence criterion; however, distributions for the variances of gamma did not appear to converge. Three bundles did not converge at the 50% quantile for the model with covariates and one bundle did not converge for the model without covariates. Most bundles did not meet convergence criterion at the 97.5% quantile for both models. Similar to the other models, the b parameter appeared to converge more quickly for the model that incorporated a covariate in the estimation process for this parameter. Conversely, based solely on the number of bundles that did not converge at the 50% quantile, the covariate for theta appeared to hinder the estimation process. In other words, the model that incorporated covariates in the estimation process had two more bundles that did not converge than the model that did not incorporate covariates. While this evidence is far from substantial, it may suggest that using the number of hints a student accessed during the assessment process as a covariate for theta may not be useful for the TRM. The value of this covariate is further discussed in a later section.

Based on the convergence results for the TRMs, it appeared that some estimates of gamma were unreasonably large (i.e., approached infinity) which was inevitably making convergence an unattainable goal. A series of additional steps and analyses were conducted to try to rectify this problem unfortunately, with no avail. While a solution was not found, these steps provided some potentially helpful insight into the problem. For instance, it was confirmed that the problem did not stem from one or two “troublesome” bundles. Each time a bundle was removed that was associated with extremely large gammas, another bundle would produce “infinite” gammas in the place of previously “normal” estimates. Possible explanations for these seemingly infinitely large testlet effects may reside in the nature of the

Assistments system itself and the manner in which data were structured for this study. These suggested explanations are provided in the next section.

The amount of time required for the TRMs to complete was more than eight times that of the polytomous model approach and even then, it did not successfully converge. Furthermore, assessing convergence based on multiple chains was more challenging than the other models due to the fact that for chains with very large number of iterations, each chain had to be run independently and PSRF values had to be calculated from output of each. Clearly the complexities associated with this type of model contributed to the convergence issues encountered in this study. In any case, using the output of an MCMC algorithm that has not converged may lead to incorrect inferences about the model and its utility (Sinharay, 2004). A summary of parameter estimates was presented in the results portion of this study and is discussed below in order to assist in suggesting possible explanations for the findings. However, interpreting these parameters for the purposes of making decisions about scoring models is *not* warranted.

Overall, as the model increased in complexity, the amount of time required for convergence also increased. All other model criteria aside, the value of accounting for the scaffolding features would need to be weighed against the time needed for calibration in order to determine which model to choose. In other words, while ignoring the scaffolding features in the Assistments system can result in a model that is quick and easy to calibrate in IRT, the value of accounting for those features may be more evident in assessing the precision of students' scaled scores obtained from those models.

Descriptive Statistics. Based on the percent correct scores for all of the items, many items were relatively easy for students. There were also a large number of items that

appeared to be more difficult for most students. Overall, the average percent correct score on main items was about 70% which is probably fairly typical for a low stakes assessment that is intended to be formative in nature. In other words, teachers are presumably providing instruction on assessed concepts either directly prior to the assessment or even throughout the process. The environment in which students are taking these assessments inevitably impacts their performance and items that are more difficult in a standardized testing environment may appear to be easier in a non-standardized environment.

Item difficulty parameter estimates for each of the evaluation models suggested that in general, items required a below average ability level to have a 50% chance of correctly responding to an item. The average difficulty parameter estimate for all of the models was below zero. Again, this was expected given that average percent correct score across items was fairly high. As the models increased in complexity, the average b parameters decreased such that the ORMs had a greater negative value than the 2PL_AllItems models which had a greater negative value than the 2PL_MainItems model. A comparison of average b parameters from the 2PL_MainItems model to the 2PL_AllItems models suggests that, on average, less ability was needed to correctly respond to the full set of items than what was needed, on average, to respond to the main items alone. The scaffold items in general, were easier than the main items, which is an expected finding in that the scaffold items are by nature subcomponents of the main items. The differences between b parameters from the ORMs and those from the 2PL_AllItems models were slight. On the other hand, the average b parameter for the TRMs was much lower than any of the other models indicating that a much lower level of ability was required to answer the items correctly. This may imply that after accounting for a potentially large “bundle ability parameter”, the amount of “overall”

ability required to correctly respond to an item is minimal. This explanation is also consistent with the large estimates of the variances of gamma. These findings may indicate that ignoring local dependence inflates b parameter estimates. However, this is not consistent with previous research that has found estimation of b parameters to be relatively stable regardless of whether or not local dependence is accounted for (Wainer & Wang, 2000; Wang & Wilson, 2005). Given the relatively typical testlet contexts described in the previous research, it is not clear how b parameters would be impacted in the context of extremely large amounts of local dependence.

Item discrimination parameter estimates for each model were, on average, higher than what is typically expected for any assessment but particularly unexpected given the relatively low difficulty parameter estimates. As the typical range of discrimination parameters is from 0.0 to +2.0, and they are rarely greater than 2.0 (Hambleton, Swaminathan & Rogers, 1991), average a parameters of 2.8 and 3.8 in the 2PL_MainItems model and 2PL_AllItems models, respectively, are quite surprising. The increase in average a parameters from the 2PL_MainItems model to the 2PL_AllItems model appears to suggest that the scaffold items also discriminate students very well. When local dependence is accounted for in the ORMs and TRMs, these estimates decrease and the average discrimination parameter estimates are within the range that is typically found. Wainer & Wang (2000) investigated changes in item parameters from a dichotomous model to a testlet model and found that for some testlets, a parameters were over-estimated when local dependence was ignored and for other testlets, discrimination was under-estimated. These researchers suggested that testlet characteristics may help explain these differences. In the present study, all of the discrimination parameters were larger when local dependence was ignored than when it was accounted for. Future

research that describes how characteristics of testlets impact these parameters may help provide an explanation for these findings.

While summaries of the parameter estimates were presented for the models with covariates and the models without covariates, these estimates were not meaningfully different. That is, adding the covariate into the estimation process of the b parameter did not appear to meaningfully impact the resulting estimates. This is to be expected if model convergence was achieved. Since covariates were brought into the model via the mean of the prior distribution of the item parameters (Wang, Bradlow & Wainer, 2005), the posterior expectations (means) are unaffected after the distribution of the parameter has stabilized.

As mentioned previously, estimates of gamma and variances of gamma were strikingly large. The smallest variance of gamma was approximately 26 and the largest variance was a number to the power of 73. Again, the distributions of the variances of gamma did not converge, it is not appropriate to make conclusions about these parameters; however, it is clear that the testlet effects account for a substantial amount of local dependence in the data. Possible explanations for this finding may lie in the nature of the assessment system from which the data were collected. A typical testlet situation is one in which a stimulus or stem is presented to a student such as a reading passage, which is followed by a series of items that are related to the particular stem. Thus, the items that refer and relate to the same stem are not locally independent of one another and a testlet model may be appropriate. The Assistments' scaffolding process creates bundles of items that are clearly not locally independent; however, this process may create even more dependencies in the data than the typical testlet situation for a couple of reasons. First, responses to the scaffold items are dependent on responses to the main items. That is, a correct response to a main item

automatically directs the student to the next item without going through the scaffolding process. For the purposes of this study, the data were re-coded such that scaffold items were assigned correct scores for every correct response to a main item. While theoretically, this re-coding is justifiable it may also have created a false inflation of local item dependence. Furthermore, response patterns that did contain variation inevitably had a zero score at the beginning of each sequence or bundle. In other words, only students that responded incorrectly to the main item were directed through the scaffolding process. There may also be students that even though they know the correct answer to the main item, choose to break the item into steps and go through the scaffold items. These students could contribute to further dependency in the data if they answer all of the scaffold items correctly even though they technically received an incorrect response to the main item. In any case, responses to items in the Assistments system depend on previous responses to items which undoubtedly impacts the amount of local dependence between items.

Second, the main items may be considered the “stem” or common stimulus to which the scaffold items refer or relate to, not unlike the typical testlet situation. However, unlike the typical testlet, the scaffold items are actually parts of the main item or stem. In other words, the scaffold items break down the main items into subcomponents to try and decrease the cognitive load on the student. In a sense, the scaffold items simply repeat different parts of the main item. Thus, the degree to which the content of the main items differs from the scaffold items is minimal; the same is true for differences between scaffold items within a bundle which also potentially contributes to a greater amount of local dependence than what is found in a typical testlet situation. The combination of multiple levels of dependence (i.e., content-based dependency, response-based dependency) coupled with the imputation of

correct response patterns to scaffold items for those that responded correctly to a main item, most likely can account for the testlet effect sizes seen in this study.

Model Fit. The model fit statistic used in this study was the DIC which is defined by two terms that represent model deviance and model complexity. Smaller values of DIC indicate better-fitting models; therefore, both of the models that did not incorporate covariates into the estimation process fit the data better than their counterpart models that did incorporate covariates. This finding was somewhat surprising as previous research has demonstrated that adding covariates to a model can decrease the pD and the DIC as a result of the covariates explaining a substantial amount of variation in the model (Spiegelhalter, et al., 1998). This suggests that the covariates used in the estimation process for item difficulty parameters and thetas simply do not help explain variations in model parameters.

The pD terms for model complexity were, not surprisingly, vastly different across the models. As explained previously, while the models are based on the same sample of students, the data were restructured or recoded according to each type of model. For the purposes of evaluating a model that did not account for scaffolding features, the 2PL_MainItems model only consisted of responses to main items while the 2PL_AllItems further accounted for the 108 additional scaffold items. For the ORMs, responses from all items were recoded into summed scores for each bundle. Thus, each type of model was calibrated on different “versions” of the same dataset. Thus, the prior information provided for each model calibration may have interacted differently with each restructured dataset. While the pD value was the largest for the 2PL_MainItems model, it most closely resembles the true number of effective parameters in the model (θ 's for each student plus item parameters). In other words, this is the amount of information that is expected to be lost when the point estimates are used

as expectations of the saturated model given the number of parameters being estimated (Spiegelhalter, et al., 1998; Spiegelhalter, et al., 2002). However, when the scaffold items are added to the calibration process for the 2PL_AllItems models, it appears that there may have been a strong conflict between the specified priors (which were the same for the previous model) and the data. As Spiegelhalter, et al. (2002) point out, negative pD values typically indicate data/prior conflict or use of a poor estimator of the distribution. Thus, the prior information regarding the scaffold items may need to be specified differently to decrease the compromise with the data.

Finally, when local dependence is accounted for and item bundles are scored polytomously, the pD value is positive but smaller than expected given the additional threshold parameters estimated for the score category functions. Spiegelhalter, et al., (1998) found that when covariates were added to a model, the pD decreased as a result of the covariates explaining a substantial amount of variation in the model. Thus, when the model fully explains the data, the pD value is small. In this respect, it may be that the ORMs fit the data quite well since model complexity appeared to be underestimated. However, it is not possible to draw conclusions about model fit without a directly comparable model using the same dataset.

Overall, based on the DIC statistics derived for the groups of models that were comparable, the models without covariates fit the data better than those with covariates. While this model evaluation criterion provided information regarding the utility of the covariates, or lack thereof, it did not assist with decisions regarding the various model types.

Information. The main finding from the item and test information functions for the various scoring models was that the 2PL_AllItems models provided considerably more

information when theta was between -2.0 and zero than the 2PL_MainItems model or the ORMs. The ORMs, on the other hand, provided more information than the other two types of scoring models at the lower end of the ability scale. This suggests that when the bundles of dichotomous items were scored polytomously, they tended to be more precise for very low performing examinees than when the bundles were not scored polytomously. However, for examinees that were just below average, the polytomous scoring of bundles reduced the amount of information that could be provided. For many bundles, the amount of information provided when the bundles were scored polytomously was less than or the same as the amount of information provided by the main items alone.

The fact that information for the ORMs was generally less than the 2PL_AllItems models was not surprising. Previous research has demonstrated that when dichotomous items in a testlet are summed and scored polytomously, information in the exact response patterns is lost which results in a lower information curve (Keller, Swaminathan & Sireci., 2003; Wainer & Wang, 2000; Wang, Bradlow & Wainer, 2002). While items that are designed to be scored polytomously typically contain more information than items designed to be scored dichotomously, a testlet formed from several dichotomous items may not have as much information as the set of dichotomous items across the ability scale (Keller, Swaminathan & Sireci., 2003). This has been a main criticism of the polytomous approach to accounting for local dependence within bundles of items and it appears to hold true for the present study as well. At the same time, research has also demonstrated that fitting standard item response models to groups of interdependent items may result in an overestimation of test information (Wang & Wilson, 2005a; Wang, Cheng & Wilson, 2005). Thus, while some amount of information is expected to be lost when groups of dichotomous items are scored polytomously,

it is not clear how much is actually lost relative to artificially inflated test information obtained using the dichotomous approach.

The solution to this problem that is typically offered in the research is the testlet model approach which utilizes exact response patterns *and* accounts for local dependence (Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000; Wainer & Kiely, 1987; Wainer & Wang, 2000). While theoretically this solution should apply quite well to the current context, unfortunately, the testlet model was not estimable using the Assisments data.

Research Question 2: Is there a relationship between student ability estimates derived from the scoring models and a criterion measure of student achievement?

The scaled scores from each of the scoring models as well as percent correct scores were assessed against a criterion measure of student ability. Statistical relationships between the different scoring metrics and students' corresponding state assessment scores indicated that the percent correct score on all of the items yielded the strongest correlation with the criterion. However, scaled scores from the 2PL_AllItems model correlated almost as strongly with state test scores as the percent correct score. In fact, the coefficients for all of the scoring models, except the 2PL_MainItems and the TRMs, were within rounding error of one another.

While scores from the 2PL_MainItems model had a relatively strong relationship with state test scores, it was weaker than the other scoring metrics. Similarly, the relationship between 2PL_AllItems scores and the criterion was slightly weaker than the relationship between the percent correct score on all the items and the criterion. The comparisons between the percent correct scores and the 2PL models were somewhat surprising as one of the main benefits of using IRT is the ability to account for item parameters such as difficulty and discrimination in the calibration process. Furthermore, research on Assisments has

specifically demonstrated that accounting for problem difficulty provides more efficient estimates of student ability which ultimately leads to more accurate predictions of performance on the state accountability test than when percent correct scores were used to make these predictions (Ayers & Junker, 2008). It is possible that if prediction models were specifically derived using this data, results and conclusions drawn from those results may be different. Correlation coefficients are limited as statistical indices for validity evidence in that they represent an average prediction based on content or skill similarity. However they don't reflect the accuracy of identifying students that will pass or fail the criterion measure of ability. Moreover, there appeared to be a relatively small group of students that had high percent correct scores and high state assessment scores that was driving this higher than expected correlation. This averaged with the other group of students that did not appear to have a relationship between these measures produced a moderately strong correlation for the entire sample. In general, the relationships between the scaled scores and the criterion did not produce this pattern; rather they were consistently moderate for the entire sample which lends support for the use of IRT calibration. That is, accounting for item parameters provides a more accurate measurement of student ability that is independent of the items from which it was calibrated.

As the statistical relationships between scores from each of the models correlate approximately equally well with the criterion measure, selecting or rejecting a model based solely on this information is not necessarily warranted. Again, this information should be used in conjunction with the other measures of model adequacy and precision to determine the most appropriate model for the Assisments data.

Model Summary & Selection

In order to determine the optimal scoring model for the Assistments system, it is useful to revisit the purpose and goals of the system. The Assistments system is a formative learning tool that is intended to provide students and teachers feedback regarding students' strengths and weaknesses while simultaneously offering instructional assistance throughout the assessment process. All of which is intended to help students achieve proficiency on the end-of-year accountability test in mathematics. As such, the Assistments system needs a scoring paradigm that accounts for student performance as well as the assistance that a student needed during the assessment process that can be used to predict performance on the state assessment.

The 2PL_MainItems model was clearly the most efficient model to calibrate with respect to the time needed to converge. However, this model did not account for any scaffolding features, it provided the least amount of total test information across the ability scale, and scaled scores obtained from this model had the weakest relationship with the criterion measure of student ability relative to the other scoring models. The 2PL_AllItems models were the second most efficient models to calibrate in terms of convergence time. These models accounted for the scaffold items but ignored local dependence that is created by the scaffolding process. However, it provided the most total test information and produced scores that had the strongest relationships with the state test scores. The ORMs were the second least efficient models to calibrate with respect to successful model completion and convergence. These models most accurately represent the nature of the data in that they accounted for the scaffold items and the local dependence that is created by the scaffolding

process. The ORMs came in second to the 2PL_AllItems models with regards to the amount of information they provided and the strength of relationships their scaled scores had with the criterion.

The TRMs were clearly not the optimal models for this particular dataset.

Theoretically, it was presumed that these models might have been the optimal models given their ability to account for local dependence without losing response pattern information.

However, these models can become quite complex to calibrate as more testlets are added to a model which may potentially render them impractical to use in operational contexts that involve a large number of testlets. Even had the model converged after 100,000 iterations, it would have been by far the least efficient model in terms of the time required for calibration. Furthermore, while parameter estimates based on models that did not successfully converge may not be reflective of true parameters, theta estimates that were provided by these models had no relationship with the criterion. Although as noted above, making these comparisons is not necessarily appropriate.

It was also determined that using the average number of hints for persons and items as a covariate in the estimation process of the item difficulty parameter and person parameters was not that useful. While the number of hints for items appeared to facilitate convergence of the difficulty parameters, overall the models that incorporated covariates fit the data worse than those that didn't. Moreover, using the average number of hints as a covariate for person parameters, did not strengthen the relationships between these parameters and the criterion. It was decided that from a practical standpoint, using the number of hints as a covariate in the estimation process was more effort than any potential benefit; therefore, these models were removed from the list of models to consider.

In comparing the remaining models, the Figure 27 below rank ordered each of the model evaluation criteria relative to the other models. Model fit evaluations were not included as these comparisons were only appropriate between the covariates versus non-covariates models. Since there were three remaining models, each criterion was ranked from one to three to show which model was the “best” and which was the “worst”. For example, the model that provided the most information was ranked a three while the model that provided the least was ranked a one. The 2PL_AllItems model was ranked the highest with respect to the amount of information it provided and the strength of relationship it had with the criterion measure of student ability. While it inevitably took longer to converge than the 2PL_MainItems model, the value of accounting for the additional scaffold items appears to be worth the extra calibration time. Therefore, based on these evaluation criteria the 2PL_AllItems model was determined to be the optimal scoring model for the Assisments data relative to the other scoring models evaluated in this study. However, it should be noted that the difference between the correlations with the criterion for the 2PL_AllItems and the ORM were practically indistinguishable. Furthermore, while information is inevitably lost in the polytomous approach, the loss may be worth the benefit if the violation against the local independence assumption is deemed unacceptable. Therefore, the ORM may be a valuable alternative.

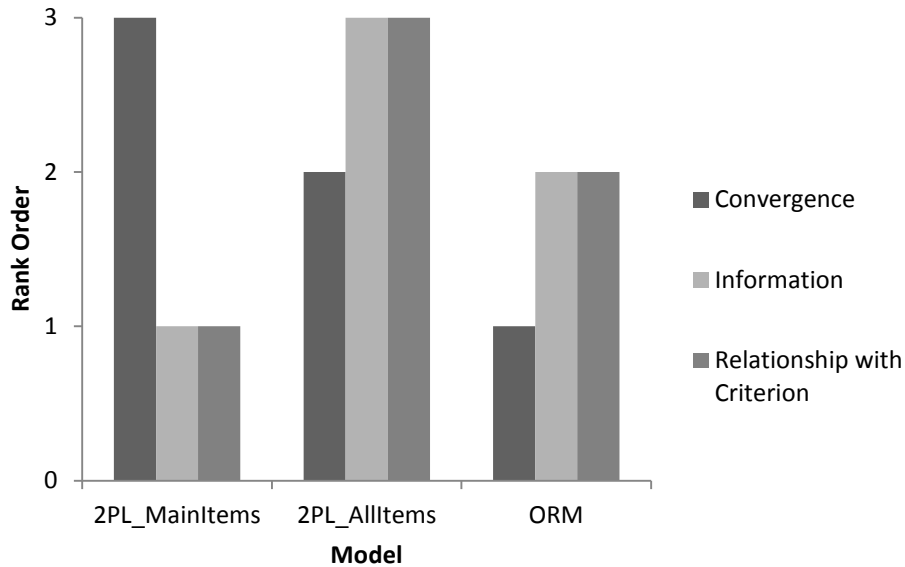


Figure 27. A rank ordered comparison of models by each evaluation criterion. Convergence = time required for model to successfully complete and converge; Information = total test information; Relationship with Criterion = based on correlation coefficients which were almost indistinguishable between the 2PL_AllItems model and the ORM.

Given the purpose of the Assistment system which is a formative tool that also provides instructional opportunities to students during the assessment process, the advantages of applying any of these scoring models from a measurement perspective may not justify the practical disadvantages. For instance, the percent correct score may be completely dependent on the specific items that a student took but it is relatively simple to understand and compute. On the other hand, scaled scores from an IRT model are independent of the items from which they were calibrated from, but ability estimates are more complex to understand and derive. As the Assistments system is a low stakes environment that is mostly geared towards learning, the benefits of the scoring models presented in this study need to be weighed against the practical constraints in an operational setting with respect to time, cost and resources.

Limitations & Future Research

There are several limitations to this study, in addition to those already mentioned, that are worth discussing. First, the results of this study are highly dependent on the specific nature of the Assistments system and cannot necessarily be generalized to other types of scaffolded assessments. The models chosen to be evaluated in the present analysis were based on the match between the characteristics of the Assistments data and the theoretical framework that supports those models. Other assessment systems that incorporate scaffolds may be characterized very differently and thus, may not be appropriate for the models used in this study. While it is reasonable to think that other assessment systems that share similar scaffolding features could benefit from the types of models presented here, it is not possible to make those generalizations without an empirical investigation. Furthermore, given the formidable estimation issues associated with applying the testlet response model to the Assistments data despite the theoretical consistency between the data and the model, it is certainly worth examining the use of this type of model in other similar contexts.

Second, and similar to the limitation noted by Bolt & Lall (2003) in their evaluation of competing IRT models, the model comparison approach conducted in this study assumed that one scoring model would be optimal for all items or bundles and for all students. Particularly given that the results of the information functions which were somewhat dependent on location of the ability scale, it is conceivable that the optimal “model” is a combination of scoring models that maximize information across the ability scale for every bundle of items. Of course, the practical constraints in this scenario may far outweigh the benefits. In any case, it should not be assumed that the chosen scoring model is the optimal one for every

situation. A mixtures model approach may actually provide the optimal solution and may be an interesting direction for future research.

Finally, an ongoing concern throughout this study was related to the reliance on others' data. The data were obtained from a pre-existing database based on an operational assessment system. While such assessment systems can provide a wealth of data and information for a variety of purposes, using others' data resources also increases the chances for misunderstandings about the data and difficulty in interpreting the data. Even though every effort was made to ensure the data were cleaned and well understood, it cannot be determined with absolute certainty that this was the case.

Conclusions

Overall, the goal of model selection to identify the least complex model that adheres to the purpose of the assessment system and adequately accounts for the essential features of the dataset (Pitt, Kim & Myung, 2003). The Assistments system is a formative learning tool that is intended to help teachers gauge student progress towards the state accountability assessment while simultaneously providing students opportunities for additional instruction. As such, the costs in terms of time, resources and interpretability of employing a more complex model need to be considered against the benefits associated with a given scoring model. For instance, the percent correct score is indeed the simplest indication of student performance; however, the benefits of applying an IRT model to assessment data often outweigh the simplicity of such measures. On the other hand, the local dependence that is inherent within the scaffolding process may be considered an essential feature that needs to be accounted for within a more complex scoring paradigm such as a polytomous model.

However, given the increased time, effort and complexity of the model, the costs associated with applying a polytomous model may outweigh the benefit of accounting for that feature.

In this study, the dichotomous model that accounted for both the main items and the scaffold items but ignored local dependence was identified as the optimal model. This selection was made on the basis of several criterion including relatively efficient model calibration time, maximum information for most ability levels, and a relatively strong relationship between its scaled scores and a criterion measure of student ability. While this model does not account for all of the scaffolding features within the Assistments system and it ignores an assumption made by the model, it appears to be the simplest model that remains consistent with the purposes of the system. A scoring model that does not account for the scaffold items is one that arguably ignores an essential feature of the Assistments and provides the least precise estimates of student ability. The additional complexity that accompanies the inclusion of the scaffold items appears to be worth the cost for this assessment system.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* (19), 716-723.
- Andrich, D. (1985). A latent-trait model for item with response dependences: Implications for test construction and analysis. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 245-275). New York: Academic Press.
- Anozie, N. O., & Junker, B. W. (2006). Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. In *Proceedings of the American Association for Artificial Intelligence Workshop on Educational Data Mining (AAAI-06)*, Boston, MA. Menlo Park, CA: AAAI Press.
- Ayers, E., & Junker, B. W. (2006). Do skills combine additively to predict task difficulty in eighth grade mathematics? In *Proceedings of the American Association for Artificial Intelligence Workshop on Educational Data Mining (AAAI-06)*, July 17, 2006, Boston, MA (Technical Report WS-06-05, pp. 14-20). Menlo Park, CA: AAAI Press.
- Ayers, E., & Junker, B. (2008). IRT modeling of tutor performance to predict end-of-year exam scores. *Educational and psychological measurement*, 68(6), 972-987.
- Almond, P., Winter, P., Cameto, R., Russell, M., Sato, E., Clarke-Midura, J., Torres, C., Haertel, G., Dolan, R., Beddow, P., & Lazarus, S. (2010). Technology-Enabled and Universally Designed Assessment: Considering Access in Measuring the Achievement of Students with Disabilities—A Foundation for Research. *Journal of Technology, Learning, and Assessment*, 10(5). Retrieved January 1, 2011 from <http://www.jtla.org>.
- Arter, J. (2003). *Assessment for Learning: Classroom Assessment To Improve Student Achievement and Well-Being*. U.S. Dept. of Education: Educational Research Information Center.
- Azevedo, R., Cromley, J. G., & Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia?* 1. *Contemporary Educational Psychology*, 29(3), 344-370.
- Bechard, S., Sheinker, J., Abell, R., Barton, K., Burling, K., Camacho, C., Cameto, R., Haertel, G., Hansen, E., Johnstone, C., Kingston, N., Murray, E., Parker, C.E., Red_eld, D., and Tucker, B. (2010). Measuring Cognition of Students with Disabilities Using Technology-Enabled Assessments: Recommendations for a Research Agenda. *Journal of Technology, Learning, and Assessment*, 10(4). Retrieved January 1, 2011 from <http://www.jtla.org>.

- Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2007). Problem Solving in Technology-Rich Environments. A Report from the NAEP Technology-Based Assessment Project, Research and Development Series. NCES 2007-466. *National Center for Education Statistics*, 196.
- Bennett, R. E., & Gitomer, D. H. (2009). Transforming K–12 assessment: Integrating accountability testing, formative assessment and professional support. *Educational assessment in the 21st century*, 43-61.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. R. Novick (Eds.), *Statistical theories of mental scores* (pp. 395-479). Reading, MA: Addison-Wesley.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in education*, 5(1), 7-74.
- Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1), 29-51.
- Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, 27(6), 395-414.
- Box, G. E. P. 1976. *Science and statistics*. J. Am. Stat. Assoc. 71:791-799.
- Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64, 153-168.
- Brown, J., Hinze, S. & Pellegrino, J. W. (2008). in 21st Century Education, T. Good, Ed. (Sage, Thousand Oaks, CA, 2008), vol. 2, chap. 77, pp. 245–255.
- Bruner, J. (1985). Vygotsky: A historical and conceptual perspective. *Culture, communication, and cognition: Vygotskian perspectives*, 21-34.
- Bruner, J. (1986). *Actual minds, possible worlds*: Harvard University Press.
- Burnham, K. P., and D. R. Anderson (2003). *Model selection and mul-timodel inference: A practical information-theoretic approach*, 2nd ed. Springer-Verlag, New York.
- Cagiltay, K. (2006). Scaffolding strategies in electronic performance support systems: types and challenges. *Innovations in education and Teaching International*, 43(1), 93.
- Camacho, C. (2009). *CPAA Implementation for the State of Mississippi: CPAA State Date Report for Spring 2009*. Children’s Progress: Mississippi.

- Cazden, C. B. (1979). Language in education: Variation in the teacher-talk register. *Language in public life*, 144-162.
- Chang, K. E., Sung, Y. T., & Chen, S. (2001). Learning through computer based concept mapping with scaffolding aid. *Journal of Computer Assisted Learning*, 17(1), 21-33.
- Children's Progress, Inc. (2009). *Children's progress: Insights for educators. Success for students (1999-2010)* (Children's Progress Academic Assessment [CPAA] Technical Report). Retrieved January 12, 2011 from <http://www.childrensprogress.com/resources/research.shtml>.
- Cook, K. F., Dodd, B. G., & Fitzpatrick, S. J. (1999). A Comparison of Three Polytomous Item Response Theory Models in the Context of Testlet Scoring. *Journal of Outcome Measurement*, 3(1), 1-20.
- Council of Chief State School Officers, The (CCSSO). (2007). *Formative assessment and CSSO: A special initiative-a special opportunity*. Retrieved June 5th, 2007 from http://www.ccsso.org/projects/SCASS/Projects/Formative_Assessment_for_Students_and_Teachers/.
- Curtis, S. M. K. (2010). Bugs code for item response theory. *Journal of Statistical Software*, 36(c01).
- Davis, E. A., & Linn, M. C. (2000). Scaffolding students' knowledge integration: Prompts for reflection in KIE. *International Journal of Science Education*, 22, 819-837.
- Dodd BG, De Ayala RJ, Koch WR (1995) Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, 19, 5-22.
- Feng, M., Heffernan, N., & Koedinger, K.R. (2006a). *Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required* in M. Ikeda, K. Ashley, and T.-W. Chan (Eds.): ITS 2006, 31 - 40.
- Feng, M., Heffernan, N.T, Koedinger, K.R. (2006b) Addressing the Testing Challenge with a Web-Based E-Assessment System that Tutors as it Assesses. *In Proceedings of the 15th International World Wide Web Conference*. pp. 307-316. New York, NY: ACM Press. 2006.
- Fitzpatrick, S.J. & Dodd, B.G. (1997). *The effect on information of a transformation of the parameter scale*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
- Forster, M., and E. Sober. (1994). *How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions*. *Br. J. Phil. Sci.* 45:1-35.

- Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74, 153-160.
- Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In J. M. Bernardo, J. O.
- Gelman, A., & Rubin, D. B. (1993). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.
- Guzdial, M. (1994). Software-realized scaffolding to facilitate programming for science learning. *Interactive Learning Environments*, 4(1), 1-44.
- Guzdial, M., Hohmann, L., Konneman, M., Walton, C., & Soloway, E. (1998). Supporting Programming and Learning-to-Program with an Integrated CAD and Scaffolding Workbench*. *Interactive Learning Environments*, 6, 1(2), 143-179.
- Guzdial, M., & Kehoe, C. (1998). Apprenticeship-based learning environments: A principled approach to providing software-realized scaffolding through hypermedia. *Journal of Interactive Learning Research*, 9, 289-336.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE.
- Hannafin, M. J., & Land, S. M. (1997). The foundations and assumptions of technology-enhanced student-centered learning environments. *Instructional Science*, 25(3), 167-202.
- Hannafin, M. J., Land, S. & Oliver, K. (1999). Open learning environments: foundations, methods, and models, in: C. M. Reigeluth (Ed.) *Instructional design theories and models: a new paradigm of instructional theory* (vol. II) (London, Lawrence Erlbaum Associates), 115–140.
- Heffernan, N. & Heffernan, C. (2008). Assistments: Teacher's Manual. Retrieved from <http://teacherwiki.Assistment.org/wiki/images/8/8b/Teachermanualsinglesided.pdf>.
- Holton, D., & Clarke, D. (2006). Scaffolding and metacognition. *International journal of mathematical education in science and technology*, 37(2), 127-143.
- Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods*, 2(3), 261.
- IBM SPSS Statistics (2010). Base 19.0.

- Jackson, S. L., Krajcik, J., & Soloway, E. (1998). *The design of guided learner-adaptable scaffolding in interactive learning environments*. Proceedings from the Conference on Human Factors in Computing Systems in Los Angeles, CA.
- Jackson, S. L., Stratford, S.J., Krajcik, J.S., & Soloway, E. (1998). *Model-It: A Case Study of Learner-Centered Design Software for Supporting Model Building*. Proceedings from the Conference on Human Factors in Computing Systems in Los Angeles, CA.
- Kang, T., Cohen, A. S., & Sung, H. J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement, 33*(7), 499-518.
- Kass, R.E. & Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*(430), 773-795.
- Keller, L. A., Swaminathan, H., & Sireci, S. G. (2003). Evaluating Scoring Procedures for Context-Dependent Item Sets1. *Applied Measurement in Education, 16*(3), 207-222.
- Kim, J. S., & Bolt, D. M. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice, 26*(4), 38-51.
- Kingston, N.M. & Nash, B. (2011). The efficacy of formative assessment: a meta-analysis. *Educational Measurement: Issues and Practice, 30*(4), 28-37.
- Koedinger, K. R., & Alevan, V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. *Educational Psychology Review, 19*(3), 239-264.
- Koedinger, K. R., McLaughlin, E. A., & Heffernan, N. T. (2010). A Quasi-Experimental Evaluation of an On-line Formative Assessment and Tutoring System. *Journal of Educational Computing Research, 43*(4), 489-510.
- Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., . . . Ryan, M. (2003). Problem-based learning meets case-based reasoning in the middle-school science classroom: Putting Learning by Design™ into practice. *Journal of the Learning Sciences, 495-547*.
- Linn, M. C. (1995). Designing computer learning environments for engineering and computer science: The Scaffolded Knowledge Integration framework. *Journal of Science Education and Technology, 4*, 103–126.
- Lord, F. (1952). A theory of test scores. *Psychometric monographs*. Iowa City, IA: Psychometric Society.
- McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences, 15*(2), 153.

- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741.
- Muraki, E. (1992). A generalized partial credit model: Application of an E-M algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Muraki, E., & Bock, D. (2003). PARSCALE for windows (Version 4.1). *Chicago: Scientific Software International*.
- No Child Left Behind Act of 2001. Public Law.107-110, 115 Stat. 1425.
- Pardos, Z. A., Heffernan, N.T., Anderson, B. & Heffernan, C. L. (2006). *Using fine-grained skill models to fit student performance with Bayesian networks*. Submitted to the ITS2006 Education Data Mining Workshop, Taiwan ROC, June 26, 2006.
- Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *The journal of the learning sciences*, 13(3), 423-451.
- Pitt, M. A., Kim, W., & Myung, I. J. (2003). Flexibility versus generalizability in model selection. *Psychonomic Bulletin & Review*, 10(1), 29-44.
- Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5), 793-808.
- Puntambekar S, Hubscher R. (2005). Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed? *Educational Psychologist*, 40, 1–12.
- Puntambekar, S., & Kolodner, J. L. (2005). Toward implementing distributed scaffolding: Helping students learn science from design. *Journal of Research in Science Teaching*, 42(2), 185-217.
- Quellmalz, E., & Haertel, G. D. (2005). Use of technology-supported tools for large-scale science assessment: Implications for assessment practice and policy at the state level. *Washington, DC: National Research Council*.
- Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and testing. *Science*, 323(5910), 75.

- Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., Kyza, E., Edelson, D. & Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences*, 337-386.
- Rasch, G. (1960). Probabilistic models for some intelligence and achievement tests. *Copenhagen, Denmark Danish Institute for Educational Research*.
- Reise, S.P., Ainsworth, A.T., & Haviland, M.G. (2005). Item response theory: fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science*, 14(2), 95-101.
- Reiser, B. J. (2002). *Why scaffolding should sometimes make tasks more difficult for learners*. Proceedings from the Annual Computer-Supported Collaborative Learning Conference.
- Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences*, 273-304.
- Reiser, B. J., Tabak, I., Sandoval, W. A., Smith, B., Steinmuller, F., & Leone, A. J. (2001). BGuILE: Strategic and conceptual scaffolds for scientific inquiry in biology classrooms. In S.M. Carver & D. Klahr (Eds.), *Cognition and Instruction: Twenty five years of progress* (pp. 263–305). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*: Oxford University Press.
- Rosenbaum, P. R. (1988). Item bundles. *Psychometrika*, 53, 349-359.
- Rupp, A. A. (2003). Item response modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing*, 3(4), 365-384.
- Samejima, F. (1996). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*, 85-100. New York: Springer.
- Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Silverfrost Ltd. (2007). FTN95 (Version 3.41).
- Sinharay, S. (2004). Experiences with MCMC convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics*, 29 (4), 461-488.
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28, 237-247.
- Sharpe, T. (2006). 'Unpacking' Scaffolding: Identifying Discourse and Multimodal Strategies that Support Learning. *Language and Education*, 20(3), 211-231.

- Shepard, L. A. (2005). Linking Formative Assessment to Scaffolding. *Educational Leadership*, 63(3), 5.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. *Research report*, 98(009).
- Spiegelhalter, D. J., Best, N., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, 64, 583-640.
- Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS User Manual, Version 1.4. MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology and Public Health, Imperial College School of Medicine, UK.
- Spiegelhalter, D.J., Thomas, A., Best, N., & Gilks, W. (1995). *BUGS 0.5*: Bayesian inference using Gibbs sampling manual (version ii)*. Cambridge, UK: MRC Biostatistics Unit.
- Stiggins, R. (2005). From Formative Assessment to Assessment For Learning: A Path to Success in Standards-Based Schools. *Phi Delta Kappan*, 87(4).
- Stone, C. (1998a). The Metaphor of Scaffolding. *Journal of Learning Disabilities*, 31(4), 344.
- Stone, C. (1998b). Should we salvage the scaffolding metaphor? *Journal of Learning Disabilities*, 31(4), 409.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51(4), 567-577.
- Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiplecategorical-response models. *Journal of Educational Measurement*, 26, 247-260.
- Thissen-Roe, A., Hunt, E., & Minstrell, J. (2004). The DIAGNOSER project: Combining assessment and learning. *Behavior Research Methods*, 36(2), 234-240.
- Tucker, B. (2009). *Beyond the bubble: Technology and the future of student assessment: Education Sector*.
- Van Der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement*, 25(3), 273.
- Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*: Springer New York.

- Veldkamp, B. P. (2003). Item selection in polytomous CAT. *New developments in psychometrics*, 207-214.
- Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. In (M. Cole, S. Scriber, V. Johns-Steiner & E. Souberman, Eds.). Cambridge, MA: Harvard University Press (Original work published 1930).
- Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 LawSchool AdmissionsTest as an example. *Applied Measurement in Education*, 8, 157-186.
- Wainer, H., Bradlow, E., & Du, Z. (2000). Testlet response theory: An analog for the 3PL useful in adaptive testing. *Computerized adaptive testing: Theory and practice*, 245-270.
- Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24, 185-202.
- Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, 15(1), 22-29.
- Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, 37, 203-220.
- Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement*, 26, 109-128.
- Wang, X., Bradlow, E. T., & Wainer, H. (2005). User's guide for SCORIGHT (Version 3.0): a computer program for scoring tests built of testlets including a module for covariate analysis. www.ets.org/research/contact.html.
- Wang, W. C., Cheng, Y. Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and psychological measurement*, 65(1), 5-27.
- Wang, W. C., & Wilson, M. (2005a). The Rasch Testlet Model. *Applied Psychological Measurement*, 29(2), 126-149.
- Wang, W. C., & Wilson, M. (2005b). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, 29(4), 296-318.
- Weakliem, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research*, 27(3), 359-397.

- Wilson, M. R., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, 60, 181-198.
- Wood, D. (1988). *How children think and learn*: Oxford.
- Wood, D., & Bruner, J. ,& Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child psychology and psychiatry*, 17(2), 89-100.
- Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). ConQuest: Generalized item response modeling software [Computer software and manual]. Camberwell, Victoria: Australian Council for Educational Research.
- Zhang, O., Shen, L. & Cannady, M. (2010). *Polytomous IRT or Testlet Model: An Evaluation of Scoring Models in Small Testlet Size Situations*. Paper Presented at Annual Meeting of the 15th International Objective Measurement Workshop, Boulder, CO.
- Zimowski, M., Muraki, E., Mislevy, R., & Bock, R. (1996). BILOG-MG [Computer software]. *Chicago: Scientific Software International*.
- Zhu, X. (2009). *Assessing Fit of Item Response Models for Performance Assessments using Bayesian Analysis*. Available from ProQuest Dissertations and Theses database. (UMI No. 3400484)
- Zhu, L., and Carlin, B. P. (2000), Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Statistics in Medicine*, 19, 2265–2278.

Appendix A

Table 20.
Item Fit Statistics for the 1PL and 2PL

Item	1PL	2PL	$\Delta\chi^2$	Item	1PL	2PL	$\Delta\chi^2$
ITEM0001	16.8	15.6	1.2	ITEM0071	23.6	8.2	15.4
ITEM0002	4.5	2.9	1.6	ITEM0072	18.4	5.7	12.7
ITEM0003	7.7	5.9	1.8	ITEM0073	20.5	5.9	14.6
ITEM0004	1.7	2.7	-1	ITEM0074	21.1	11.4	9.7
ITEM0005	14.8	6	8.8	ITEM0075	8	1.3	6.7
ITEM0006	2.7	1.5	1.2	ITEM0076	4.2	2.5	1.7
ITEM0007	6	3.4	2.6	ITEM0077	1.3	0.9	0.4
ITEM0008	6.6	2.4	4.2	ITEM0078	50.3	43.6	6.7
ITEM0009	9.5	7.3	2.2	ITEM0079	6.1	4.4	1.7
ITEM0010	33.9	9.9	24	ITEM0080	4.8	3.4	1.4
ITEM0011	16.3	24.7	-8.4	ITEM0081	39.6	40.4	-0.8
ITEM0012	3.4	11.3	-7.9	ITEM0082	62.9	57	5.9
ITEM0013	6.2	4.8	1.4	ITEM0083	22.1	27.8	-5.7
ITEM0014	3.9	6.4	-2.5	ITEM0084	11.2	18.5	-7.3
ITEM0015	5.4	2.6	2.8	ITEM0085	5.3	11.1	-5.8
ITEM0016	5.1	2.6	2.5	ITEM0086	31	35	-4
ITEM0017	2.2	4.2	-2	ITEM0087	38.6	17.9	20.7
ITEM0018	4.9	2	2.9	ITEM0088	1.4	6.2	-4.8
ITEM0019	7.5	3.2	4.3	ITEM0089	1.1	0.8	0.3
ITEM0020	0.1	3.7	-3.6	ITEM0090	6.5	8.5	-2
ITEM0021	10.3	7.7	2.6	ITEM0091	9.9	4	5.9
ITEM0022	2.7	3.3	-0.6	ITEM0092	3.3	7.2	-3.9
ITEM0023	3.3	3.6	-0.3	ITEM0093	3.1	7.6	-4.5
ITEM0024	12.3	1.2	11.1	ITEM0094	6.3	9.7	-3.4
ITEM0025	1.7	3.7	-2	ITEM0095	0.8	2.2	-1.4
ITEM0026	7.6	5.1	2.5	ITEM0096	2.8	6	-3.2
ITEM0027	0.2	0.9	-0.7	ITEM0097	2.1	4.9	-2.8
ITEM0028	1.1	2.9	-1.8	ITEM0098	7.5	4.9	2.6
ITEM0029	99.5	3.2	96.3	ITEM0099	1.9	1.7	0.2
ITEM0030	15.1	13.5	1.6	ITEM0100	42.6	15.5	27.1
ITEM0031	15.3	10.2	5.1	ITEM0101	15.8	6.5	9.3
ITEM0032	2	6.6	-4.6	ITEM0102	20.1	7.3	12.8
ITEM0033	18.4	5.7	12.7	ITEM0103	10	5.6	4.4
ITEM0034	5.2	9.6	-4.4	ITEM0104	5.9	4.6	1.3
ITEM0035	2.4	2	0.4	ITEM0105	12.5	15.3	-2.8
ITEM0036	2.1	4.5	-2.4	ITEM0106	11.5	8.9	2.6

Item	1PL	2PL	$\Delta\chi^2$
ITEM0037	2.3	4.3	-2
ITEM0038	1	1.3	-0.3
ITEM0039	6.4	5	1.4
ITEM0040	0.3	2.3	-2
ITEM0041	11	8.9	2.1
ITEM0042	1.2	3.9	-2.7
ITEM0043	0.9	3.9	-3
ITEM0044	5.3	2.8	2.5
ITEM0045	11.6	9.4	2.2
ITEM0046	3.8	1.9	1.9
ITEM0047	1.1	2.1	-1
ITEM0048	0.2	0.7	-0.5
ITEM0049	29	15.7	13.3
ITEM0050	6.6	3.9	2.7
ITEM0051	4.6	5.4	-0.8
ITEM0052	1.9	0.7	1.2
ITEM0053	1.7	3	-1.3
ITEM0054	28.9	8.7	20.2
ITEM0055	10.8	5.4	5.4
ITEM0056	17.3	8.8	8.5
ITEM0057	10.5	6.5	4
ITEM0058	9.1	6	3.1
ITEM0059	0.3	1.4	-1.1
ITEM0060	3.7	2.7	1
ITEM0061	2	2.2	-0.2
ITEM0062	59.1	67.9	-8.8
ITEM0063	30.4	33.2	-2.8
ITEM0064	58	56.3	1.7
ITEM0065	21.5	8.6	12.9
ITEM0066	5.3	5	0.3
ITEM0067	4.6	3.6	1
ITEM0068	3.5	4.5	-1
ITEM0069	29.7	24.9	4.8
ITEM0070	29.6	16.3	13.3

Item	1PL	2PL	$\Delta\chi^2$
ITEM0107	14.1	13.7	0.4
ITEM0108	10.8	4.1	6.7
ITEM0109	2.4	4.5	-2.1
ITEM0110	18.9	4.5	14.4
ITEM0111	4.2	8.5	-4.3
ITEM0112	9.5	12.5	-3
ITEM0113	32.8	5.4	27.4
ITEM0114	16.4	3.6	12.8
ITEM0115	0.9	0.3	0.6
ITEM0116	30	5	25
ITEM0117	9	3.6	5.4
ITEM0118	4.4	5.8	-1.4
ITEM0119	9.9	13.8	-3.9
ITEM0120	8.4	8.3	0.1
ITEM0121	16.3	7.8	8.5
ITEM0122	17.4	12.5	4.9
ITEM0123	2.5	5.2	-2.7
ITEM0124	4.6	5.8	-1.2
ITEM0125	0.9	2.8	-1.9
ITEM0126	9.5	7.2	2.3
ITEM0127	10	6.3	3.7
ITEM0128	2.6	3.1	-0.5
ITEM0129	4.5	6	-1.5
ITEM0130	3.5	4.6	-1.1
ITEM0131	31.2	5.7	25.5
ITEM0132	19.2	7.3	11.9
ITEM0133	17.3	2.3	15
ITEM0134	40.1	7.5	32.6
ITEM0135	6.9	3.3	3.6
ITEM0136	15.9	5.2	10.7
ITEM0137	15.1	2.7	12.4
ITEM0138	13.5	9.1	4.4
ITEM0139	40.4	12	28.4
ITEM0140	3021.3	600.4	2420.9

Appendix B

PROGRAM DIC_DichotItems

INTEGER :: NITEMS,NDRAWS,NSTDS

REAL :: P1,P2,L1,L2,SUM_L1,SUM_L2,AVG_L1

REAL,ALLOCATABLE :: ADRAWS(:,:),BDRAWS(:,:),THETADRAWS(:,:)

REAL,ALLOCATABLE :: APARS(:),BPARS(:),THETAS(:)

INTEGER,ALLOCATABLE :: RESP_ARRAY(:,:)

OPEN (3,FILE="t_DrawsC.txt")

OPEN (4,FILE="a_DrawsC.txt")

OPEN (5,FILE="b_DrawsC.txt")

OPEN (6,FILE="2PL_AllItems.dat")

OPEN (7,FILE="thetas.txt")

OPEN (8,FILE="ItemPars.txt")

OPEN (9,FILE="DIC.OUT")

NITEMS=140; NDRAWS=500;

NSTDS=2745

SUM_L1=0.0; AVG_L1=0.0

SUM_L2=0.0

ALLOCATE

(ADRAWS(NDRAWS,NITEMS),BDRAWS(NDRAWS,NITEMS),THETADRAWS(NDRAWS,NSTDS))

ALLOCATE (RESP_ARRAY(NSTDS,NITEMS))

ALLOCATE (APARS(NITEMS),BPARS(NITEMS),THETAS(NSTDS))

!***READ DATA***

DO i=1,NDRAWS

 READ (3,20) (THETADRAWS(i,k),k=1,NSTDS)

 20 FORMAT (2745f11.6)

END DO

DO i=1,NDRAWS

 READ (4,30) (ADRAWS(i,j),j=1,NITEMS)

 30 FORMAT (140f11.6)

END DO

DO i=1,NDRAWS

 READ (5,40) (BDRAWS(i,j),j=1,NITEMS)

 40 FORMAT (140f11.6)

END DO

DO k=1,NSTDS

 READ (6,50) (RESP_ARRAY(k,j),j=1,NITEMS)

 50 FORMAT (140i1)

END DO

```

!***CALCULATE MEAN DEVIANCE***
DO i=1,NDRAWS
  DO k=1,NSTDS
    DO j=1,NITEMS
      P1=0.0; L1=0.0
      P1 = 1/(1+EXP(-((1.7*ADRAWS(i,j))*(THETADRAWS(i,k)-BDRAWS(i,j))))))
      IF (RESP_ARRAY(k,j) == 1) THEN
        L1 = (-2*(LOG(.000001+P1)))
      ELSE IF (RESP_ARRAY(k,j) == 0) THEN
        L1 = (-2*(LOG(.000001+(1-P1))))
      ELSE IF (RESP_ARRAY(k,j) == 9) THEN
        L1 = 0
      END IF
      SUM_L1 = SUM_L1 + L1
    END DO
  END DO
END DO

AVG_L1 = SUM_L1/NDRAWS

!***READ DATA***
DO k=1,NSTDS
  READ (7,60) THETAS(k)
  60 FORMAT (t7,f11.4)
END DO

DO j=1,NITEMS
  READ (8,70) APARS(j),BPARS(j)
  70 FORMAT (t7,f11.4,t29,f11.4)
END DO

!***CALCULATE DEVIANCE OF POSTERIOR EXPECTATIONS***
DO k=1,NSTDS
  DO j=1,NITEMS
    P2=0.0; L2=0.0
    P2 = 1/(1+EXP(-((1.7*APARS(j))*(THETAS(k)-BPARS(j))))))
    IF (RESP_ARRAY(k,j) == 1) THEN
      L2 = (-2*(LOG(.000001+(P2))))
    ELSE IF (RESP_ARRAY(k,j) == 0) THEN
      L2 = (-2*(LOG(.000001+(1-P2))))
    ELSE IF (RESP_ARRAY(k,j) == 9) THEN
      L2 = 0
    END IF
    SUM_L2 = SUM_L2 + L2
  END DO
END DO

!***CALCULATE DIC***
WRITE (9,80)

```

```
80 FORMAT (3x,"Dbar    D(thetabar)    pD    DIC    "/)
WRITE(9,90) AVG_L1,SUM_L2,AVG_L1-SUM_L2,(AVG_L1+(AVG_L1-SUM_L2))
90 FORMAT (4f12.4)
```

```
END PROGRAM
```

Appendix C

```
PROGRAM DIC_PolyItems
```

```
INTEGER :: NDRAWS,NSTDS,NBUNDLES
```

```
REAL :: L1,L2,SUM_L1,SUM_L2,AVG_L1
```

```
REAL,ALLOCATABLE :: ADRAWS(:,:),THETADRAWS(:,:),CUTDRAWS(:,,:),BDRAWS(:,:)
```

```
REAL,ALLOCATABLE :: APARS(:),THETAS(:),CUTPARS(:,:),BPARS(:),TEMP(:)
```

```
INTEGER,ALLOCATABLE :: RESP_ARRAY(:,:),ITEMCATS(:),MAXSCORE(:)
```

```
OPEN (3,FILE="t_drawsC.txt")
```

```
OPEN (4,FILE="a_drawsC.txt")
```

```
OPEN (5,FILE="b_drawsC.txt")
```

```
OPEN (6,FILE="dr_drawsC.txt")
```

```
OPEN (7,FILE="ORM.dat")
```

```
OPEN (8,FILE="itemcats.TXT")
```

```
OPEN (9,FILE="thetas.txt")
```

```
OPEN (10,FILE="itempars.txt")
```

```
OPEN (11,FILE="cutpars.txt")
```

```
OPEN (12,FILE="DIC_poly.OUT")
```

```
NDRAWS=500; NSTDS=2745
```

```
NBUNDLES=32
```

```
SUM_L1=0.0; AVG_L1=0.0
```

```
SUM_L2=0.0
```

```
L1=0.0; L2=0.0
```

```
ALLOCATE
```

```
(ADRAWS(NDRAWS,NBUNDLES),THETADRAWS(NDRAWS,NSTDS),CUTDRAWS(NDRAWS,NBUNDLES,10),BDRAWS(NDRAWS,NBUNDLES))
```

```
ALLOCATE
```

```
(RESP_ARRAY(NSTDS,NBUNDLES),ITEMCATS(NBUNDLES),MAXSCORE(NBUNDLES))
```

```
ALLOCATE
```

```
(APARS(NBUNDLES),THETAS(NSTDS),CUTPARS(NBUNDLES,10),BPARS(NBUNDLES))
```

```
!***READ DATA***
```

```
DO i=1,NDRAWS
```

```
  READ (3,20) (THETADRAWS(i,k),k=1,NSTDS)
```

```
  20 FORMAT (2745f11.6)
```

```
END DO
```

```
DO i=1,NDRAWS
```

```
  READ (4,25) (ADRAWS(i,j),j=1,NBUNDLES)
```

```
  25 FORMAT (32f11.6)
```

```
END DO
```

```
DO i=1,NDRAWS
```

```
  READ (5,30) (BDRAWS(i,j),j=1,NBUNDLES)
```

```

30 FORMAT (32f11.6)
END DO

DO j=1,NBUNDLES
  READ (8,35) ITEMCATS(j)
  MAXSCORE(j) = ITEMCATS(j)+2
  35 FORMAT (i1)
END DO
iSUM = SUM(ITEMCATS)
ALLOCATE(TEMP(iSUM))

CUTDRAWS=0.0
DO i=1,NDRAWS
  40 FORMAT (107f10.6)
  READ (6,40) (TEMP(j),j=1,iSUM)
  iCOUNT=0
  DO j=1,NBUNDLES
    DO l=2,ITEMCATS(j)+1
      iCOUNT=iCOUNT+1
      CUTDRAWS(i,j,l) = TEMP(iCOUNT)
    END DO
  END DO
END DO
END DO

DO k=1,NSTDS
  READ (7,45) (RESP_ARRAY(k,j),j=1,NBUNDLES)
  45 FORMAT (32i1)
END DO

!***CALCULATE MEAN DEVIANCE***
DO i=1,NDRAWS
  DO k=1,NSTDS
    DO j=1,NBUNDLES
      IF (RESP_ARRAY(k,j)>0) THEN
        L1=0.0
        T=0.0
        T=1.7*ADRAWS(i,j)*(THETADRAWS(i,k)-BDRAWS(i,j))
        IF (RESP_ARRAY(k,j)==MAXSCORE(j)) THEN
          PSTAR_BIG=1.0
        ELSE
          PSTAR_BIG=1/(1+EXP(-(CUTDRAWS(i,j,RESP_ARRAY(k,j))-T)))
        END IF
        IF (RESP_ARRAY(k,j)==1) THEN
          PSTAR_SMALL = 0.0
        ELSE
          PSTAR_SMALL=1/(1+EXP(-(CUTDRAWS(i,j,RESP_ARRAY(k,j)-1)-T)))
        END IF
        L1 = -2*LOG(PSTAR_BIG - PSTAR_SMALL)
        SUM_L1 = SUM_L1 + L1
      END IF
    END DO
  END DO
END DO

```

```

        END IF
    END DO
END DO

AVG_L1 = SUM_L1/(NDRAWS)

!***READ DATA***
DO k=1,NSTDS
    READ (9,50) THETAS(k)
    50 FORMAT (t7,f11.4)
END DO

DO j=1,NBUNDLES
    READ (10,55) APARS(j),BPARS(j)
    55 FORMAT (t7,f11.4,t29,f11.4)
END DO

CUTPARS=0.0
DO j=1,NBUNDLES
    READ (11,60) (CUTPARS(j,l),l=2,ITEMCATS(j)+1)
    60 FORMAT (1000f11.5)
END DO

!***CALCULATE DEVIANCE OF POSTERIOR EXPECTATIONS***
DO k=1,NSTDS
    DO j=1,NBUNDLES
        IF (RESP_ARRAY(k,j)>0) THEN
            L2=0.0
            R=0.0
            R=1.7*APARS(j)*(THETAS(k)-BPARS(j))
            IF (RESP_ARRAY(k,j)==MAXSCORE(j)) THEN
                PSTAR_BIG=1.0
            ELSE
                PSTAR_BIG=1/(1 + EXP(-(CUTPARS(j,RESP_ARRAY(k,j)) - R)))
            END IF
            IF (RESP_ARRAY(k,j)==1) THEN
                PSTAR_SMALL = 0.0
            ELSE
                PSTAR_SMALL=1/(1 + EXP(-(CUTPARS(j,RESP_ARRAY(k,j)-1) - R)))
            END IF
            L2 = -2*LOG(PSTAR_BIG - PSTAR_SMALL)
            SUM_L2 = SUM_L2 + L2
        END IF
    END DO
END DO

!***CALCULATE DIC***
WRITE(12,65)
65 FORMAT (3x,"Dbar    D(thetabar)    pD    DIC    "/)

```

```
WRITE(12,70) AVG_L1,SUM_L2,AVG_L1-SUM_L2,(AVG_L1+(AVG_L1-SUM_L2))  
70 FORMAT (4f12.4)
```

```
END PROGRAM
```


Appendix D

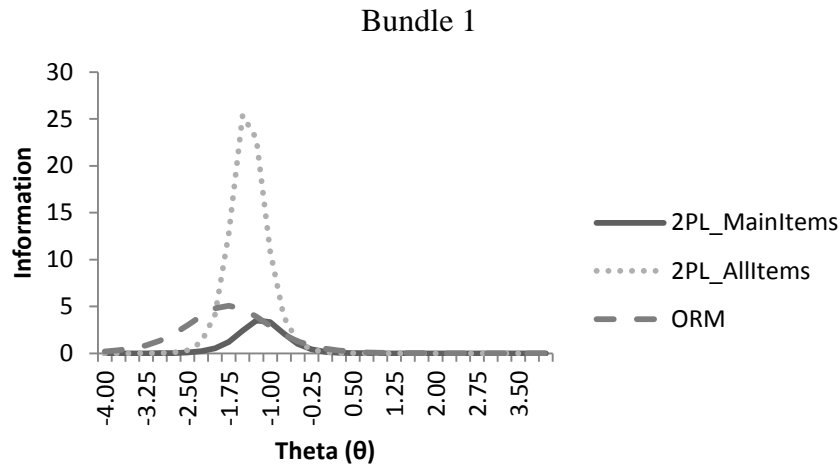


Figure 28. Information for Bundle 1. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

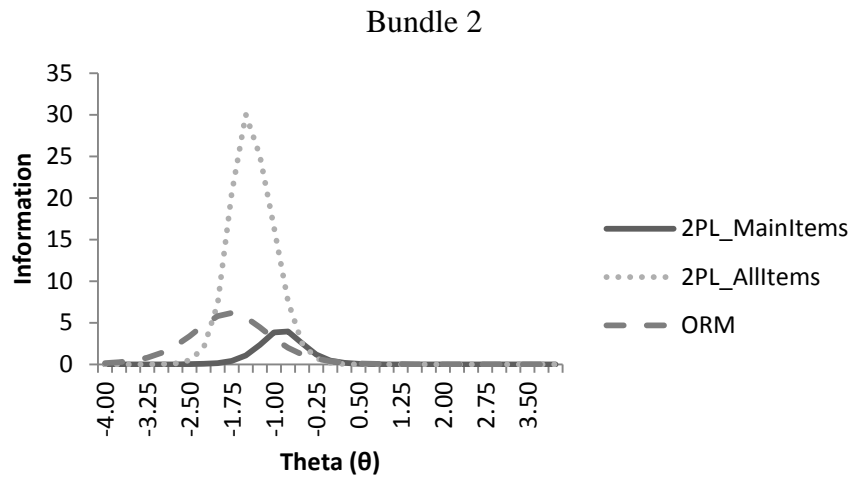


Figure 29. Information for Bundle 2. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

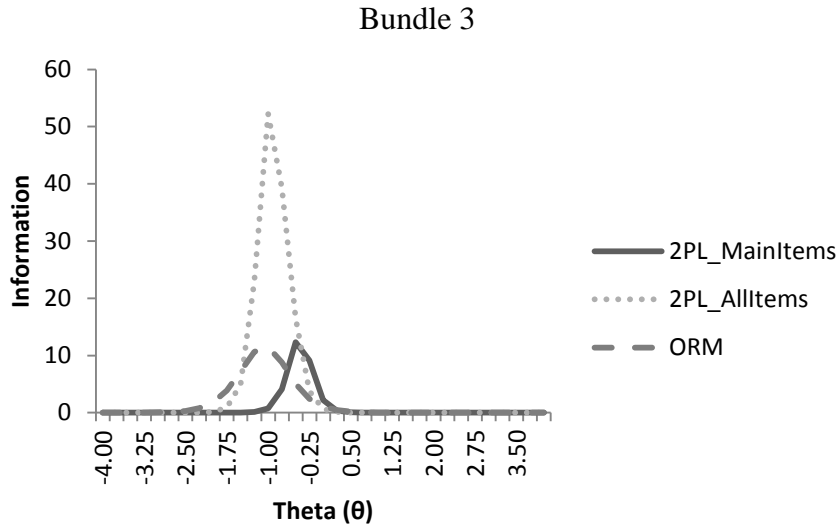


Figure 30. Information for Bundle 3. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

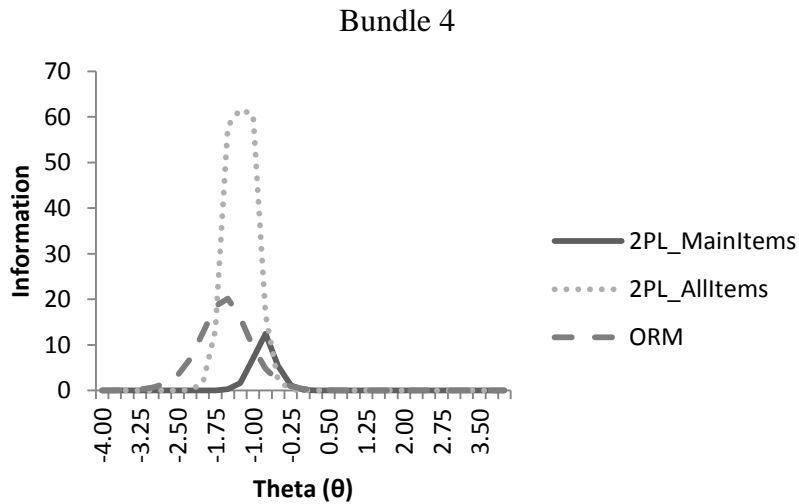


Figure 31. Information for Bundle 4. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

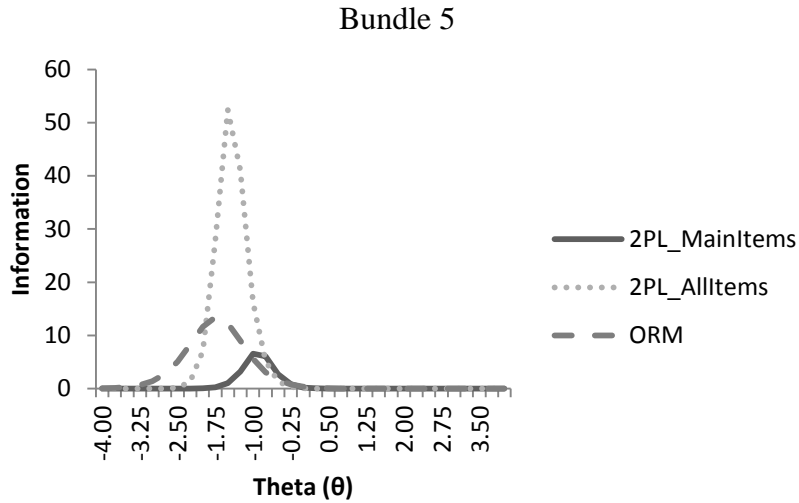


Figure 32. Information for Bundle 5. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

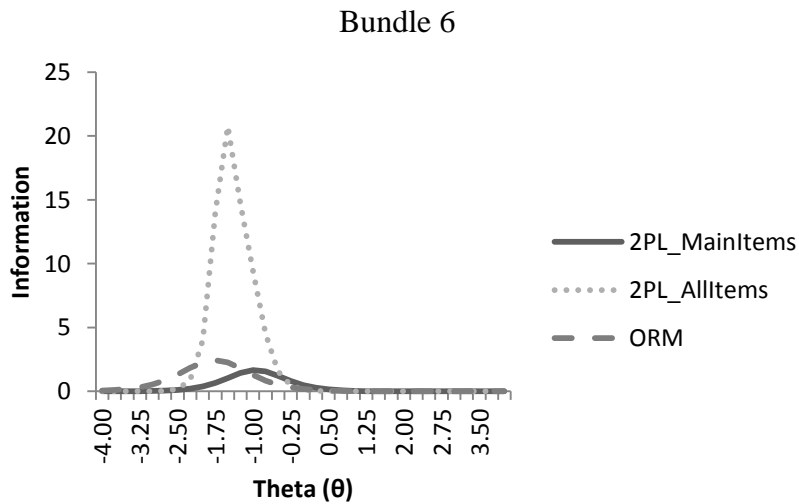


Figure 33. Information for Bundle 6. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

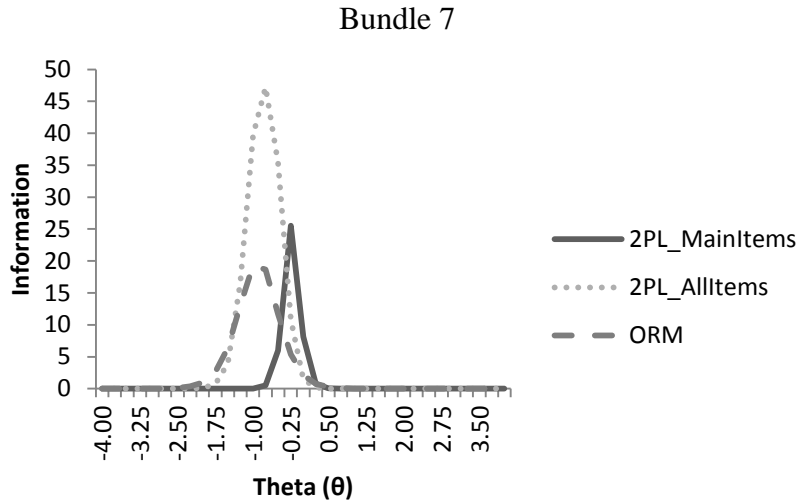


Figure 34. Information for Bundle 7. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

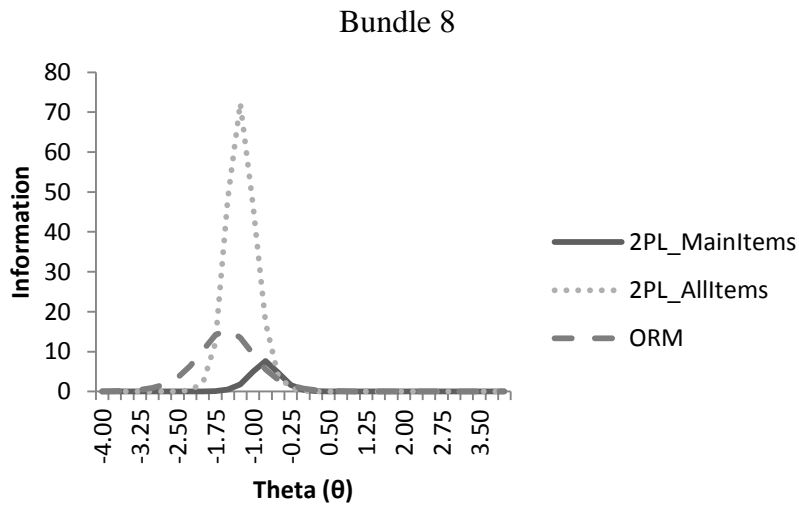


Figure 35. Information for Bundle 8. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

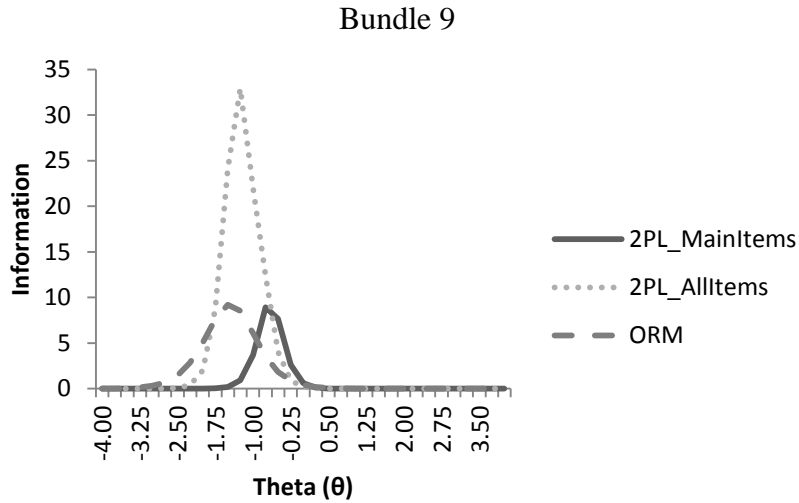


Figure 36. Information for Bundle 9. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

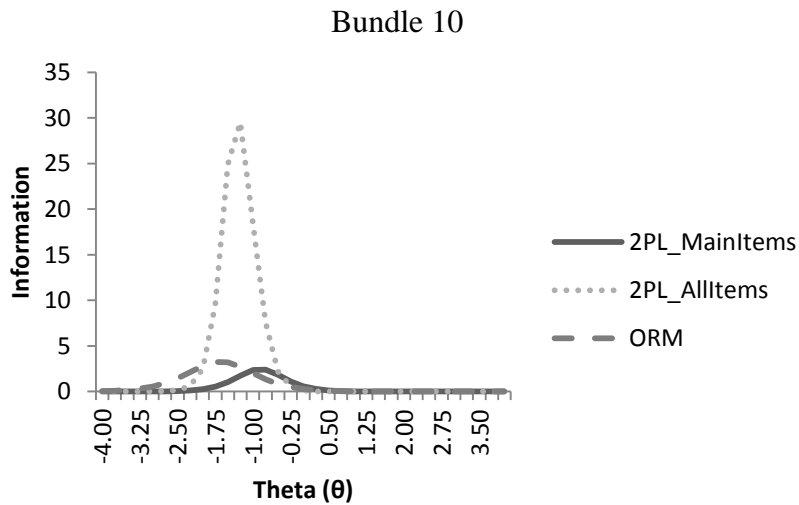


Figure 37. Information for Bundle 10. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

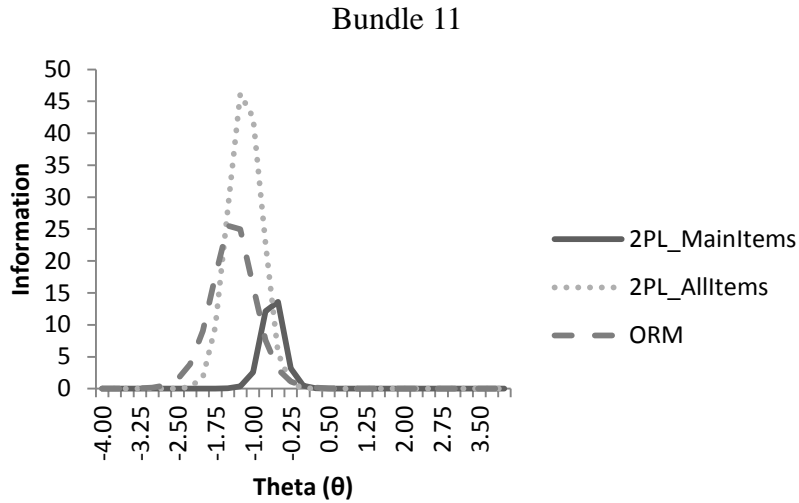


Figure 38. Information for Bundle 11. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

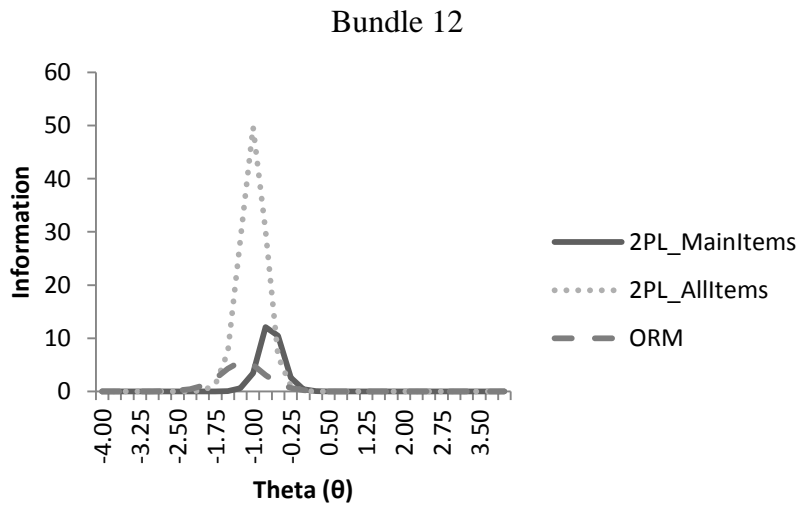


Figure 39. Information for Bundle 12. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

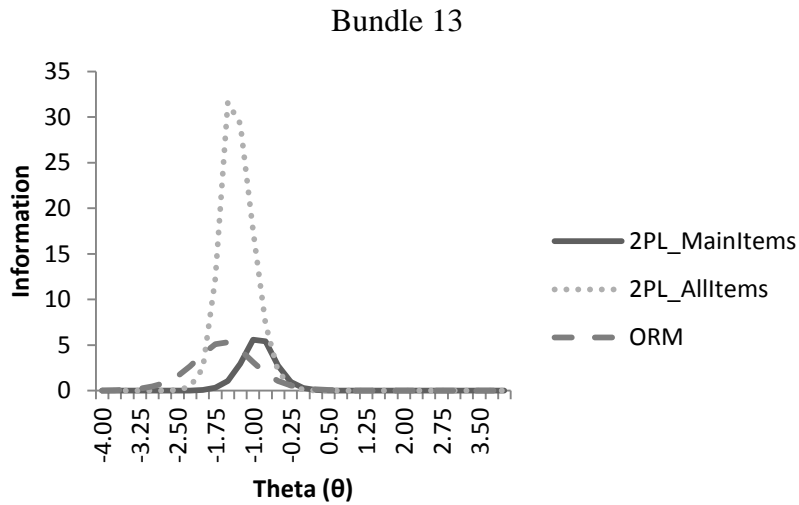


Figure 40. Information for Bundle 13. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

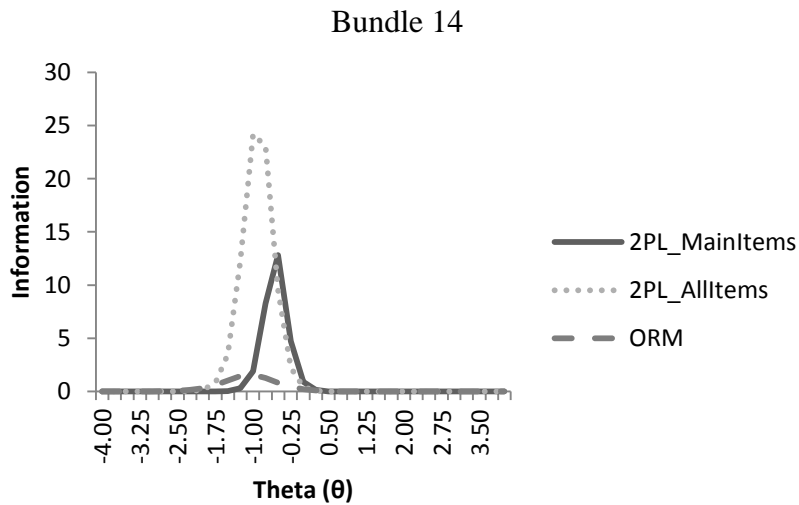


Figure 41. Information for Bundle 14. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

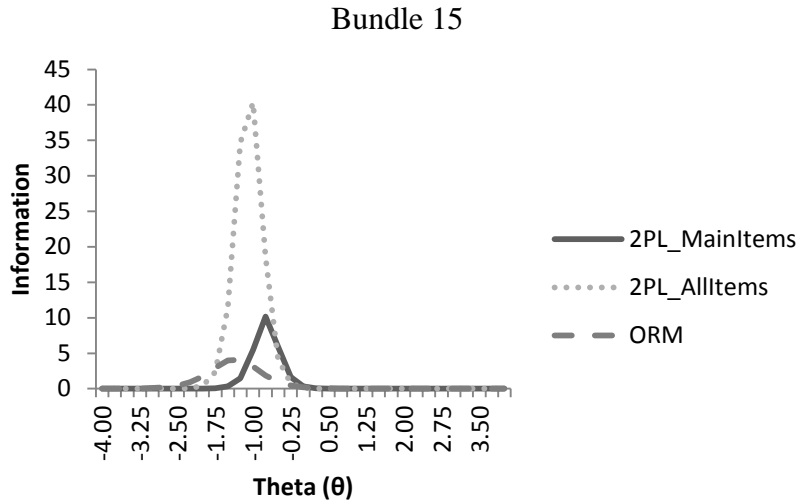


Figure 42. Information for Bundle 15. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

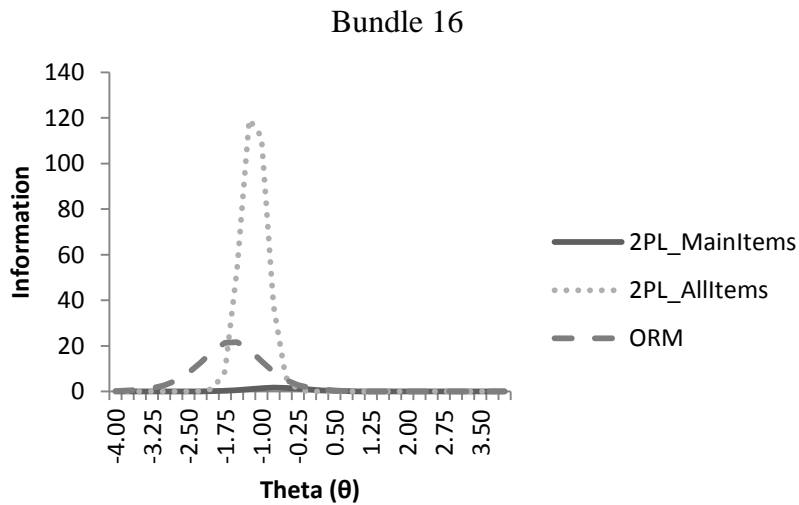


Figure 43. Information for Bundle 16. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

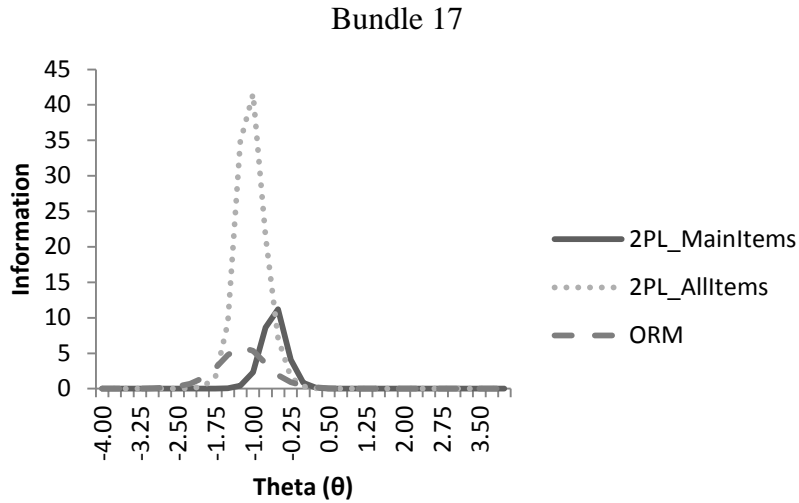


Figure 44. Information for Bundle 17. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

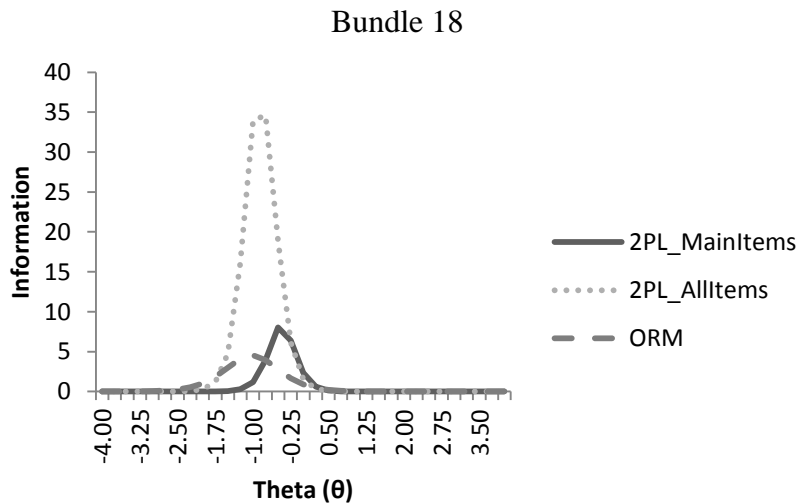


Figure 45. Information for Bundle 18. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

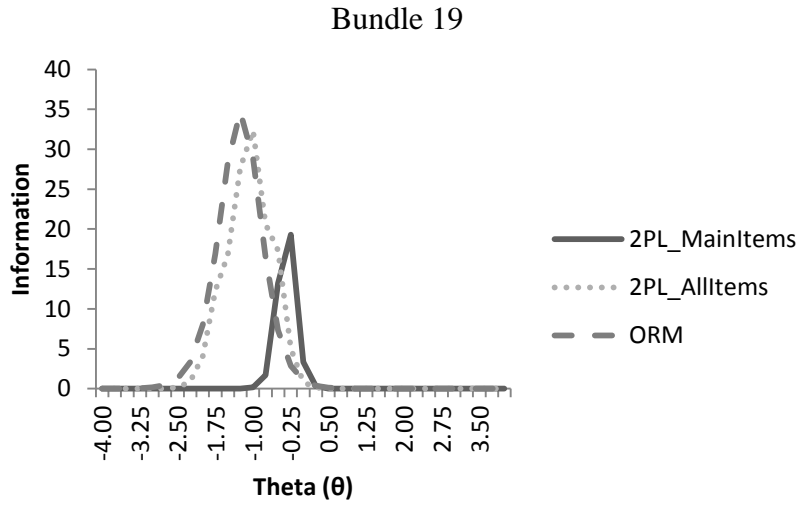


Figure 46. Information for Bundle 19. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

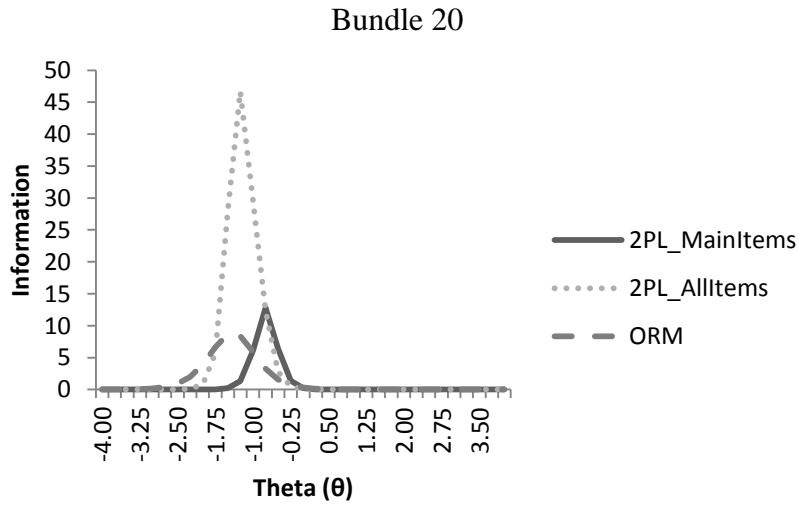


Figure 47. Information for Bundle 20. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

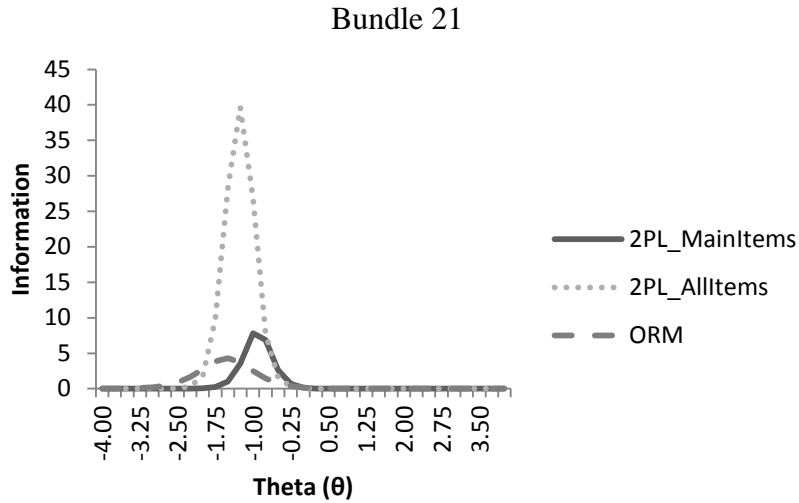


Figure 48. Information for Bundle 21. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

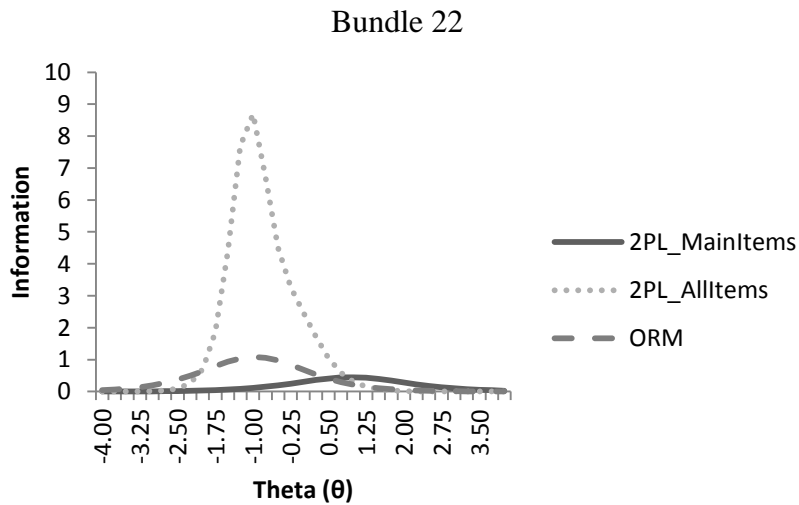


Figure 49. Information for Bundle 22. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

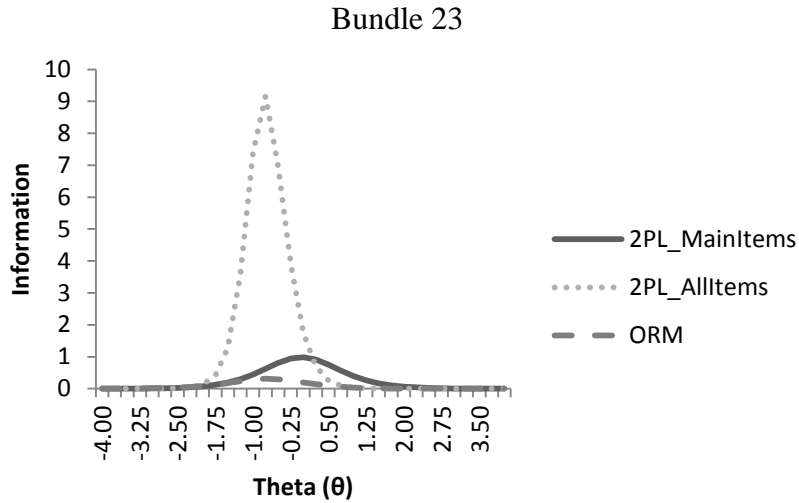


Figure 50. Information for Bundle 23. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

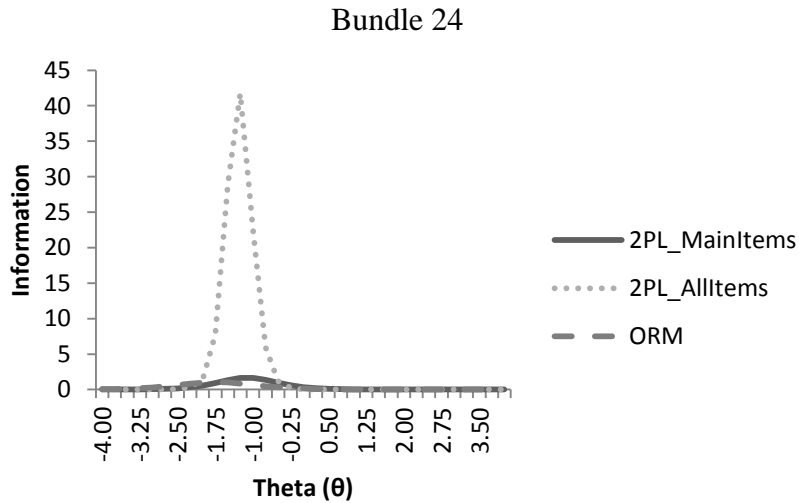


Figure 51. Information for Bundle 24. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

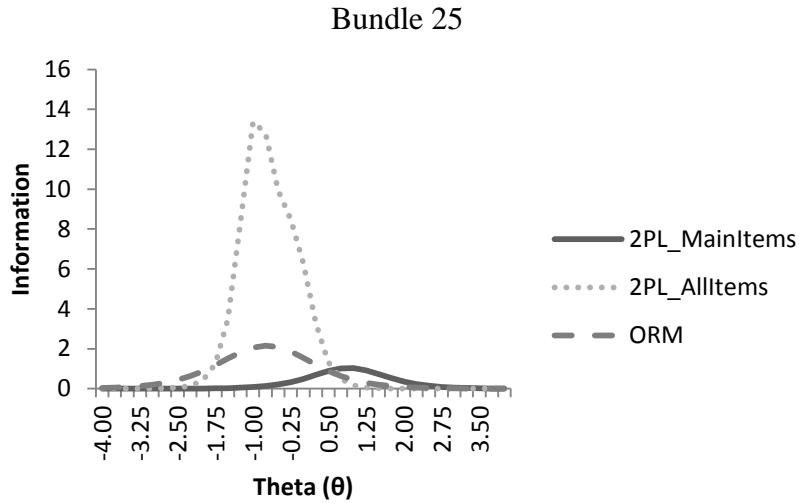


Figure 52. Information for Bundle 25. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

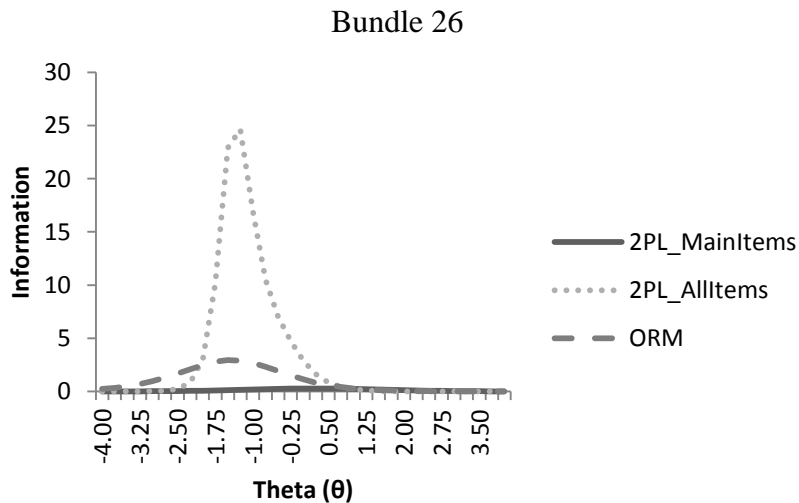


Figure 53. Information for Bundle 26. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

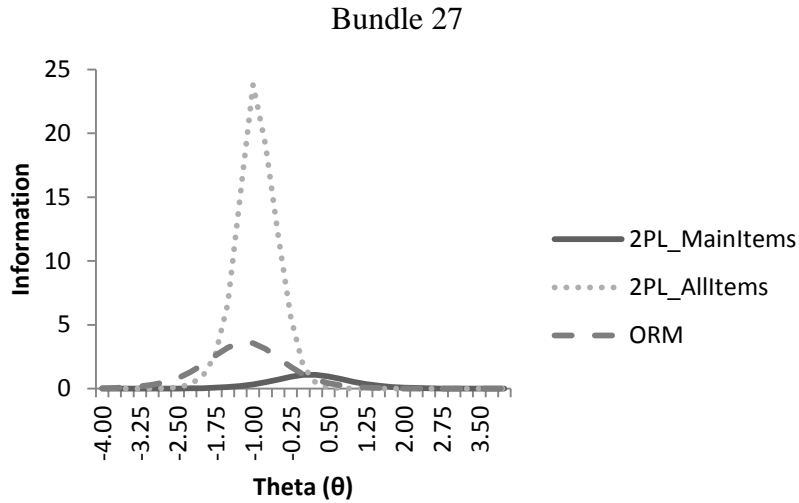


Figure 54. Information for Bundle 27. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

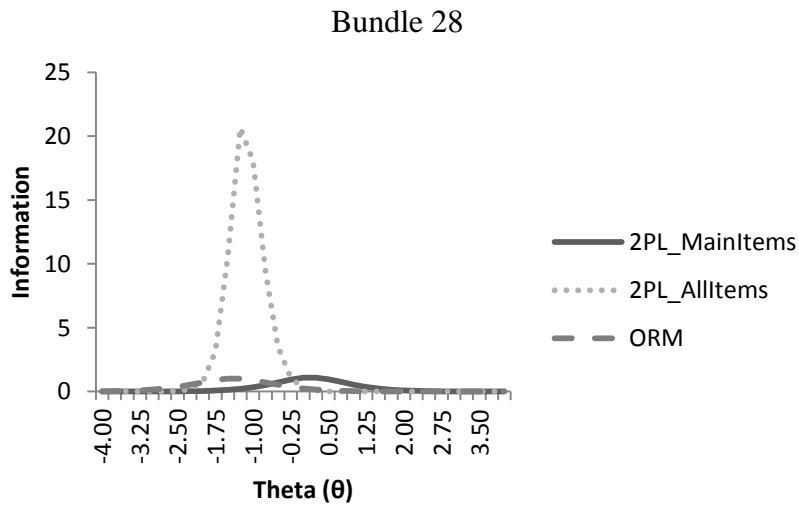


Figure 55. Information for Bundle 28. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

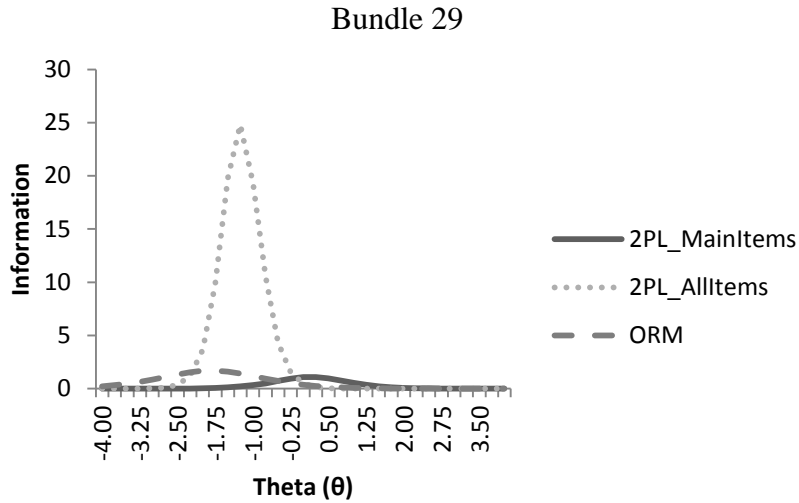


Figure 56. Information for Bundle 29. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

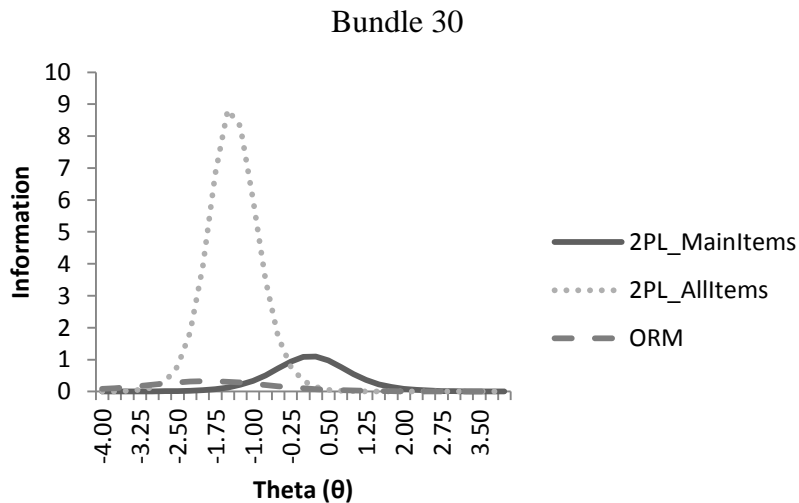


Figure 57. Information for Bundle 30. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

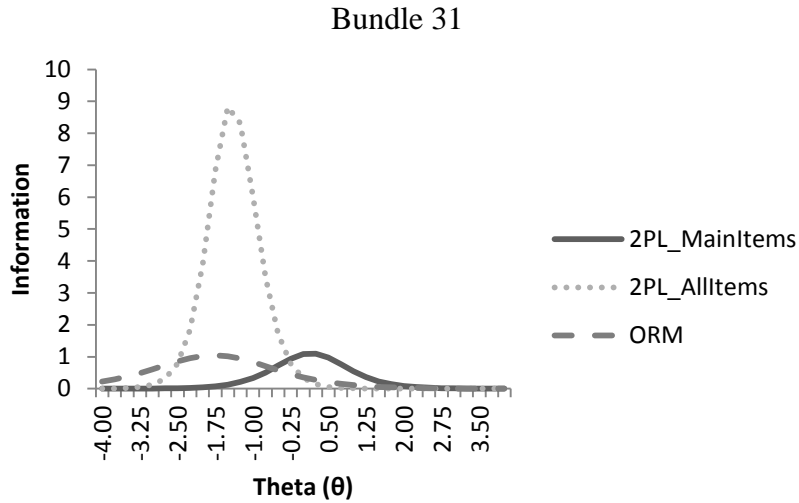


Figure 58. Information for Bundle 31. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.

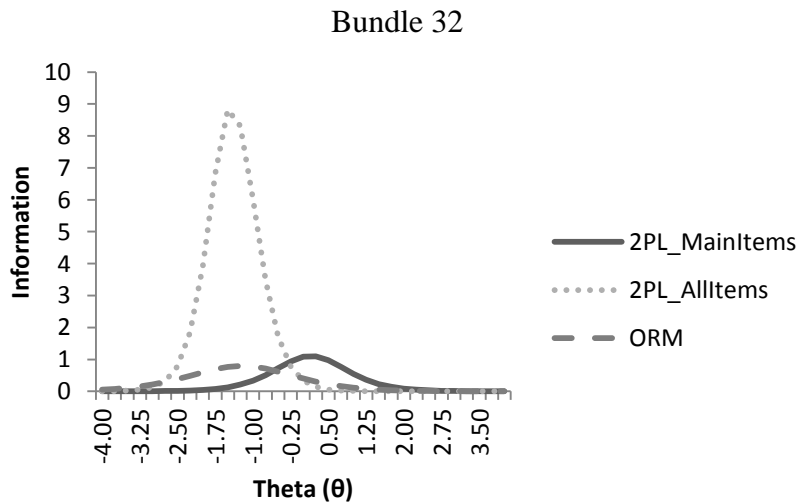


Figure 59. Information for Bundle 32. Bundle defined by a single main item in the 2PL_MainItems model, the main item plus scaffold items ignoring local dependence in the 2PL_AllItems model, or the summed scores of main and scaffold items in the ORM.