VALIDATION OF A COGNITIVE DIAGNOSTIC MODEL

ACROSS MULTIPLE FORMS OF A READING COMPREHENSION ASSESSMENT

by

Amy K. Clark

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: Neal Kingston

_____
William Skorupski

_____
Vicki Peyton

_____
Bruce Frey

_____
Phil McKnight

Date Defended: August 29, 2013

The Dissertation Committee for Amy Clark certifies that this is the approved version of the following dissertation:

VALIDATION OF A COGNITIVE DIAGNOSTIC MODEL

ACROSS MULTIPLE FORMS OF A READING COMPREHENSION ASSESSMENT

_____

Chairperson: Neal Kingston

Date Approved: August 29, 2013

Abstract

The present study sought to fit a cognitive diagnostic model (CDM) across multiple forms of a passage-based reading comprehension assessment using the attribute hierarchy method. Previous research on CDMs for reading comprehension assessments served as a basis for the attributes in the hierarchy. The two attribute hierarchies were fit to data from three forms of a large-scale assessment, one consisting of nine attributes and the other eleven attributes. Model-data fit of the two models was evaluated to determine if a single model of reading comprehension could be applied to multiple operational testing forms.

Keywords: cognitive diagnostic model, assessment, reading comprehension, attribute hierarchy

Acknowledgements

I would first like to extend my gratitude to Dr. Neal Kingston for serving as an exceptional mentor. His knowledge and passion for the field of measurement have served as a great inspiration during my doctoral career. I would also like to thank my committee members, Dr. William Skorupski, Dr. Bruce Frey, Dr. Vicki Peyton, and Dr. Phil McKnight for their thoughtful feedback on previous drafts of this work. Their insight has pushed me to think of the project from various angles and ultimately served to improve the final product presented here.

I would also like to thank the team of coders who dedicated their time to assisting with this project: Lauren Adams, Angela Broaddus, and Russell Swinburne. Their knowledge of the English language arts domain served as the foundation upon which all other work was based, and without their insight all aspects of the research conducted as a part of this project would have suffered.

In addition, I would like to thank Sukkeun Im for sharing his expertise in Fortran programming. He was not only willing to share previous iterations of code for fitting a model using the attribute hierarchy method, but was also available to answer questions that arose during the process. His collaboration and insight were both greatly appreciated.

Additionally, I would like to extend thanks to The College Board for providing me with the opportunity to participate in The College Board Fellowship Program, which served as the basis for this work. Their prompt distribution of data and test forms, as well as feedback on earlier drafts of the work was greatly appreciated. Without their assistance, the dissertation in its current format would not have been possible.

Finally, I would like to thank my fiancé and family for their unwavering love and encouragement as I completed my doctoral work. Their support helped to make the process all the more enjoyable!

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1 – Introduction

Background

In education, there has been an increasing demand for assessments that can provide greater information in terms of identifying and reporting examinee ability. Educators have expressed an interest in using the results of assessments to inform instruction (Huff & Goodman, 2007; Trout & Hyde, 2006), and researchers have urged the educational measurement field to move beyond reporting on a single latent construct towards richer reporting practices (Snow & Lohman, 1989).

In response to these demands, cognitive diagnostic modeling has become increasingly popular in the fields of cognitive psychology and educational measurement. One particular reason for attention in recent years is the unique ability to "diagnose" or identify examinee strengths and weaknesses with regard to the cognitive processes underlying performance on an assessment (Gierl, 2007; Yang & Embretson, 2007). As such, a wide audience values the information these models can provide. Testing organizations can use the models to provide more detailed score reports, and students, parents, and educators can use these reports to inform subsequent study and instructional approaches.

Because of this utility, operational testing programs have begun implementing cognitive diagnostic modeling with tests that are already in use. In this process, a cognitive model is retrofit to a test form using a specified hierarchy of skill acquisition for the particular domain. Once the model has been fit to the data, students are classified into knowledge states. These classifications can then be used to construct diagnostic score reports that provide detailed description of student levels of mastery by attribute.

While the implementation of diagnostic models can produce valid information regarding student ability, retrofitting a unique model for a single test form is both time consuming and impractical in large testing organizations. Rather than retrofitting a diagnostic model to a single test form at a time, operational testing programs would benefit from being able to apply a single cognitive model across multiple forms within a domain.

Purpose of the Study

A number of cognitive diagnostic models have previously been identified for passage-based items included on specific forms of the Critical Reading SAT (Sheehan, 1997; VanderVeen et al., 2007; Wang & Gierl, 2011). This study seeks to consolidate those models and apply them to items on the Critical Reading section of the PSAT. Rather than creating a new model from the items on a single test form, as is commonplace with retrofitting diagnostic models, cognitive models that have previously been validated in the domain of reading comprehension will be synthesized to create a diagnostic model that can be applied to multiple test forms. By implementing a comprehensive model of reading comprehension across multiple test forms, this study expands on the literature in the area of cognitive diagnostic models.

Research questions. The following research questions pertain to a passage-based reading comprehension assessment primarily administered to high school students.

1. Which hierarchy of cognitive skills related to responding to passage-based reading comprehension provides the best fit?
2. Can a single cognitive diagnostic model of reading comprehension be fit to multiple forms of a critical reading assessment using the attribute hierarchy method?

Hypotheses. It was hypothesized that the more parsimonious hierarchy proposed in this study will provide better fit to the data, as evidenced by classification to the knowledge states. It was also hypothesized that a single cognitive diagnostic model of reading comprehension based on that hierarchy would be established to have good model fit for multiple forms of a passage-based critical reading assessment.

Importance of the Study

This study expands on current research in a number of ways. First, validating a cognitive diagnostic model across multiple test forms provides test developers with validity evidence by accounting for the underlying cognitive processes represented in the Critical Reading section of the PSAT. A model validated for multiple test forms could also be applied to future test forms. In addition, the validation of the diagnostic model can also be used to create detailed score reports that reflect student strengths and weaknesses based on their probability of attribute mastery. These reports could be provided to students to use as a remediation template for improving areas of weakness prior to taking the SAT and in focusing their schooling efforts. Finally, having a validated cognitive diagnostic model that aligns with multiple Critical Reading forms will impact future iterations of the assessment by allowing item writers to develop new items that directly assess the specified cognitive processes and thus create more valid testing instruments in the future.

Summary

This study compared two models of passage-based reading comprehension: a more parsimonious and a slightly more complex model. Fit of the two models was compared in order to determine which model provided better fit to the data. In addition, model-data fit

was analyzed to determine if a single cognitive diagnostic model of reading comprehension can be fit to multiple test forms.

Chapter 2 – Review of the Literature

Cognitive Diagnostic Modeling

Cognitive diagnostic models first originated with Fischer (1973) and the use of the linear logistic test model to create a statistical model for a cognitive diagnostic assessment. During the 1980's the literature on cognitive diagnostic models continued to expand with Tatsuoka's (1983) development of the rule space model. However, it was not until Snow and Lohman (1989) wrote their seminal chapter on the interdependency of cognitive psychology with psychometrics that the area began to expand rapidly.

Since that time, a variety of names have been used to classify this group of models. Namely, Rupp and Templin (2008) prefer to refer to this class of models as diagnostic classification models, as they are intended to classify test takers into knowledge states that indicate their level of mastery of the construct. Despite the breadth of models available, all make use of observed response patterns for diagnostic purposes. Items are each associated with cognitive processes or "attributes" that are required for mastery. Students who have mastered a given attribute are predicted to respond correctly to items measuring that attribute. Likewise, students who have not mastered the attribute are predicted to respond incorrectly to items measuring the attribute. Based on item response patterns, diagnostic score reports can be provided that specify a student's level of mastery on each of the attributes.

Cognitive diagnostic assessment. There are several different ways to implement cognitive diagnostic models as a means of classifying test takers into knowledge states. Perhaps the most valuable implementation strategy in terms of classification consistency and diagnostic feedback involves the creation of a cognitive diagnostic assessment based

on the specified model. Using this approach, a series of attributes are determined that are required to solve problems within construct of interest. Once the attributes have been identified they are validated, often using think-aloud protocols (e.g. Wang & Gierl, 2011). The next step is to create a bank of assessment items to measure each attribute and assemble a test form. After the test form has been administered to examinees and model fit has been determined, student responses are then used to create detailed score reports based on the attributes in the model.

This approach of first specifying the attributes to be measured by the assessment, then developing items to assess the attributes has several benefits. First, writing items to the attributes ensures that an adequate number of items are available to assess each attribute. Furthermore, using this procedure, item writers can write distractors that are examples of common misconceptions associated with the construct. In situations where the examinee responds incorrectly, the response option selected provides additional information to be included in the diagnostic score report.

Retrofitting models. For large-scale assessments that are already in use, cognitive diagnostic models can be retrofit to existing test forms in an effort to identify underlying cognitive skills. The process is very similar to the procedure used to create a cognitive diagnostic assessment (Gierl & Cui, 2008; Leighton & Gierl, 2007; Rupp & Templin, 2008). The form is examined and the attributes necessary for solving the items are specified. The items are then coded for the attributes by multiple raters and inter-rater reliability is calculated. Next, the computational model is applied to the data. Fit is determined and attributes that do not have quality statistical properties are removed.

Previous studies have identified the following qualities as indicative of poor statistical quality when retrofitting models: too few items coded for the attribute, low or negative correlation of the attribute with total score, negative or non-significant Beta weight in a regression predicting total score, or if the attribute correlates too highly with another attribute (Buck, Tatsuoka, & Kostin, 1997; Buck & Tatsuoka, 1998; Svetina, Gorin, & Tatsuoka, 2011; VanderVeen et al., 2007). Should these examples be present, the model is revised and model fit is determined again. If the fit of the model is acceptable, diagnostic score reports are then constructed for examinees.

One complaint against retrofitting models to assessments already in use is that a Q matrix must be created based on the attributes that happen to be evident among the test items on the form (Gierl, 2007). With such an approach, items are typically examined for the content and skills required to demonstrate mastery. These skills then create the attributes that may or may not form a hierarchy of skill acquisition. However, the items alone inform the attribute selection, rather than being driven by theory or previous research. If the measure omits important skills or content, the attribute hierarchy does as well.

One way to improve upon the retrofitting approach is by first specifying the cognitive attributes necessary within a domain using theory and previous research. Items are then coded for evidence of those attributes, rather than the items forming the list of attributes. While still constrained by the limitations of retrofitting models, the attribute hierarchy is at least based on sound research. Several studies have employed the practice of first consulting the research, then modifying the attribute list based on characteristics of the items (Buck et al., 1997; Buck & Tatsuoka, 1998; Buck et al., 1998). Wang and Gierl

(2011) used an approach similar to this as they drew upon the actual attributes specified by VanderVeen et al. (2007), and ultimately expanded the cognitive model from five attributes to eleven. Similarly, de la Torre and Douglas (2008) used a Q matrix from Mislevy (1996), rather than develop a new set of attributes based on the item characteristics alone. Each of these examples demonstrates an improvement to the typical retrofitting technique by using a previously validated hierarchy of skills, rather than item characteristics alone, to retrofit a cognitive model to existing data.

A similar criticism related to retrofitting models to data from existing test forms relates to the inclusion of items for the diagnostic modeling process. Since the items were not originally intended for use as a part of a diagnostic assessment, items on a particular form may not all relate to a cohesive set of attributes. In some instances, only items that relate to the attributes of interest are selected from a pool of items, rather than modeling items from a form for diagnostic purposes (Embretson & Wetzel, 1987; Leighton, Cui, & Cor, 2009). Similarly, researchers may choose to include only those items that demonstrate high rater agreement when fitting a diagnostic model to the data (Wang, Gierl, & Leighton, 2006). While such approaches provide valuable information for future construction of diagnostic assessments, their utility for diagnostic reporting for the form as a whole is limited.

While there is great utility in retrofitting cognitive models to existing data for large-scale assessment programs, there are several other drawbacks associated with retrofitting models. One of the biggest issues is that the items were not originally written with the intention of classifying examinees into cognitive states. For this reason, the distractors may not necessarily provide additional diagnostic information regarding student

misconceptions. Furthermore, when using an existing test form, attributes can only be coded for the items that are already present, regardless of what is found in the research regarding the hierarchical nature of skill acquisition within that particular domain. This becomes an especially important issue when too few items are present to reliably assess the attribute. Each of the aforementioned limitations has an impact on model fit and classification consistency.

Rule space method. While there are a number of methods used for creating cognitive diagnostic models, the rule space method is one of the earliest models and also one of the most commonly used. Originally developed by Tatsuoka (1983), the model accounts for latent ability by comparing observed student responses to an ideal response pattern in order to estimate the probability that a student's responses were drawn from a particular knowledge state (Birenbaum & Tatsuoka, 1993; Tatsuoka, 1993, 2009). These comparisons result in the classification of each examinee to a knowledge state that allows for inferences to be made regarding the examinee's ability.

To determine these knowledge states, or expected response patterns, a series of matrices are created that explain the cognitive processes underlying the items (Tatsuoka, 1991, 1993). The adjacency (A) matrix, specified by k by k attributes, represents the direct relationships between attributes using binary coding. Placing a 1 in location (j, k) indicates attribute j is a prerequisite of attribute k. The reachability (R) matrix specifies both direct and indirect prerequisites for a given attribute. In the rule space method, it is not a requirement that attributes have prerequisites specified. The incidence (Q) matrix, consists of k attributes by p possible combinations of attributes, where $p = 2^k-1$. The Q matrix indicates all the possible combinations of attributes, without the influence of any hierarchy

on the structure. Finally, the reduced incidence (Q_r) matrix is created. It consists of only the vectors of the W matrix that follow the attribute hierarchy specified in the R matrix. In cognitive diagnostic assessment, the Q_r matrix serves as the test plan for item construction. Items can be specifically written to target each of the attribute combinations. To the extent the data fits the specified matrices, model fit will be strengthened or weakened.

Attribute hierarchy method. The attribute hierarchy method (AHM), developed by Leighton, Gierl, and Hunka (2004), was developed to expand on the rule space method. While the AHM is similar to the rule space method in that it follows a similar approach of comparing observed responses to expected responses derived from a series of matrices, the method differs in that it requires that attributes be ordered into a hierarchy. Since cognitive skills do not develop in isolation, but rather form an interconnected web of ability, the attribute hierarchy method more accurately reflects human cognition and the way skills are learned and applied.

The procedure for specifying matrices in the AHM follows the protocol outlined by Tatsuoka (1991, 1993), with the inclusion of the set of A, R, Q, and Q_r matrices. Whereas with the rule space method, the specification of a hierarchy of related skills is not required, the AHM necessitates that prerequisite skills be identified in the A and R matrices. This allows the attributes to be ordered into a hierarchy of dependent skills that can then be used to classify test takers into knowledge states.

When using the attribute hierarchy method to retrofit a diagnostic assessment to an existing form, the process has been applied to a single test form at a time. Attributes are specified then fit to student responses from the form to determine if classification consistency and model-data fit are evident. Most of the applications of the attribute

hierarchy method have been with data from the SAT. Two studies have fit an attribute hierarchy to data from the Critical Reading section of the March 2005 administration of the SAT (Wang & Gierl, 2007; Wang & Gierl, 2011). Similarly, the attribute hierarchy method has also been applied to data from the mathematics section of the March 2005 administration of the SAT (Gierl, Cui, & Hunka, 2007; Gierl, Leighton, et al., 2009; Gierl, Wang, & Zhou, 2008; Gierl, Zheng, & Cui, 2008; Leighton et al., 2009). Beyond the SAT, the attribute hierarchy method has also been retrofit to data from a 2002 administration of the College English Test in China (Wang et al., 2006). However, for purposes of generalizing to operational testing programs, future research should explore the extent that the attribute hierarchy method can be retrofit to multiple parallel or alternate test forms and be found to demonstrate good model-data fit and classification consistency for each.

Attribute coding. When a diagnostic model is retrofit to an existing assessment, it is necessary for raters to code each item to indicate whether or not the item requires a particular attribute in order for the examinee to provide a correct response. This process is in contrast to that used when creating a cognitive diagnostic assessment, where items are specifically written to assess each attribute. For the attribute hierarchy method in particular, items are coded for the direct attributes assessed by the item as well as their prerequisites.

Correct specification of the Q matrix is essential in order to establish model-data fit. When a Q matrix is incorrectly specified, model-data fit is impacted, which can result in low classification rates (Svetina et al., 2011), poor discrimination between masters and non-masters, (DiBello, Roussos, & Stout, 2007), spuriously high or low expected scores (Liu, Douglas, & Henson, 2009), or inflated slipping and guessing parameters (Junker & Sijtsma,

2001). In order to ensure that cognitive diagnostic models correctly classify examinees into knowledge states, it is imperative that the Q matrix be correctly coded. However, there is not a single agreed upon procedure for use in the coding process when retrofitting an attribute hierarchy to an existing assessment.

One way in which the coding approach differs by study is in the selection of coders. In many applications of diagnostic models to existing data, the authors of the study are included as coders (Buck et al., 1997; Gierl, Leighton, et al., 2009). While this approach may impact the results of the coding, the authors of one particular study specified that due to funding issues, outside coders could not be included (Buck & Tatsuoka, 1998). In contrast, other applications of diagnostic modeling to existing forms have made use of content experts for coding the items for required attributes (von Davier, 2008; Wang et al., 2006). Still others recruited graduate students to assist with the coding process (Jang, 2005; Wang & Gierl, 2011). Regardless of the approach taken for selecting coders, it is imperative that coders accurately assign attributes to items in order to ensure accuracy of the Q matrix.
An additional area that differs across studies with regard to coding is the number of coders that assign attribute to items. The most common number of raters is two (Birenbaum & Tatsuoka, 1993; Buck et al., 1998). One study in particular made use of five coders (Jang, 2005). In some instances, only a single rater is used to assign codes (Buck & Tatsuoka, 1998; Leighton et al., 2009). However, using a single coder could be especially problematic, as there is no evidence of inter-rater reliability or consensus among coders, attribute codes are confounded with the sole coder's level of consistency. For this reason, it is recommended that more than one coder be used to assign attribute codes to each item.

When multiple coders are used to assign attribute codes, the number of items coded by each coder sometimes differs. In some examples, each coder assigns attribute codes to all items (VanderVeen et al., 2007; Wang & Gierl, 2011). Although time consuming, this approach allows for the greatest potential inter-rater reliability because no items are excluded. In contrast, other studies make use of a single coder to code all items, while a second coder reviews the codes or independently coded a subset of the items (Svetina et al., 2011). While there are time benefits to having a second rater code only a subset of items or simply review the codes, including requiring less time and monetary resources, this approach may again impact the reliability of the codes assigned to items. Much like the situation in which only a single coder was implemented, the use of a second rater only coding a subset of the items could drastically impact the codes assigned to each item. In such a situation it becomes essential that the first coder assign codes to the items in a consistent manner.

When more than one rater is used to code items for cognitive attributes, indices of rater agreement are often calculated between the coders. Typically these raters code the attributes independently without discussing assignments, and the level of agreement is calculated after all items have been coded with their requisite attributes. The most commonly used metric when codes are dichotomously assigned is percent agreement (Birenbaum & Tatsuoka, 1993; Buck et al., 1998). When attributes are coded continuously, such as for count variables, Pearson correlation values may be used to provide an estimate of rater agreement (Buck & Tatsuoka, 1998). Other measures of rater agreement include Cohen's kappa for two raters, and Fleiss's kappa and intraclass correlation for groups of

raters. Each of these indices provides an estimate of the level of agreement between raters and can in some cases inform the final selection of items or attributes.

In contrast, rather than estimate levels of agreement, the goal of many studies is to attain a consensus regarding the coding of attributes to items. This consensus then forms the final Q matrix used to fit the diagnostic model to the data. In such cases, raters may independently code the items for attributes, but meet to discuss codes and reach an agreement for any items with discrepancies (Gierl, Leighton, et al., 2009; VanderVeen et al., 2007). While levels of agreement or inter-rater reliability are not reported, this approach allows for a single Q matrix to be used that retains all items and attributes, as disagreement is resolved prior to its construction.

A final way coding procedures has differed when retrofitting cognitive diagnostic models is with regard to the extent of the instructions provided to the coder(s). Coders may first meet to code a subset of items together in order to establish a common agreement regarding the application of attributes to the items (Buck et al., 1997; Wang & Gierl, 2011). In contrast, coders may be provided with a set of instructions without discussing the attributes or reaching an initial consensus regarding their application (Jang, 2005). The depth of instructions may also differ from including a simple instruction to code all present attributes to including a detailed set of coding instructions that contains examples and explanations of the attributes (Buck et al., 1997; Svetina et al., 2011). The level of instruction provided to the coders may impact their ability to accurately code items for attribute, ultimately impacting the specification of the Q matrix and model-data fit.

Fitting the attribute hierarchy model. Once the Q matrix has been established, an attribute hierarchy model can be fit to the data. This process begins by first constructing an attribute pattern matrix, which lists all possible combinations of attributes that provide a unique response pattern. The attribute pattern matrix is constructed by transposing the Q_r matrix and inserting a row of 0s at the top to represent the lack of mastery of any attributes in the hierarchy. Each subsequent row provides a unique combination of attributes that any given examinee may have mastered. This matrix informs the creation of a final matrix, an expected response matrix (E), which consists of the sequential expected item response pattern for a student having mastered any combination of attributes.

Following construction of the E matrix, the model can be fit to the data. Item response theory parameter estimates are obtained for the items included on the assessment. These values are then used to obtain ability estimates associated with each row of the E matrix, or each of the possible knowledge states an examinee could be classified into. Finally, the item parameter estimates and ability estimates associated with each knowledge state are used to fit the model to the data.

Determining model fit. Model fit for the AHM can be evaluated using the hierarchy consistency index (HCI; Cui, 2007). The HCI can be calculated with the following equation:

$$HCI_i = 1 - \frac{2 \sum_{j \in S_{correct_i}} \sum_{g \in S_j} X_{i_j}(1 - X_{i_g})}{N_{c_i}}, \quad (1)$$

where

$S_{correct_i}$ includes items that are answered correctly by examinee i,

X_{i_j} is examinee i's score (1 or 0) to item j,

S_j includes items that require the subset of attributes measured by item j,

X_{i_g} is examinee i's score (1 or 0) to item g where $g \in S_j$, and

N_{c_i} is the total number of comparisons for all the items that are answered correctly by examinee i.

The HCI compares the observed person responses to the Q_r matrix to determine the degree that observed responses align with the specified hierarchy. By dividing the number of slips, or instances the examinee responded incorrectly when expected to respond correctly, by the number of correct classifications, the possible values range from -1 to 1. These values are averaged over persons to obtain a value indicative of model-data fit. Leighton et al. (2009) specified the range of 0.8 to 1.0 as excellent fit, 0.6 to 0.8 as moderate fit, and values below 0.6 as indicative of poor fit.

Cognitive Processes for Reading Comprehension

There have been several diagnostic models constructed to account for the cognitive processes students engage in during passage-based reading comprehension assessments. Two such studies specifically examined reading processes examinees engaged in while responding to passage-based items on a second language reading assessment. Buck et al. (1997) identified 27 cognitive and linguistic attributes for 40 items on the Test of English for International Communication. Attributes included vocabulary skills, using background

knowledge, and making inferences. After classifying examinees to knowledge states using the rule space method, the list was reduced to 16 of attributes and 8 interactions. Similarly, Jang (2009) identified nine processing skills evident in the 76 items from field tests of a prototype of the Next Generation Test of English as a Foreign Language. Similarly, the nine attributes included such skills as discerning word meaning, making inferences, and selecting relevant information. Items were then analyzed using the Fusion Model to determine fit.

While the models in the previous two studies were specified for tests for non-native English speakers, similar attributes were identified for passage-based reading comprehension assessments that did not assess English as a second language. Svetina et al. (2011) specified 22 cognitive attributes that came from five larger skillsets: location, vocabulary, complex text processing, making inferences, and pragmatic and compensatory skills. These attributes were coded to items from a national exam used for scholarship award selection. After fitting the model with the rule space method, 7 attributes were removed from the model to improve classification consistency.

Two additional studies examining cognitive attributes required for passage-based reading comprehension employed a unique approach in that cognitive attributes were specified based on items sampled from a larger pool of items rather than discrete forms. Embretson and Wetzel (1987) selected 46 items from a larger pool of items for the Armed Services Vocational Aptitude Battery. The authors specified 15 cognitive variables to account for differences in the level of cognitive processing required by the items. These cognitive variables fell within the larger categories of propositional analysis and decision process. Items were analyzed using the linear logistic trait model to account for differences

in item difficulty. Gorin and Embretson (2006) expanded on this research, applying 10 cognitive variables to a collection of 200 items from released forms of the GR

Three studies specifically examined items from the SAT. Sheehan (1997) used a tree-based approach to provide diagnostic feedback for 78 verbal items on the SAT, 40 of which were passage-based reading comprehension items. A total of nine skills were identified as necessary components for responding to the items, including identifying an author's purpose, making inferences, and determining word meaning. Similarly, VanderVeen et al. (2007) analyzed SAT items for processes that would enable classification of test takers into instructionally relevant profiles. The researchers identified five large-grained text-processing skills that encompassed all 67 critical reading items on the SAT. These five areas included the following: word meaning, sentence meaning, situation model, global text meaning, and pragmatic meaning. Items were analyzed for these five dimensions of reading comprehension using confirmatory factor analysis rather than with a diagnostic modeling approach.

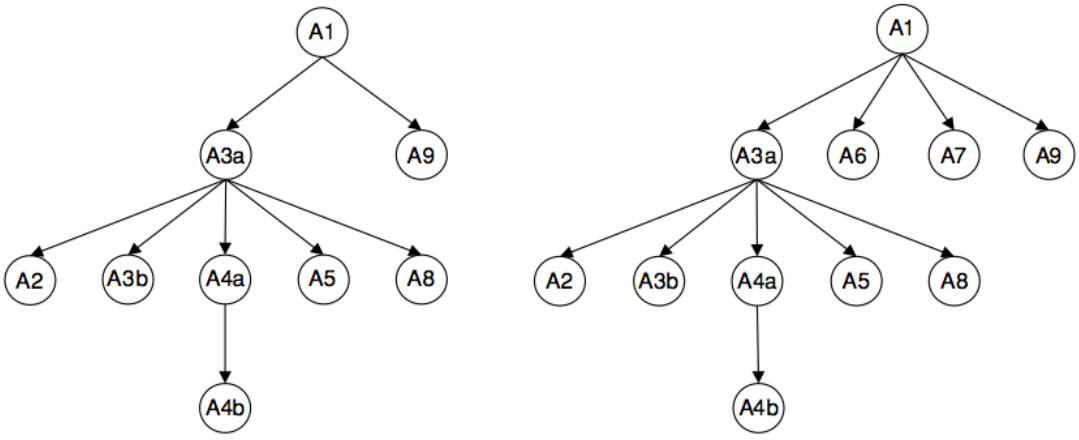
In the most recent analysis of SAT items, Wang and Gierl (2011) expanded on the five-attribute model specified by VanderVeen et al. (2007). The authors used the attribute hierarchy method to account for hierarchical reading processes on a shortened 20-item version of the Critical Reading section of the SAT. In addition, the two of the dimensions from the VanderVeen study were expanded to a more fine-grained level, and additional attributes were added based on think-aloud protocols with students. Three hierarchies were examined in the study: hierarchy one included 9 attributes, hierarchy two included 11 attributes, and hierarchy three included the 5 attributes included in the VanderVeen study. Table 1 lists the attributes included in these hierarchies. Hierarchy one contained

attributes A1-A5, A8 and A9, hierarchy two contained all the attributes listed, and hierarchy three included only attributes A1-A5, with a single attribute for A3 and a single attribute for A4. Figure 1 provides a pictorial representation of the hierarchies. After examining model-data fit through a cross-validation technique, the third hierarchy was eliminated. The remaining two hierarchies were found to have similar HCI fit values, between 0.66 and 0.68, indicating moderate model-data fit. For their study, the authors elected to use hierarchy one because the fit of the two models was equivalent and hierarchy one contained fewer attributes and thus represented a more parsimonious model.

Table 1

Summary of Cognitive Attributes

Number	Attribute
A1	Basic language knowledge, such as word recognition and basic grammar
A2	Determining word meaning by referring to context
A3a	Literal understanding of sentences with minimal amount of inference
A3b	Understanding sentences by making inferences based on the reader's experience and background knowledge
A4a	Literal understanding of larger sections of text with minimal amount of inference
A4b	Understanding larger sections of text by making inferences based on the reader's experience and world knowledge; building coherence across, summarizing, and evaluating larger sections of text
A5	Analyzing author's purposes, goals, and strategies
A6	Understanding text with difficult vocabulary
A7	Understanding text with complex syntactic structure
A8	Using rhetorical knowledge
A9	Evaluating response options



Hierarchy One Hierarchy Two

Figure 1. Attribute hierarchies of reading comprehension

The literature related to diagnostic models for passage-based reading comprehension assessments reveals many similarities in model specification. For each of the aforementioned assessments, between 9 and 24 attributes were included in the model, with an average of eight items associated with each attribute. Similarly, across the models for passage-based reading items, many of the same attributes listed in Table 1 were consistently identified for inclusion in the model, indicating that these attributes likely represent important constructs related to passage-based reading comprehension. One way to improve upon the attributes specified in Table 1 is to apply them to a larger number of items to determine if model-data fit can be improved.

Chapter 3 – Research Methods

Overview of Research

Using previous research on diagnostic models for reading comprehension as a guide, the attributes specified in Table 1 were used to construct a cognitive diagnostic model of reading comprehension. The two attribute hierarchies were fit to data from three forms of a large-scale assessment, and model-data fit was evaluated to determine if a single model of reading comprehension could be applied to multiple operational testing forms.

The previously outlined research was conducted in two stages. The first stage included specification of the attribute hierarchy and expert coding to associate attributes with the items. The second stage of research included psychometric analysis of the items and evaluation of model-data fit for each of the hierarchy by form combinations.

Instrumentation

A diagnostic model of reading comprehension was applied to the 35 items included on the Critical Reading section of the PSAT/National Merit Scholarship Qualifying Test (NMSQT). The section is speeded, allotting test takers 25 minutes to complete the passage-based, multiple-choice items. Scores from the measure are used as an initial screening for scholarships sponsored by the National Merit Scholarship Corporation (Marini, Mattern, & Shaw, 2011). The assessment also serves to prepare students for future administrations of the SAT. Score reports provide students with feedback to highlight areas of strength and those in need of improvement. In addition, score reports are made available to schools at the individual student and school level that provide overviews of student performance.

Participants

A random sample of 100,000 students was obtained from each of three administered forms of the PSAT NMSQT. For comparability, students were drawn from the pool of examinees that were administered the exam on October 14, 2009, October 13, 2010, and October 12, 2011. Examinees resided in all 50 states, Puerto Rico and Canada, among other locations. Demographic information such as grade level, gender, and ethnicity was also collected at the student level. Table 2 includes a summary of the demographic information.

Table 2

Summary of Demographic Data

Demographics		2009	2010	2011
Sex	Female	51,416	51,458	51,261
	Male	47,516	47,779	47,919
	No Response	1,068	763	820
Racial/ Ethnic Group	American Indian or Alaska Native	743	771	688
	Asian, Asian American, or Pacific Islander	6,992	6680	7052
	Black or African American	16,636	15,369	14,496
	Mexican or Mexican American	8,305	8403	9008
	Puerto Rican	1,949	1834	1857
	Other Hispanic, Latino, or Latin American	9,443	9394	9379
	White	47,670	46,240	46,877
	Other	3,641	3649	3614
	No Response	4,621	7660	7029
Grade Level	Not yet 8th	269	241	249
	Eighth	1,850	1521	1576
	Ninth	10,641	10,410	10,519
	Tenth	45,001	45,766	45,605
	Eleventh	41,013	41,530	41,366
	Twelve	175	156	133
	Other	22	16	16

No Response	1,029	360	536
Total	100,000	100,000	100,000

Human subjects approval was sought for this research study. The Human Subjects Committee Lawrence Campus reviewed the proposal and approved the research under the expedited procedure. See Appendix A for a copy of the Human Subjects Committee approval.

Variables

The variables included in this study vary by which hierarchy is imposed. Referencing Hierarchy 1, the independent variables in this study include the 9 attributes in the hierarchy: A1, A2, A3a, A3b, A4a, A4b, A5, A8, and A9. The dependent variables are the latent classes that examinees could be classified into based on their observed response pattern. Using Hierarchy 2, the independent variables in this study include the 11 attributes in the hierarchy: A1, A2, A3a, A3b, A4a, A4b, A5, A6, A7, A8, and A9. The dependent variables are the latent classes that examinees could be classified into based on their observed response pattern.

Stage One Procedures

Matrix specification. The review of the literature pertaining to CDMs constructed in the area of passage-based reading comprehension revealed the nine previously mentioned studies. Across these identified studies, the eleven attributes included in the study by Wang and Gierl (2011) were consistently identified. Table 3 includes a detailed breakdown of support for the attributes from other studies. For this reason, and since the hierarchies were validated using think-aloud protocols with examinees, the two

hierarchies specified by Wang and Gierl (2011) were selected for use in the current study. The attributes form two hierarchies, a more parsimonious model that includes nine cognitive attributes, and a more detailed model that includes those nine attributes as well as two more. The 11 attributes are listed in Table 1, and a visual representation of the two hierarchies is depicted in Figure 1.

Table 3

Support for Cognitive Attributes

Attribute	Description	Reference
A1	Basic language knowledge, such as word recognition and basic grammar	Wang & Gierl, 2011; VanderVeen et al., 2007
A2	Determining word meaning by referring to context	Wang & Gierl, 2011; VanderVeen et al., 2007; Svetina, 2011; Sheehan, 1997; Buck et al., 1998; Jang, 2009
A3a	Literal understanding of sentences with minimal amount of inference	Wang & Gierl, 2011; Sheehan, 1997; Jang, 2009
A3b	Understanding sentences by making inferences based on the reader's experience and background knowledge	Wang & Gierl, 2011; VanderVeen et al., 2007; Svetina, 2011; Buck, Tatsuoka, & Kostin, 1997; Jang, 2009
A4a	Literal understanding of larger sections of text with minimal amount of inference	Wang & Gierl, 2011; VanderVeen et al., 2007; Svetina, 2011; Buck, Tatsuoka, & Kostin, 1997; Buck et al., 1998; Jang, 2009
A4b	Understanding larger sections of text by making inferences based on the reader's experience and word knowledge; building coherence across, summarizing, and evaluating larger sections of text	Wang & Gierl, 2011; VanderVeen et al., 2007; Svetina, 2011; Buck, Tatsuoka, & Kostin, 1997; Jang, 2009
A5	Analyzing author's purposes, goals and strategies	Wang & Gierl, 2011; VanderVeen et al., 2007; Svetina, 2011; Sheehan, 1997; Jang, 2009
A6	Understanding text with difficult vocabulary	Wang & Gierl, 2011; Svetina, 2011; Buck, Tatsuoka, & Kostin, 1997; Buck et al., 2008; Embretson & Wetzel, 1987
A7	Understanding text with complex syntactic structure	Wang & Gierl, 2011; VanderVeen et al., 2007; Svetina, 2011; Buck et al., 1998; Embretson & Wetzel, 1987
A8	Using rhetorical knowledge (imagery, metaphor, parallelism)	Wang & Gierl, 2011; Sheehan, 1997
A9	Evaluating response options	Wang & Gierl, 2011; Buck, Tatsuoka, & Kostin, 1997; Embretson & Wetzel, 1987

Based on the two specified hierarchies, a set of matrices was constructed for use with the attribute hierarchy method. The A matrix for hierarchy one was of order 9 x 9, to indicate the direct prerequisites for each attribute based on the hierarchical structure specified in Figure 1. This matrix is shown in Figure 2 below. In looking at the figure, one can determine that attribute A2 requires A3a as a prerequisite. Similarly, the A matrix for hierarchy two was of order 11 x 11, as shown in Figure 3 that follows.

	A1	A2	A3a	A3b	A4a	A4b	A5	A8	A9
A1	0	0	1	0	0	0	0	0	1
A2	0	0	0	0	0	0	0	0	0
A3a	0	1	0	1	1	0	1	1	0
A3b	0	0	0	0	0	0	0	0	0
A4a	0	0	0	0	0	1	0	0	0
A4b	0	0	0	0	0	0	0	0	0
A5	0	0	0	0	0	0	0	0	0
A8	0	0	0	0	0	0	0	0	0
A9	0	0	0	0	0	0	0	0	0

Figure 2. A matrix for hierarchy one.

	A1	A2	A3a	A3b	A4a	A4b	A5	A6	A7	A8	A9
A1	0	0	1	0	0	0	0	1	1	0	1
A2	0	0	0	0	0	0	0	0	0	0	0
A3a	0	1	0	1	1	0	1	0	0	1	0
A3b	0	0	0	0	0	0	0	0	0	0	0
A4a	0	0	0	0	0	1	0	0	0	0	0
A4b	0	0	0	0	0	0	0	0	0	0	0
A5	0	0	0	0	0	0	0	0	0	0	0
A6	0	0	0	0	0	0	0	0	0	0	0
A7	0	0	0	0	0	0	0	0	0	0	0
A8	0	0	0	0	0	0	0	0	0	0	0
A9	0	0	0	0	0	0	0	0	0	0	0

Figure 3. A matrix for hierarchy two.

Next the R matrix was constructed for each of the two hierarchies to indicate the direct and indirect prerequisites for attributes in the model. The R matrix for hierarchy one was of order 9 x 9 as shown in Figure 4 that follows. In looking at the figure, one can determine that all attributes directly or indirectly require attribute A1, while no attributes require A9 as a direct or indirect prerequisite. The R matrix for hierarchy two was of order 11 x 11, as shown in Figure 5 that follows.

	A1	A2	A3a	A3b	A4a	A4b	A5	A8	A9
A1	1	1	1	1	1	1	1	1	1
A2	0	1	0	0	0	0	0	0	0
A3a	0	1	1	1	1	1	1	1	0
A3b	0	0	0	1	0	0	0	0	0
A4a	0	0	0	0	1	1	0	0	0
A4b	0	0	0	0	0	1	0	0	0
A5	0	0	0	0	0	0	1	0	0
A8	0	0	0	0	0	0	0	1	0
A9	0	0	0	0	0	0	0	0	1

Figure 4. R matrix for hierarchy one.

	A1	A2	A3a	A3b	A4a	A4b	A5	A6	A7	A8	A9
A1	1	1	1	1	1	1	1	1	1	1	1
A2	0	1	0	0	0	0	0	0	0	0	0
A3a	0	1	1	1	1	1	1	0	0	1	0
A3b	0	0	0	1	0	0	0	0	0	0	0
A4a	0	0	0	0	1	1	0	0	0	0	0
A4b	0	0	0	0	0	1	0	0	0	0	0
A5	0	0	0	0	0	0	1	0	0	0	0
A6	0	0	0	0	0	0	0	1	0	0	0
A7	0	0	0	0	0	0	0	0	1	0	0
A8	0	0	0	0	0	0	0	0	0	1	0
A9	0	0	0	0	0	0	0	0	0	0	1

Figure 5. R matrix for hierarchy two.

Since a retrofitting approach was used in the current study, the Q matrix was generated using an item-by-attribute coding process rather than generating the matrix from all possible combinations of attributes by items, as would be performed in the case of

creating a diagnostic assessment. The process for coding the items for the cognitive attributes required for mastery is described in the section that follows.

Coding process. Raters were recruited to participate in the process of coding the items for whether or not they require mastery of any of the 11 attributes in order to provide a correct response. Outside raters will be included in the study in lieu of the researcher coding all items as an effort to eliminate any potential bias that may interfere with the coding process. Rather than using a single rater, multiple raters will be included in the coding process to ensure consensus is reached regarding the coding for each item.

Three coders were selected due to their roles in English language arts test development for a Midwestern operational testing center offering K-12 accountability assessments. As the measure included in the current study is created for examinees in middle and high-school, the coders have knowledge and familiarity with assessment items that are of similar content and difficulty.

In preparation for the coding process, the researcher prepared test booklets for each of the three forms. The booklets included the 35 passage-based items and their respective passages. Items that were not passage-based, such as analogy items, are not a focus of the current study and thus were not included in the booklets. A codebook was also created that included the 11 cognitive attributes, the description, along with an expanded explanation of each attribute. Since the researcher did not develop the cognitive attributes, but rather drew from previous research, descriptions of the meaning of each attribute were compiled from the literature. In addition to the codebook, the researcher prepared an answer key for each of the three forms and a coding spreadsheet for use during the rating sessions.

A series of four meetings were held with the three selected content experts and the researcher. The first meeting began with a brief overview of the study and the purpose for the coding. The researcher stressed that the purpose of coding was to determine whether or not an item *requires* the cognitive attribute in order to obtain a correct response, as opposed to whether or not an attribute *could* be used to correctly respond to the item. Next, the group reviewed the list of 11 cognitive attributes and discussed the meaning of each. Although the researcher facilitated the meeting, additional input regarding the coding was not provided. The three coders were encouraged to share their interpretations with one another in order to come to a consensus as to how they knew when to code the particular attribute as a requirement for a correct response.

After a consensus was reached regarding the meaning of the cognitive attributes, the group prepared to code the items. All coders were provided with the testing booklet for form A of the assessment. The raters began with the first item, reading the item and necessary parts of the passage to determine which of the cognitive attributes are required for a correct response. The researcher will share the keyed answer for the item with the raters. Each coder then independently coded the first item for each of the cognitive attributes, and then the group discussed the codes. Beginning with the first attribute on the list, the coders discussed whether or not the attribute was required for a correct response to the item. Any areas of disagreement were discussed, and group members explained divergent thinking. Once a consensus was reached, the researcher documented the decision on the coding spreadsheet, recording a 0 if the item did not require the attribute or a 1 if the item did require the attribute. The group then began discussing the next attribute, moving through the complete list of attributes. This process was repeated for each item on

the form. Subsequent meetings were held following the same procedure to reach coding consensus for forms B and C.

Following the meetings to obtain the coding for the cognitive attributes required by the items the coding spreadsheet was used to construct the final attribute matrix. The codes were modified to reflect the hierarchical nature of skill acquisition, as specified by the direct and indirect prerequisites in the R matrix. Cells were examined to ensure that a 1 was present for attributes that are a prerequisite skill of an attribute coded as 1. For example, if A3a is coded as 1, the matrix needs to reflect that A1 is a prerequisite and as such, is also coded as a 1. In addition, a second coding sheet was constructed to only include the attributes included in Hierarchy 1. The final Q matrices based on the expert coding process are presented in Tables 5 – 9.

Stage Two Procedures

Expected response generation. Once the items were coded for the cognitive attributes, expected response patterns were generated across each of the three test forms for both hierarchies. First, an expected response matrix was constructed for form A reflecting hierarchy one using an Excel spreadsheet. The 35 items were listed in sequential order in column A. The cognitive attributes required by each item were specified in column B. This included between one and nine attributes. Next the items were sorted in ascending order from items that required the fewest attributes to items that required the most attributes for mastery. Items that required identical attributes were grouped together. For example, the items in the first two lines may require only A1, while the last item may require mastery of all nine attributes. A new column, C, was created to document the items a student would correctly respond to if they had only mastered the attributes specified in

column B. In the previous example, the first two items required only A1, so column C would list item 1 and 2 to indicate that a person having only mastered A1 is expected to respond correctly to only those two items. Moving down the list, items require gradually more attributes for mastery. Continuing the example, the next item might require A3a as well as A1, so column C would list that item, as well as 1 and 2 to indicate that a student having mastered A1 and A3a is expected to correctly respond to three items. See Table 4 for an example. This process was conducted for all 35 items on all three forms, reflecting both hierarchy one and hierarchy two.

Table 4

Example of Spreadsheet Used to Construct Item by Attribute Hierarchy

Item	Attributes required	Items answered with attributes required
1	A1	1, 2
2	A1	1, 2
44	A1 A3a	1, 2, 44
31	A1 A3a A3b	1, 2, 44, 31
35	A1 A3a A3b A9	1, 2, 44, 31, 35, 48
48	A1 A3a A3b A9	1, 2, 44, 31, 35, 48

After all 35 items were documented on the spreadsheet, the list of possible attribute combinations was expanded using the Q_r matrix. Recall that the Q_r matrix includes all possible combinations of attributes given the constraints of the attribute hierarchy. However, not all of these combinations were observed given the codes assigned by the content. For example, if the raters coded the items and consistently across the form A9 was

always coded with A3b, then that particular form would not be able to provide accurate feedback regarding mastery of A9 in the absence of A3b. For this reason, the Q_r matrix had to be further narrowed for each form to only include those attribute combinations that were possible given the nature of the codes for each item. Due to the potential for researcher error in listing all possible combinations of cognitive attributes, the final set of expected response patterns was double checked for accuracy and to ensure that no duplicates were included in the list.

Once all the possible attribute combinations were identified, the expected response matrix was constructed for each form and hierarchy combination. The matrix for each form included 35 columns and a number of rows equal to the number of unique attribute combinations identified for the form. Each row consisted of zeros and ones to represent the expected response pattern across the 35 items given a particular combination of cognitive attributes. Each unique pattern of zeros and ones represented a knowledge state an examinee might be classified into. The first row of the expected response matrix included 35 zeros to represent a student that has mastered none of the attributes and thus responds incorrectly to all items. Similarly, the last row of the matrix included 35 ones to represent a student who has mastered all attributes.

This process of constructing all possible attribute combinations was then repeated for form B and C using hierarchy one. Then the entire process was repeated again for forms A, B, and C using hierarchy two. In total, six unique sets of expected response patterns were generated.

Assignment to groups. For each of the three test forms, participants were randomly assigned to two groups of 10,000 examinees in order to cross validate the HCI

values of model fit. Random assignment to groups was conducted using Excel with the full file of 100,000 examinees. All examinees who took Form A were assigned a random number using random number generation. The file was sorted on the random numbers to ensure cases were in a random order. Then the file was split in half and 10,000 examinees were selected from the first half. After removing the remaining cases from the first half, the process was repeated to obtain a second group of 10,000 examinees from the second half of the total 100,000 examinees. This process was then repeated for Forms B and C. For each form, these two groups were used to cross validate the fit of the two hierarchies to the data. A total of twelve analyses were conducted during the data analysis process. Figure 6 shows the organizational structure of the 12 analyses that were conducted as a part of this study.

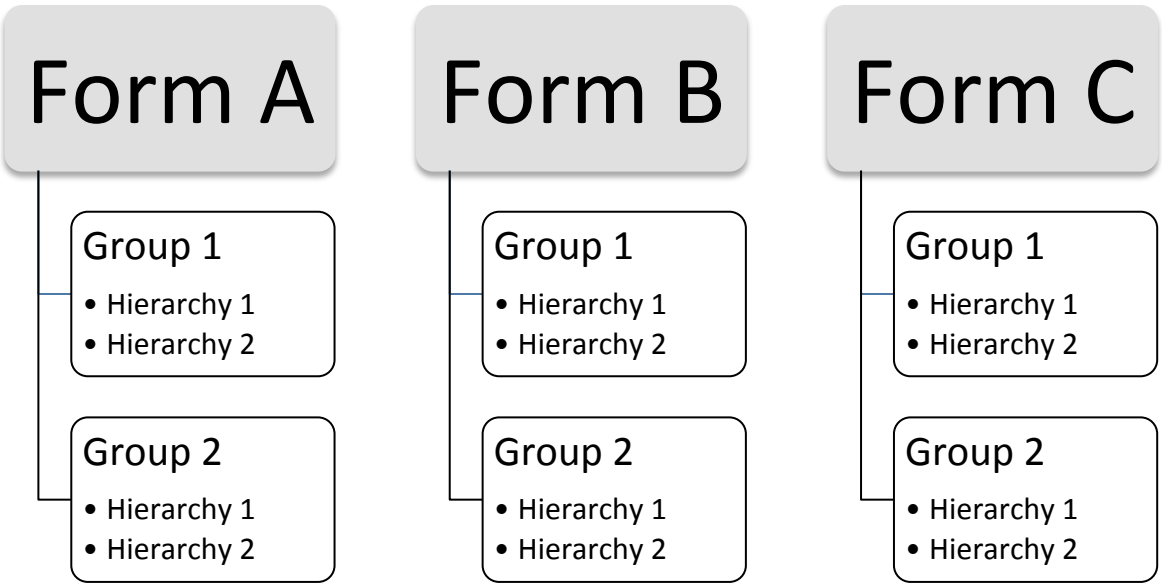


Figure 6. Organization of analyses conducted by form, group, and hierarchy

Item response theory. Prior to obtaining ability estimates for examinees using item response theory, the data were explored to determine if the assumptions of item response theory were met. Specifically, an exploratory factor analysis was conducted using SPSS to confirm that a single dimension underlies performance on the 35 items. To confirm a unidimensional model, the scree plot for each hierarchy by form combination was examined. In addition, the ratio of adjacent Eigenvalues was compared.

Next, model-data fit for the two- and three-parameter logistic models was compared. Both models were fit to the data for each hierarchy by form combination using IRTPRO (Cai, Thissen, & du Toit, 2011). Two compare model-data fit, AIC and BIC values were recorded.

Using the model that was determined to provide better fit to the data, the parameter estimates obtained from fitting the data from each form. All item parameter estimates were computed using maximum likelihood estimation. Item parameter values were saved to a separate output file. Scoring was done using expected a posteriori estimation.

Once the item parameter estimates were obtained, they were used to obtain ability estimates for each of the knowledge states associated with each form using IRTPRO. Ability estimates for each of the knowledge states were computed using expected a posteriori estimation. A total of six sets of ability estimates were computed for the knowledge states associated with each hierarchy across the three forms.

Fitting the attribute hierarchy model. In order to fit the diagnostic model using the attribute hierarchy method, a program was written in Fortran. The program requires three input files when fitting the model for each test form. The expected response file includes a column with the attributes associated with the knowledge state, followed by the

response pattern associated with that knowledge state, and finally the ability estimate associated with the knowledge state. A second file, the data file, includes the response pattern for each examinee. Finally, a parameter file is included that contains the *a*, *b*, and *c* parameter values for each item. The Fortran program reads in the three files and writes several output files, including a difference file, a probability file, an expected theta file, and a person by probability file.

First the difference file was created. The program loops over persons, items, and knowledge states to determine the difference between the expected and observed responses, and writes these values to a file. In the file, each person will has *k* rows, equal to the number of possible knowledge states. Each row consists of 35 values, representing each item. Each of the 35 columns includes a -1, 0, or 1. A value of -1 indicates an instance where the student responded correctly when predicted to respond incorrectly, or the equivalent of a guess. A value of 1 indicates an instance where the student responded incorrectly when predicted to respond correctly, or the equivalent of a slip. A 0 indicates an instance where the student responded as predicted. The probabilities were calculated in the following way: If the difference = -1, then the probability = item difficulty. If the difference = 1, then the probability = 1 – item difficulty. If the difference is 0, then the probability = 0.

The difference file was then used to create a probability file, which includes the probability of a correct response to each item for each person by *k* knowledge states. For example, person one was associated with *k* rows by 35 columns, with the cell values representing the probability of a correct response for that item, given the knowledge state.

Next the expected theta file was created. This file lists each student's observed response pattern, followed by the theta and probability associated with each of the

knowledge states. The value with the highest probability reflects the knowledge state the student will be classified into. Finally, this file was condensed into the person by probability file, which will consist of *j* persons by *k* knowledge states. Each row lists the person ID and the probability values for each of the *k* knowledge states.

Comparison of model-data fit. Finally the fit of each of the attribute hierarchies to the data was evaluated. First, classification consistency of the model was calculated to determine the proportion of examinees that were successfully classified to a single knowledge state. To be successfully classified to a single knowledge state, one probability value would be larger than all others. In those instances where an examinee had an equal probability of falling into two or more knowledge states, the model did not successfully identify their knowledge state, thus resulting in a lower classification consistency value.

Fit of the attribute hierarchies to the data was also evaluated using the HCI. For each test form, the two samples of 10,000 students were used to cross-validate the classification consistency and HCI values. Classification consistency and HCI values for each test form were then compared to determine which model provided better fit the data and whether a single attribute hierarchy was effective for use with multiple test forms.

Assumptions of the Study

There are several assumptions associated with the attribute hierarchy method (Sinharay, Puhan, & Haberman, 2009). First, the method assumes that providing a correct response to an item requires one or more cognitive processes. Second, the method assumes that these cognitive processes can be organized into a hierarchy. Third, it is assumed that a latent ability parameter can be estimated for each examinee. Finally, the method assumes

that an examinee's observed responses can be compared with expected responses to determine level of mastery for each cognitive process.

Limitations of the Study

One limitation associated with the outlined research regards the construction of a model of the cognitive processes that underlie existing items. This type of model may not be as encompassing as a model that would be used as the foundation for developing a cognitive diagnostic assessment. Another potential limitation associated with the use of a post-hoc model is that model specification or fit may be impacted, particularly if there are not an adequate number of items associated with the identified cognitive processes (Gierl & Cui, 2008; Gierl, Wang, et al., 2008; Rupp & Templin, 2008). However, basing the model on sound cognitive theory and incorporating all 35 passage-based items on the Critical Reading section of the PSAT may mitigate this effect.

Internal validity. The extent to which this study has internal validity depends on the accuracy of the coding of items for attributes required for mastery. To the extent that the items were accurately coded for the correct attributes, internal validity will be high, as the independent variables (attributes) will be linked with the dependent variables (knowledge states). To the extent that items were coded inaccurately for the attributes that are required for mastery, there may be a lack of alignment between the independent variables and the dependent variables, resulting in lower internal validity. Potential threats to internal validity were addressed through the use of random assignment to groups.

External validity. This study sought to fit a single diagnostic model of reading comprehension to multiple forms of a passage-based reading comprehension assessment that was administered to mostly high school students. Multiple time points, individuals,

and administration settings were included in the research design to increase the generalizability to future use of the measure. By fitting a single model across the three test forms, the model could potentially be generalized to additional forms of the assessment and perhaps different assessments that also assess the construct of passage-based reading comprehension. Additional research will need to be conducted in order to determine if a single diagnostic model can be applied to multiple forms of assessments in different domains.

Positive results. It was anticipated that there were two possible positive results that could occur while conducting this study. First, one attribute hierarchy would be identified as having the best fit to the data. Second, a single hierarchy would be found to have good fit across all three of the test forms. It was anticipated that if positive results were found, the single model for passage-based reading comprehension could be applied to additional test forms, and such a model could be used in the future to provide test takers with detailed score reports regarding their proficiency on each of the cognitive attributes.

Negative results. Similarly, it was anticipated that there were two possible negative results that could occur as a result of this study. First, poor model-data fit could occur on any of the test forms, which would indicate that neither hierarchy accurately explained the cognitive processes test takers engaged in while being administered this assessment. Second, a single hierarchy might not have been found to have good model-data fit across the three test forms. This was possible because each form had to be separately coded for the attributes each item requires for a correct response. If any form lacked internal validity, model-data fit would likely be impacted.

Summary

This study built on the relevant literature pertaining to cognitive diagnostic modeling of passage-based reading comprehension by combining several previously used approaches. By selecting attributes found to successfully identify underlying cognitive processes in previous applications, the model was expected to more validly reflect the cognitive processes than by using the typical approach of coding only the items on a single test form. Furthermore, by validating a model across multiple test forms, this method was intended to be beneficial to the College Board by providing information on the effectiveness of a cognitive model that could be widely applied in operational situations.

Chapter 4 - Results

Stage One

Review of relevant research yielded two hierarchies representing the acquisition of passage-based reading comprehension, including one parsimonious model and one slightly more complex model. These two models were previously presented in Table 1 and Figure 1. Based on these hierarchies, the R and A matrices were defined to demonstrate the direct and indirect prerequisites of each attribute. Next the Q matrix was specified, which required the items to be coded for the attributes necessary to provide a correct response to the item.

Attribute coding. Due to the retrofitting nature of the current study, content experts were recruited to associate the items with the cognitive attributes required to provide a correct response to the item. During the process of coding the items, the three content experts were able to reach an agreement as to the attributes required for each item on each of the three forms of the assessment. These codes were combined with the previously specified prerequisites in the A and R matrices to create a unique Q matrix for each form. The final expert coding Q matrices are presented on the following pages for hierarchies one and two. Note that the two hierarchies are the same with the exception that hierarchy two contains two additional attributes.

Table 5

Q Matrix for Form A Hierarchy One

Item	A1	A2	A3a	A3b	A4a	A4b	A5	A8	A9
A_9	1	0	1	0	0	0	0	0	0
A_10	1	0	1	0	0	0	1	0	0
A_11	1	0	1	0	0	0	1	0	0
A_12	1	0	1	0	0	0	1	0	1
A_13	1	0	1	0	0	0	1	0	1
A_14	1	0	1	0	0	0	0	0	0
A_15	1	0	1	0	1	0	1	0	1
A_16	1	0	1	0	0	0	1	0	1
A_17	1	0	1	0	1	1	1	0	1
A_18	1	1	1	0	0	0	1	0	1
A_19	1	1	1	0	0	0	1	0	1
A_20	1	1	1	0	0	0	1	0	1
A_21	1	1	1	0	0	0	0	0	0
A_22	1	0	1	0	1	0	0	0	1
A_23	1	0	1	0	0	0	0	0	1
A_24	1	0	1	0	1	0	0	0	1
A_30	1	0	1	0	0	0	0	0	1
A_31	1	0	1	0	0	0	0	0	1
A_32	1	0	1	0	0	0	0	0	0
A_33	1	0	1	1	0	0	0	0	1
A_34	1	0	1	0	1	0	0	0	1
A_35	1	0	1	0	0	0	0	0	1
A_36	1	0	1	0	0	0	0	0	1
A_37	1	0	1	1	0	0	0	0	1
A_38	1	0	1	0	0	0	0	0	1
A_39	1	0	1	0	0	0	1	1	1
A_40	1	0	1	0	0	0	0	0	0
A_41	1	0	1	0	1	1	0	0	1
A_42	1	0	1	0	0	0	1	0	0
A_43	1	0	1	0	0	0	1	0	0
A_44	1	1	1	0	0	0	0	0	0
A_45	1	0	1	0	1	0	1	0	1
A_46	1	0	1	0	0	0	1	0	1
A_47	1	0	1	0	0	0	0	0	1
A_48	1	0	1	0	0	0	0	0	1

Table 6

Q Matrix for Form B Hierarchy One

Item	A1	A2	A3a	A3b	A4a	A4b	A5	A8	A9
B_9	1	1	1	0	1	1	1	1	1
B_10	1	0	1	0	0	0	1	1	1
B_11	1	0	1	0	0	0	0	0	0
B_12	1	0	1	0	0	0	1	1	1
B_13	1	0	1	0	1	0	1	0	1
B_14	1	0	1	1	0	0	1	0	1
B_15	1	0	1	0	1	0	1	0	1
B_16	1	0	1	0	0	0	0	0	1
B_17	1	0	1	0	1	1	1	1	1
B_18	1	1	1	0	1	0	0	0	1
B_19	1	0	1	1	0	0	0	0	1
B_20	1	1	1	0	0	0	1	0	1
B_21	1	0	1	0	0	0	1	0	1
B_22	1	1	1	0	0	0	0	0	0
B_23	1	0	1	0	1	1	1	0	1
B_24	1	0	1	0	1	1	0	0	1
B_30	1	0	1	0	0	0	0	0	0
B_31	1	0	1	0	0	0	0	0	0
B_32	1	0	1	0	0	0	0	0	0
B_33	1	0	1	0	1	0	0	0	1
B_34	1	0	1	0	1	0	0	0	1
B_35	1	0	1	1	0	0	0	0	1
B_36	1	1	1	0	1	0	0	0	1
B_37	1	0	1	0	1	0	0	0	0
B_38	1	0	1	0	1	0	0	0	0
B_39	1	0	1	0	1	0	1	0	1
B_40	1	0	1	0	1	0	0	0	1
B_41	1	0	1	0	1	1	1	0	1
B_42	1	0	1	0	0	0	1	0	1
B_43	1	0	1	1	0	0	0	0	1
B_44	1	0	1	0	0	0	0	0	1
B_45	1	1	1	0	0	0	0	0	1
B_46	1	0	1	1	0	0	0	0	1
B_47	1	0	1	0	0	0	1	0	1
B_48	1	0	1	0	0	0	0	0	0

Table 7

Q Matrix for Form C Hierarchy One

Item	A1	A2	A3a	A3b	A4a	A4b	A5	A8	A9
C_9	1	0	1	0	0	0	1	0	0
C_10	1	0	1	0	1	0	1	0	1
C_11	1	0	1	0	1	0	1	0	1
C_12	1	0	1	0	1	1	0	0	1
C_13	1	0	1	0	1	0	0	0	0
C_14	1	0	1	0	1	0	0	0	1
C_15	1	1	1	0	0	0	0	0	1
C_16	1	1	1	0	0	0	0	0	1
C_17	1	0	1	0	0	0	0	0	0
C_18	1	0	1	0	1	0	0	1	0
C_19	1	0	1	0	0	0	0	0	0
C_20	1	0	1	1	0	0	0	0	1
C_21	1	0	1	0	1	0	0	0	1
C_22	1	1	1	0	0	0	0	0	1
C_23	1	1	1	0	0	0	0	0	1
C_24	1	0	1	0	0	0	0	0	0
C_30	1	0	1	1	0	0	0	0	1
C_31	1	0	1	0	0	0	0	1	0
C_32	1	0	1	0	0	0	0	0	0
C_33	1	0	1	0	0	0	0	0	0
C_34	1	0	1	0	1	0	0	0	1
C_35	1	0	1	0	1	0	0	0	1
C_36	1	0	1	0	0	0	0	0	0
C_37	1	0	1	0	1	0	0	0	0
C_38	1	1	1	0	0	0	0	0	1
C_39	1	0	1	0	0	0	0	0	1
C_40	1	0	1	0	0	0	0	0	0
C_41	1	1	1	0	0	0	0	0	1
C_42	1	0	1	0	0	0	0	0	1
C_43	1	0	1	0	1	1	0	0	1
C_44	1	0	1	0	1	1	0	0	1
C_45	1	0	1	0	0	0	0	1	1
C_46	1	0	1	0	1	0	0	0	0
C_47	1	0	1	1	0	0	0	0	1
C_48	1	0	1	0	1	1	0	0	1

Table 8

Q Matrix for Form A Hierarchy Two

Item	A1	A2	A3a	A3b	A4a	A4b	A5	A6	A7	A8	A9
A_9	1	0	1	0	0	0	0	0	0	0	0
A_10	1	0	1	0	0	0	1	0	0	0	0
A_11	1	0	1	0	0	0	1	0	0	0	0
A_12	1	0	1	0	0	0	1	0	0	0	1
A_13	1	0	1	0	0	0	1	0	0	0	1
A_14	1	0	1	0	0	0	0	0	0	0	0
A_15	1	0	1	0	1	0	1	0	0	0	1
A_16	1	0	1	0	0	0	1	1	0	0	1
A_17	1	0	1	0	1	1	1	0	0	0	1
A_18	1	1	1	0	0	0	1	0	0	0	1
A_19	1	1	1	0	0	0	1	0	0	0	1
A_20	1	1	1	0	0	0	1	1	0	0	1
A_21	1	1	1	0	0	0	0	0	0	0	0
A_22	1	0	1	0	1	0	0	0	0	0	1
A_23	1	0	1	0	0	0	0	1	0	0	1
A_24	1	0	1	0	1	0	0	0	0	0	1
A_30	1	0	1	0	0	0	0	0	0	0	1
A_31	1	0	1	0	0	0	0	0	1	0	1
A_32	1	0	1	0	0	0	0	0	0	0	0
A_33	1	0	1	1	0	0	0	0	0	0	1
A_34	1	0	1	0	1	0	0	0	0	0	1
A_35	1	0	1	0	0	0	0	0	0	0	1
A_36	1	0	1	0	0	0	0	0	0	0	1
A_37	1	0	1	1	0	0	0	0	1	0	1
A_38	1	0	1	0	0	0	0	0	0	0	1
A_39	1	0	1	0	0	0	1	0	0	1	1
A_40	1	0	1	0	0	0	0	0	0	0	0
A_41	1	0	1	0	1	1	0	1	0	0	1
A_42	1	0	1	0	0	0	1	0	0	0	0
A_43	1	0	1	0	0	0	1	0	0	0	0
A_44	1	1	1	0	0	0	0	0	0	0	0
A_45	1	0	1	0	1	0	1	0	0	0	1
A_46	1	0	1	0	0	0	1	0	0	0	1
A_47	1	0	1	0	0	0	0	0	1	0	1
A_48	1	0	1	0	0	0	0	0	0	0	1

Table 9

Q Matrix for Form B Hierarchy Two

Item	A1	A2	A3a	A3b	A4a	A4b	A5	A6	A7	A8	A9
B_9	1	1	1	0	1	1	1	0	0	1	1
B_10	1	0	1	0	0	0	1	0	0	1	1
B_11	1	0	1	0	0	0	0	0	0	0	0
B_12	1	0	1	0	0	0	1	0	0	1	1
B_13	1	0	1	0	1	0	1	0	1	0	1
B_14	1	0	1	1	0	0	1	0	0	0	1
B_15	1	0	1	0	1	0	1	0	1	0	1
B_16	1	0	1	0	0	0	0	0	0	0	1
B_17	1	0	1	0	1	1	1	0	0	1	1
B_18	1	1	1	0	1	0	0	0	0	0	1
B_19	1	0	1	1	0	0	0	0	0	0	1
B_20	1	1	1	0	0	0	1	0	0	0	1
B_21	1	0	1	0	0	0	1	0	0	0	1
B_22	1	1	1	0	0	0	0	0	0	0	0
B_23	1	0	1	0	1	1	1	0	0	0	1
B_24	1	0	1	0	1	1	0	0	0	0	1
B_30	1	0	1	0	0	0	0	0	0	0	0
B_31	1	0	1	0	0	0	0	0	0	0	0
B_32	1	0	1	0	0	0	0	0	0	0	0
B_33	1	0	1	0	1	0	0	0	1	0	1
B_34	1	0	1	0	1	0	0	0	0	0	1
B_35	1	0	1	1	0	0	0	0	0	0	1
B_36	1	1	1	0	1	0	0	0	0	0	1
B_37	1	0	1	0	1	0	0	0	0	0	0
B_38	1	0	1	0	1	0	0	0	0	0	0
B_39	1	0	1	0	1	0	1	0	0	0	1
B_40	1	0	1	0	1	0	0	0	0	0	1
B_41	1	0	1	0	1	1	1	0	0	0	1
B_42	1	0	1	0	0	0	1	0	0	0	1
B_43	1	0	1	1	0	0	0	0	0	0	1
B_44	1	0	1	0	0	0	0	1	0	0	1
B_45	1	1	1	0	0	0	0	0	0	0	1
B_46	1	0	1	1	0	0	0	0	0	0	1
B_47	1	0	1	0	0	0	1	0	0	0	1
B_48	1	0	1	0	0	0	0	0	0	0	0

Table 10

Q Matrix for Form C Hierarchy Two

Item	A1	A2	A3a	A3b	A4a	A4b	A5	A6	A7	A8	A9
C_9	1	0	1	0	0	0	1	0	0	0	0
C_10	1	0	1	0	1	0	1	0	0	0	1
C_11	1	0	1	0	1	0	1	0	0	0	1
C_12	1	0	1	0	1	1	0	0	0	0	1
C_13	1	0	1	0	1	0	0	0	0	0	0
C_14	1	0	1	0	1	0	0	0	0	0	1
C_15	1	1	1	0	0	0	0	0	0	0	1
C_16	1	1	1	0	0	0	0	0	0	0	1
C_17	1	0	1	0	0	0	0	1	0	0	0
C_18	1	0	1	0	1	0	0	0	0	1	0
C_19	1	0	1	0	0	0	0	0	0	0	0
C_20	1	0	1	1	0	0	0	0	0	0	1
C_21	1	0	1	0	1	0	0	0	0	0	1
C_22	1	1	1	0	0	0	0	0	0	0	1
C_23	1	1	1	0	0	0	0	0	0	0	1
C_24	1	0	1	0	0	0	0	0	0	0	0
C_30	1	0	1	1	0	0	0	0	0	0	1
C_31	1	0	1	0	0	0	0	0	0	1	0
C_32	1	0	1	0	0	0	0	0	0	0	0
C_33	1	0	1	0	0	0	0	0	0	0	0
C_34	1	0	1	0	1	0	0	0	0	0	1
C_35	1	0	1	0	1	0	0	0	0	0	1
C_36	1	0	1	0	0	0	0	0	0	0	0
C_37	1	0	1	0	1	0	0	0	0	0	0
C_38	1	1	1	0	0	0	0	1	0	0	1
C_39	1	0	1	0	0	0	0	0	0	0	1
C_40	1	0	1	0	0	0	0	0	0	0	0
C_41	1	1	1	0	0	0	0	0	0	0	1
C_42	1	0	1	0	0	0	0	1	0	0	1
C_43	1	0	1	0	1	1	0	1	0	0	1
C_44	1	0	1	0	1	1	0	0	0	0	1
C_45	1	0	1	0	0	0	0	0	0	1	1
C_46	1	0	1	0	1	0	0	0	0	0	0
C_47	1	0	1	1	0	0	0	0	0	0	1
C_48	1	0	1	0	1	1	0	0	0	0	1

Table 11 includes a summary of the number of items coded for each attribute by hierarchy across the three forms. For each of the forms examined, it was determined that all items assessed at least A1 and A3a. As evidenced in the table, few items were coded as measuring several of the attributes (e.g. A2, A3b, A4b, A6, A7, A8). This finding was observed across the three forms. The limited number of items coded to these attributes suggests that perhaps hierarchy one, which does not include A6 and A7, may better represent the data.

Table 11

Items Coded by Attribute

	A1	A2	A3a	A3b	A4a	A4b	A5	A6	A7	A8	A9
Form A	35	5	35	2	7	2	15	4	3	1	25
Form B	35	6	35	5	15	5	14	1	3	4	27
Form C	35	6	35	3	14	4	3	4	0	3	22

Stage Two

Expected response generation. Following the creation of the Q matrix for each form, expected response patterns were generated for each form. These patterns consisted of the possible responses that examinees might provide given their mastery of a unique combination of the attributes in the model. Thus, expected response patterns differed between hierarchy one and two due to the inclusion of two additional attributes in hierarchy two.

Table 12 includes a summary of the number of expected response patterns by form and hierarchy. These patterns represent the potential knowledge states into which

individuals could be classified. Due to having two fewer attributes, hierarchy one has substantially fewer possible knowledge states than hierarchy two across all three forms.

Table 12

Number of Knowledge States by Form

	Hierarchy 1	Hierarchy 2
Form A	40	141
Form B	41	124
Form C	57	115

Due to the large number of possible expected response patterns across the three forms, the expected response patterns are presented in Tables 23 – 28 along with the ability estimates associated with each response pattern, the results of which are as detailed in the text that follows.

Analysis of assumptions. Following the identification of the expected response patterns for each form, the process of analyzing the data began. First, in order to justify using a unidimensional IRT model, the data for each test form was evaluated for evidence of unidimensionality. Exploratory factor analysis was conducted in SPSS 20.0 for each test form. As part of the analysis, a scree plot of the eigenvalues was obtained for each form by group combination. These plots are displayed in Figures 7 – 12 that follow.

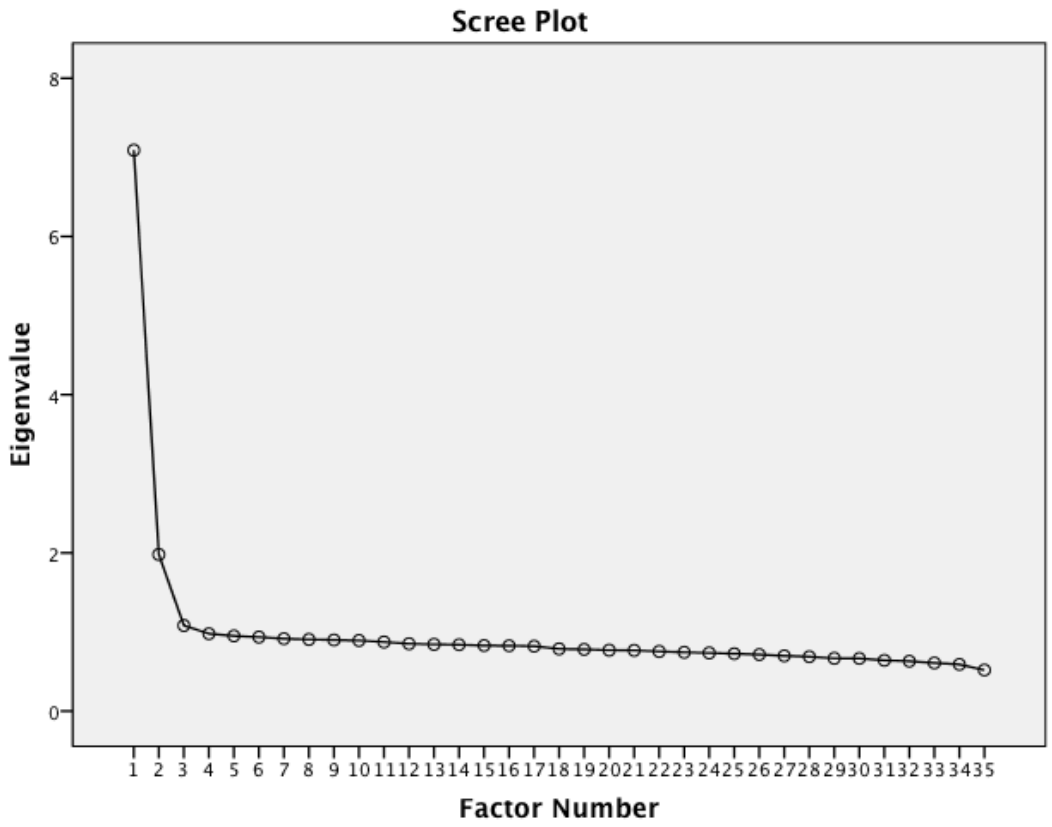


Figure 7. Scree plot of dimensions underlying Form A group 1.

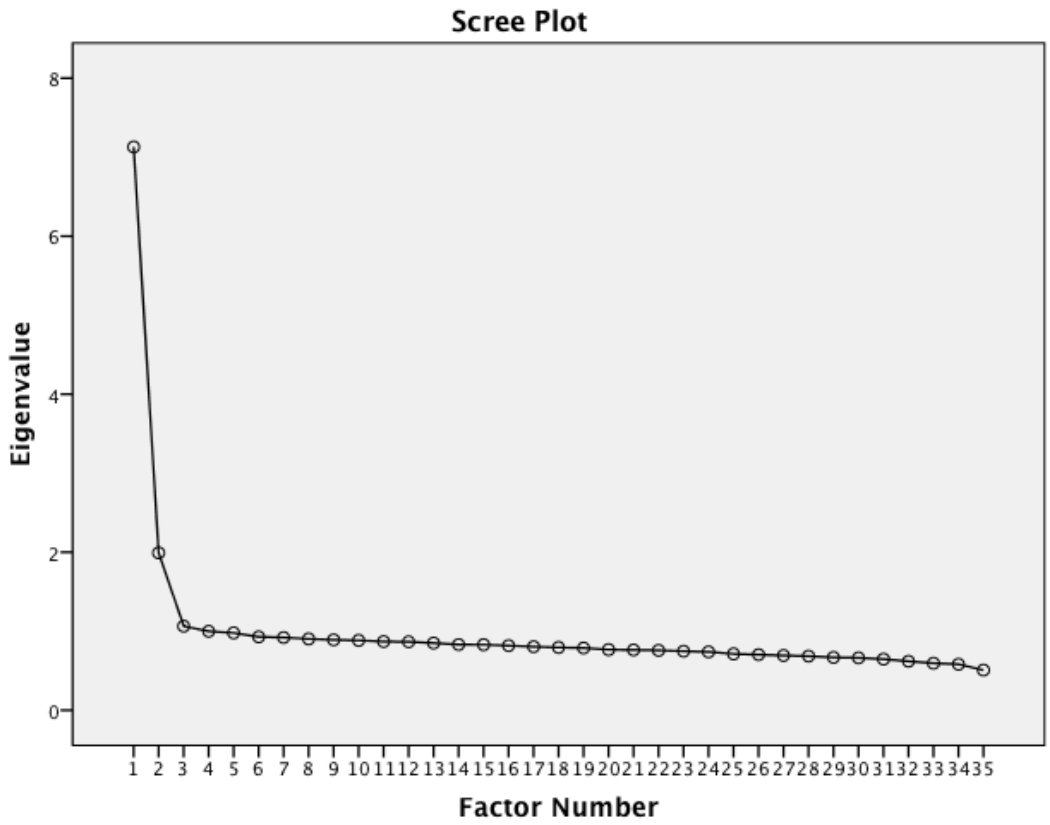


Figure 8. Scree plot of dimensions underlying Form A group 2.

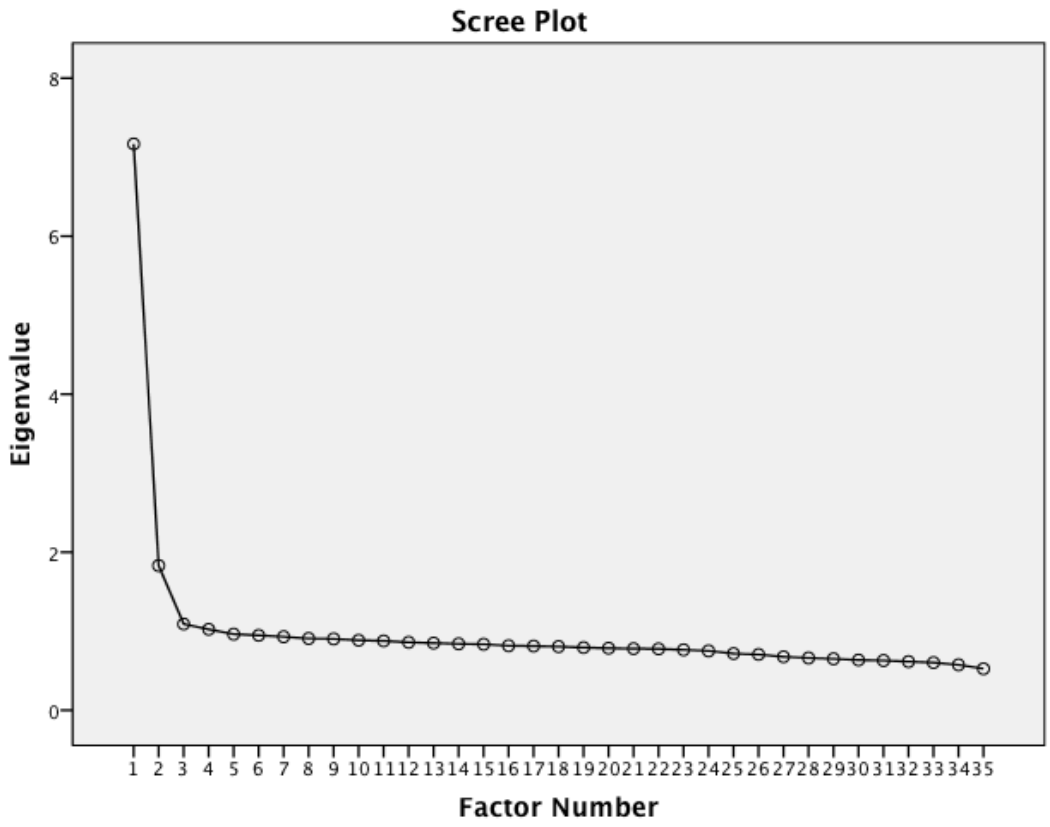


Figure 9. Scree plot of dimensions underlying Form B group 1.

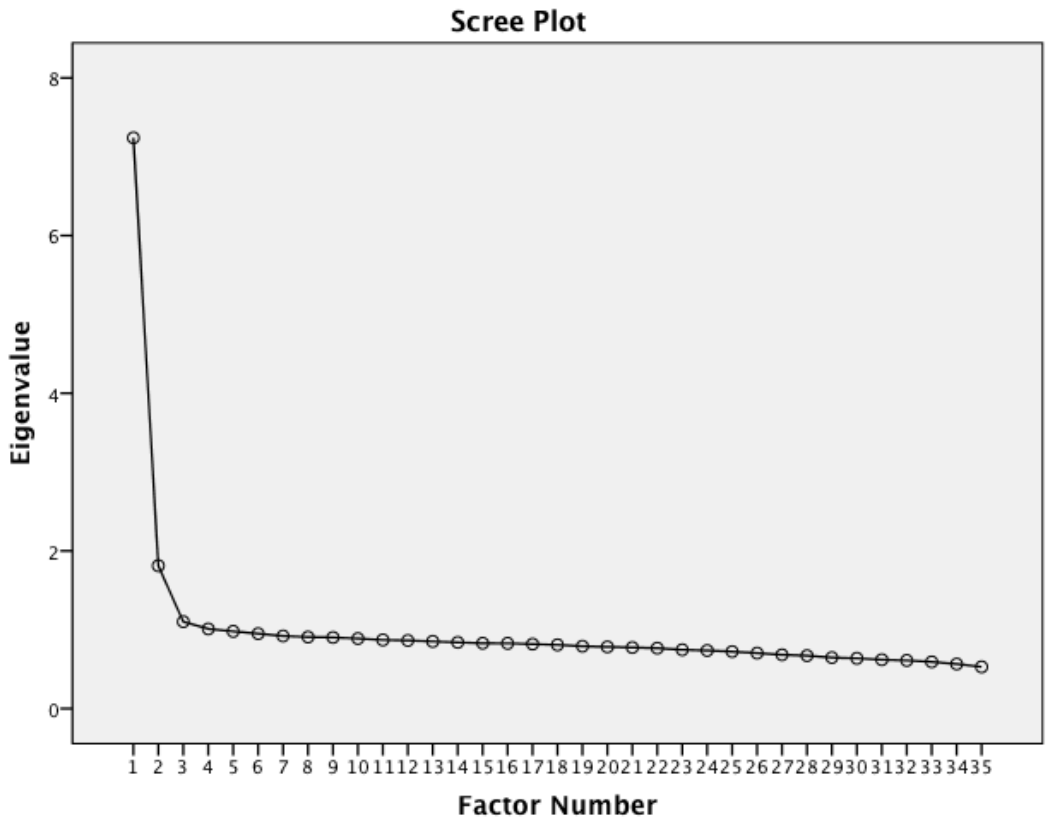


Figure 10. Scree plot of dimensions underlying Form B group 2.

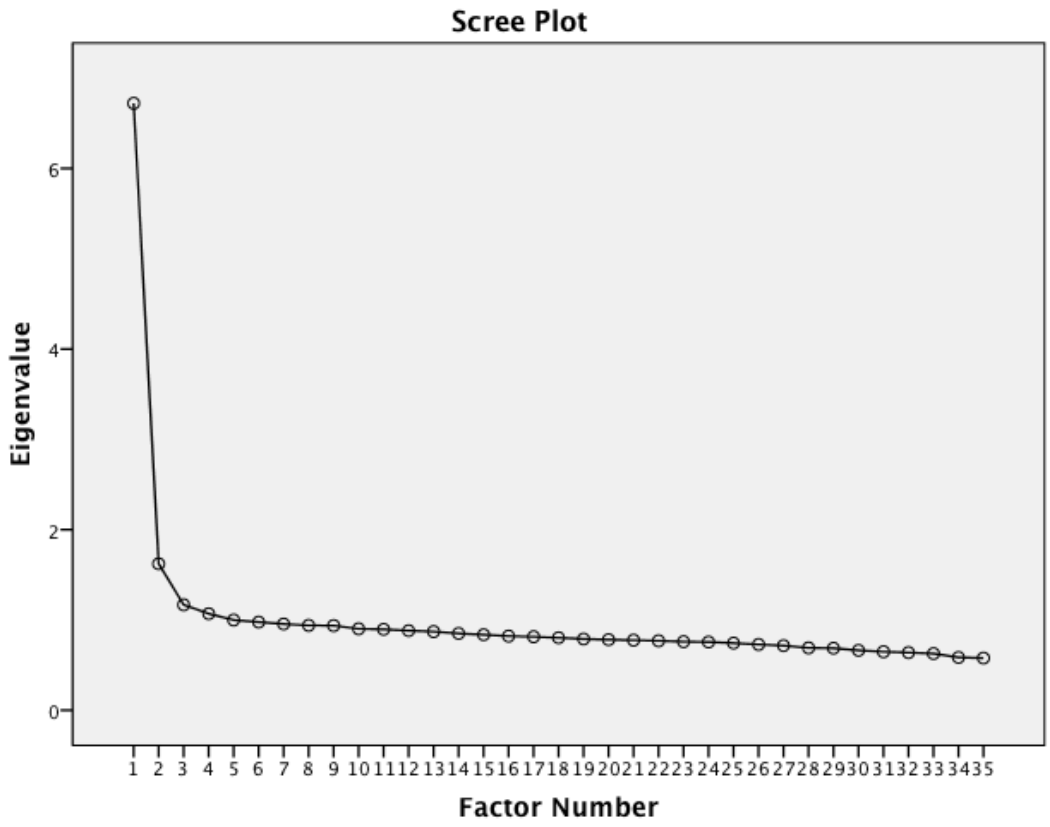


Figure 11. Scree plot of dimensions underlying Form C group 1.

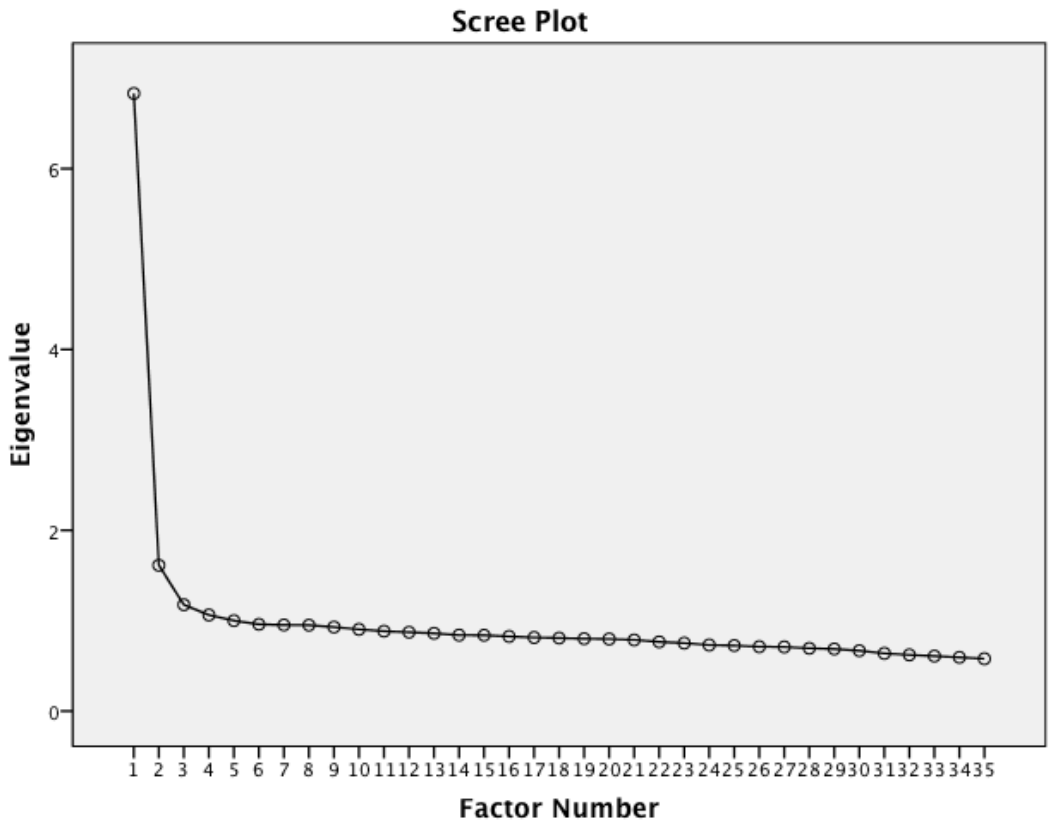


Figure 12. Scree plot of dimensions underlying Form C group 2.

Examination of the scree plots revealed that for all six form-by-group combinations, a single dominant factor was observed. This was evidenced by a single plot high on the y axis, followed by a drop in the eigenvalues associated with the remaining factors. This finding suggested that a single unidimensional model fit the data across all form-by-group combinations, and supported the use of a unidimensional item response theory model when applying the attribute hierarchy method.

In addition to evaluating the scree plots for each form-by-group combination, the ratio of adjacent eigenvalues was also calculated as further evidence of a unidimensional construct underlying the data. A similar finding was observed when examining the ratio of adjacent Eigenvalues, as presented in Table 13. The values in the first column represent the

ratio of the first eigenvalue to the second eigenvalue. The values in the second column represent the ratio of the second eigenvalue to the third eigenvalue, continuing on through the remaining columns. Inspection of the values in the table revealed that the values in the first column are must larger that those in the second, third, and fourth columns. This finding indicates that the first eigenvalue accounts for the largest amount of variance.

Table 13

Ratio of adjacent Eigenvalues

	1 to 2	2 to 3	3 to 4	4 to 5
Form A Group 1	3.58	1.83	1.11	1.03
Form A Group 2	3.58	1.87	1.07	1.02
Form B Group 1	3.92	1.68	1.07	1.06
Form B Group 2	4.00	1.64	1.09	1.03
Form C Group 1	4.15	1.39	1.09	1.07
Form C Group 2	4.24	1.37	1.11	1.06

Based on the findings observed in the scree plots and in the ratio of adjacent eigenvalues, the conclusion was made that the data are likely measuring a single underlying construct. Thus, support for use of a unidimensional item response theory model was obtained.

Comparison of item response theory models. Once support had been obtained for fitting a unidimensional item response theory model to the data, two models were compared for evidence of better fit to the data. Fit of the two- and three-parameter logistic models was compared across the test forms and groups. Fit of the two models to the data

are presented in Table 14 that follows. The -2 log likelihood, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) values for both models are presented. Inspection of the table revealed that the column including values obtained from the three-parameter logistic model (3PL) consistently included smaller values than those obtained for the two-parameter logistic model (2PL). This finding was observed across groups and test forms. As indicated by the smaller values across all three indices, the three-parameter logistic model was determined to provide the better fit to the data.

Table 14

Fit Comparison for Form A Item Response Theory Models

	Group 1		Group 2	
Statistic	2PL	3PL	2PL	3PL
-2 log likelihood	378559.0	375616.5	378096.9	375618.8
AIC	378699.0	375826.5	378236.9	375828.8
BIC	379203.7	376583.6	378741.7	376585.9

Table 15

Fit Comparison for Form B Item Response Theory Models

	Group 1		Group 2	
Statistic	2PL	3PL	2PL	3PL
-2 log likelihood	382746.6	380907.1	382966.2	381423.4
AIC	382886.6	381117.1	383106.2	381633.4
BIC	383391.3	381874.2	383610.9	382390.5

Table 16

Fit Comparison for Form C Item Response Theory Models

	Group 1		Group 2	
Statistic	2PL	3PL	2PL	3PL
-2 log likelihood	389770.5	388435.6	387240.5	385993.1
AIC	389910.5	388645.6	387380.5	386203.1
BIC	390415.2	389402.7	387885.2	386960.2

Because the three-parameter model was determined to provide better fit to the data in the current study, item parameter estimates for the three-parameter logistic model were saved for groups 1 and 2 for each form. The parameter estimate values obtained from the three parameter logistic model for groups 1 and 2 for each of the three forms are displayed in Tables 17 – 22. These values were then used for the subsequent process of fitting a diagnostic model using the attribute hierarchy method.

Table 17

Group 1 Form A Item Parameter Estimates

Item	*a*	s.e.	*b*	s.e.	*c*	s.e.
9	1.36	0.04	0.84	0.02	0.17	0.01
10	2.37	0.05	-0.52	0.02	0.28	0.01
11	2.25	0.09	1.88	0.02	0.16	0.00
12	3.23	0.12	2.03	0.02	0.09	0.00
13	1.38	0.04	0.35	0.02	0.15	0.01
14	1.80	0.04	-0.25	0.03	0.19	0.01
15	1.73	0.05	1.45	0.02	0.15	0.00
16	1.80	0.05	1.23	0.01	0.18	0.00
17	2.22	0.06	1.24	0.01	0.13	0.00
18	2.15	0.05	0.85	0.01	0.20	0.00
19	2.87	0.07	1.30	0.01	0.10	0.00
20	2.64	0.06	1.21	0.01	0.11	0.00
21	2.96	0.08	1.29	0.01	0.15	0.00
22	1.71	0.04	1.30	0.01	0.05	0.00
23	3.76	0.10	1.42	0.01	0.07	0.00
24	2.81	0.09	1.71	0.01	0.10	0.00
30	1.23	0.03	-0.26	0.04	0.04	0.02
31	0.95	0.01	0.32	0.01	0.00	0.00
32	2.12	0.04	0.08	0.02	0.18	0.01
33	1.59	0.05	0.83	0.02	0.23	0.01
34	1.50	0.02	-0.74	0.01	0.00	0.00
35	1.08	0.03	0.61	0.03	0.05	0.01
36	1.57	0.02	-0.98	0.01	0.00	0.00
37	2.95	0.09	1.56	0.01	0.12	0.00
38	1.69	0.02	-0.31	0.01	0.00	0.00
39	1.02	0.01	-0.31	0.01	0.00	0.00
40	0.97	0.04	0.71	0.04	0.10	0.01
41	1.86	0.02	-0.61	0.01	0.00	0.00
42	1.27	0.03	0.39	0.03	0.05	0.01
43	1.84	0.04	0.69	0.01	0.15	0.01
44	1.82	0.07	1.86	0.02	0.14	0.00
45	1.84	0.04	0.16	0.02	0.10	0.01
46	2.76	0.05	0.20	0.01	0.15	0.00
47	2.35	0.05	1.14	0.01	0.12	0.00
48	2.80	0.05	0.25	0.01	0.12	0.00

Table 18

Group 2 Form A Item Parameter Estimates

Item	*a*	s.e.	*b*	s.e.	*c*	s.e.
9	1.30	0.04	0.83	0.02	0.16	0.01
10	2.49	0.06	-0.51	0.02	0.28	0.01
11	2.27	0.08	1.85	0.02	0.16	0.00
12	3.34	0.13	2.03	0.02	0.09	0.00
13	1.35	0.03	0.37	0.02	0.16	0.01
14	1.77	0.04	-0.24	0.03	0.20	0.01
15	1.82	0.05	1.43	0.02	0.16	0.00
16	1.78	0.05	1.21	0.01	0.17	0.00
17	2.03	0.05	1.24	0.01	0.13	0.00
18	2.30	0.06	0.86	0.01	0.21	0.00
19	3.01	0.07	1.30	0.01	0.10	0.00
20	2.63	0.06	1.22	0.01	0.12	0.00
21	3.04	0.09	1.32	0.01	0.16	0.00
22	1.80	0.04	1.30	0.01	0.06	0.00
23	3.86	0.10	1.40	0.01	0.07	0.00
24	2.63	0.08	1.75	0.02	0.11	0.00
30	1.18	0.01	-0.36	0.01	0.00	0.00
31	0.94	0.01	0.35	0.01	0.00	0.00
32	2.14	0.04	0.06	0.02	0.17	0.01
33	1.70	0.05	0.86	0.02	0.23	0.01
34	1.53	0.02	-0.73	0.01	0.00	0.00
35	1.11	0.03	0.61	0.03	0.05	0.01
36	1.54	0.02	-1.00	0.01	0.00	0.00
37	2.83	0.08	1.57	0.01	0.11	0.00
38	1.70	0.02	-0.32	0.01	0.00	0.00
39	1.03	0.01	-0.31	0.01	0.00	0.00
40	0.98	0.03	0.68	0.04	0.09	0.01
41	1.90	0.02	-0.62	0.01	0.00	0.00
42	1.13	0.01	0.27	0.01	0.00	0.00
43	1.83	0.04	0.71	0.01	0.15	0.01
44	1.91	0.07	1.85	0.02	0.15	0.00
45	1.73	0.03	0.14	0.02	0.09	0.01
46	2.77	0.05	0.21	0.01	0.16	0.00
47	2.38	0.06	1.15	0.01	0.12	0.00
48	2.82	0.05	0.24	0.01	0.13	0.00

Table 19

Group 1 Form B Item Parameter Estimates

Item	*a*	s.e.	*b*	s.e.	*c*	s.e.
9	2.04	0.07	1.82	0.02	0.11	0.00
10	0.94	0.01	-0.47	0.01	0.00	0.00
11	1.28	0.04	0.19	0.04	0.20	0.01
12	1.03	0.04	0.43	0.04	0.13	0.02
13	1.84	0.06	1.39	0.02	0.22	0.00
14	1.20	0.04	1.43	0.02	0.15	0.01
15	2.15	0.07	1.42	0.01	0.16	0.00
16	2.03	0.04	0.09	0.01	0.08	0.01
17	1.59	0.04	1.01	0.01	0.12	0.01
18	2.37	0.05	0.30	0.01	0.20	0.01
19	3.43	0.09	1.36	0.01	0.09	0.00
20	-0.03	0.37	-65.14	25517503	0.06	88730
21	2.69	0.12	1.95	0.02	0.15	0.00
22	1.68	0.04	0.47	0.02	0.25	0.01
23	2.62	0.07	1.50	0.01	0.12	0.00
24	2.94	0.11	1.79	0.02	0.15	0.00
30	2.00	0.04	-1.07	0.04	0.04	0.03
31	1.38	0.03	0.29	0.02	0.10	0.01
32	1.64	0.04	-0.12	0.02	0.13	0.01
33	1.08	0.01	-0.29	0.01	0.00	0.00
34	1.17	0.04	-0.04	0.05	0.14	0.02
35	1.02	0.01	-0.49	0.01	0.00	0.00
36	1.71	0.04	-0.55	0.03	0.04	0.02
37	1.87	0.05	1.05	0.01	0.20	0.00
38	1.46	0.02	-0.57	0.01	0.00	0.00
39	1.27	0.04	1.66	0.02	0.07	0.00
40	1.62	0.02	-0.57	0.01	0.00	0.00
41	2.44	0.05	0.03	0.01	0.12	0.01
42	2.62	0.05	0.14	0.01	0.16	0.01
43	2.00	0.05	1.28	0.01	0.12	0.00
44	1.46	0.04	1.17	0.02	0.09	0.00
45	2.67	0.05	0.35	0.01	0.12	0.00
46	2.97	0.05	0.66	0.01	0.07	0.00
47	2.17	0.05	1.13	0.01	0.13	0.00
48	2.51	0.05	0.31	0.01	0.17	0.01

Table 20

Group 2 Form B Item Parameter Estimates

Item	*a*	s.e.	*b*	s.e.	*c*	s.e.
9	1.97	0.06	1.85	0.02	0.11	0.00
10	0.92	0.01	-0.48	0.01	0.00	0.00
11	1.17	0.04	0.06	0.04	0.17	0.02
12	1.00	0.03	0.42	0.04	0.12	0.02
13	1.88	0.06	1.39	0.02	0.22	0.00
14	1.20	0.04	1.38	0.02	0.14	0.01
15	2.31	0.07	1.43	0.02	0.16	0.00
16	1.94	0.04	0.10	0.01	0.06	0.01
17	1.55	0.04	1.01	0.02	0.11	0.01
18	2.24	0.05	0.29	0.01	0.20	0.01
19	3.55	0.10	1.35	0.01	0.09	0.00
20	0.01	0.00	381.63	783.00	0.06	0.46
21	2.68	0.12	1.96	0.02	0.15	0.00
22	1.55	0.04	0.42	0.02	0.22	0.01
23	2.69	0.08	1.52	0.01	0.13	0.00
24	3.03	0.11	1.79	0.02	0.15	0.00
30	1.99	0.04	-1.10	0.03	0.01	0.02
31	1.42	0.03	0.31	0.02	0.11	0.01
32	1.58	0.04	-0.15	0.02	0.11	0.01
33	1.06	0.01	-0.27	0.01	0.00	0.00
34	1.23	0.04	0.01	0.04	0.15	0.02
35	1.02	0.01	-0.47	0.01	0.00	0.00
36	1.71	0.04	-0.55	0.03	0.03	0.02
37	1.90	0.05	1.04	0.01	0.20	0.00
38	1.46	0.02	-0.56	0.01	0.00	0.00
39	1.21	0.04	1.67	0.02	0.06	0.00
40	1.68	0.02	-0.56	0.01	0.00	0.00
41	2.39	0.04	0.03	0.01	0.12	0.01
42	2.61	0.05	0.15	0.01	0.16	0.01
43	2.10	0.05	1.30	0.01	0.13	0.00
44	1.39	0.04	1.15	0.02	0.08	0.01
45	2.67	0.05	0.35	0.01	0.12	0.00
46	2.89	0.05	0.66	0.01	0.07	0.00
47	2.18	0.05	1.15	0.01	0.13	0.00
48	2.54	0.05	0.31	0.01	0.16	0.00

Table 21

Group 1 Form C Item Parameter Estimates

Item	*a*	s.e.	*b*	s.e.	*c*	s.e.
9	2.46	0.09	1.97	0.02	0.13	0.00
10	0.88	0.05	1.06	0.06	0.32	0.02
11	1.27	0.05	1.59	0.02	0.14	0.01
12	1.84	0.06	1.70	0.02	0.08	0.00
13	2.40	0.06	0.70	0.01	0.20	0.00
14	1.75	0.04	-0.69	0.03	0.14	0.02
15	3.10	0.15	1.85	0.02	0.19	0.00
16	2.43	0.07	1.11	0.01	0.19	0.00
17	1.41	0.04	0.54	0.02	0.10	0.01
18	1.21	0.04	1.10	0.02	0.12	0.01
19	1.81	0.04	-0.50	0.03	0.11	0.02
20	2.68	0.06	0.59	0.01	0.14	0.00
21	2.28	0.05	0.94	0.01	0.06	0.00
22	1.27	0.06	2.09	0.04	0.17	0.01
23	1.82	0.04	0.66	0.01	0.12	0.01
24	1.60	0.05	1.25	0.02	0.16	0.01
30	1.71	0.04	-1.19	0.05	0.10	0.03
31	0.81	0.01	-0.49	0.01	0.00	0.00
32	0.90	0.01	-0.53	0.01	0.00	0.00
33	1.37	0.03	-0.49	0.04	0.04	0.02
34	1.33	0.09	2.54	0.06	0.17	0.00
35	1.74	0.04	-0.56	0.03	0.02	0.02
36	1.28	0.02	-0.39	0.01	0.00	0.00
37	1.11	0.01	-0.17	0.01	0.00	0.00
38	1.95	0.04	0.84	0.01	0.12	0.00
39	1.16	0.01	0.41	0.01	0.00	0.00
40	1.11	0.01	-0.10	0.01	0.00	0.00
41	1.05	0.01	-0.72	0.01	0.00	0.00
42	0.87	0.01	0.32	0.01	0.00	0.00
43	1.68	0.05	1.10	0.01	0.12	0.00
44	2.54	0.05	0.50	0.01	0.07	0.00
45	0.13	0.26	15.40	27.53	0.06	0.11
46	1.51	0.03	0.37	0.02	0.03	0.01
47	3.02	0.07	0.74	0.01	0.14	0.00
48	1.28	0.04	1.58	0.02	0.09	0.01

Table 22

Group 2 Form C Item Parameter Estimates

Item	*a*	s.e.	*b*	s.e.	*c*	s.e.
9	2.24	0.08	1.93	0.02	0.12	0.00
10	0.90	0.05	0.99	0.06	0.33	0.01
11	1.25	0.05	1.50	0.02	0.13	0.01
12	1.74	0.05	1.67	0.02	0.08	0.00
13	2.32	0.05	0.60	0.01	0.19	0.00
14	1.80	0.04	-0.72	0.03	0.18	0.02
15	3.21	0.15	1.76	0.02	0.19	0.00
16	2.32	0.07	1.01	0.01	0.18	0.00
17	1.41	0.03	0.44	0.02	0.10	0.01
18	1.15	0.04	1.03	0.02	0.11	0.01
19	1.87	0.04	-0.57	0.03	0.13	0.01
20	2.61	0.06	0.49	0.01	0.14	0.00
21	2.35	0.05	0.86	0.01	0.06	0.00
22	1.33	0.07	2.01	0.04	0.17	0.00
23	1.72	0.04	0.58	0.01	0.12	0.01
24	1.49	0.05	1.19	0.02	0.15	0.01
30	1.76	0.04	-1.28	0.04	0.11	0.03
31	0.83	0.01	-0.58	0.01	0.00	0.00
32	0.94	0.01	-0.62	0.01	0.00	0.00
33	1.36	0.03	-0.62	0.03	0.02	0.02
34	1.15	0.07	2.45	0.06	0.16	0.01
35	1.72	0.02	-0.68	0.01	0.00	0.00
36	1.35	0.02	-0.47	0.01	0.00	0.00
37	1.14	0.14	-0.67	0.09	0.00	0.00
38	1.91	0.04	0.77	0.01	0.12	0.00
39	1.11	0.01	0.35	0.01	0.00	0.00
40	1.11	0.01	-0.19	0.01	0.00	0.00
41	1.08	0.01	-0.79	0.01	0.00	0.00
42	0.86	0.01	0.19	0.01	0.00	0.00
43	1.62	0.05	1.02	0.01	0.12	0.00
44	2.53	0.05	0.43	0.01	0.07	0.00
45	0.12	0.29	16.55	39.71	0.05	0.11
46	1.51	0.03	0.30	0.02	0.04	0.01
47	2.93	0.06	0.66	0.01	0.14	0.00
48	1.31	0.04	1.49	0.02	0.09	0.00

Using the item parameters obtained from the three-parameter logistic item response theory model, ability estimates were calculated for each knowledge state for each of the three forms. These values are displayed for both groups in Tables 23 – 28 that follow.

Table 23

Ability Estimates for Form A Hierarchy One

Attributes Mastered	Expected Response Pattern	Ability Estimate	
		Group 1	Group 2
A0	00000000000000000000000000000000	-2.00	-2.02
A13a	10000100000000000010000000100000000	-1.51	-1.53
A13a9	10000100000000101110011010100000011	-0.22	-0.25
A13a5	11100100000000000010000000101100000	-0.84	-0.87
A123a	10000100000010000010000000100010000	-1.18	-1.21
A13a59	11111101000000101110011010101100111	0.74	0.71
A13a4a9	10000100000001111110111010100000011	0.17	0.14
A13a3b9	10000100000000101111011110100000011	0.05	0.02
A123a9	10000100000010101110011010100010011	0.05	0.02
A13a589	11111101000000101110011011101100111	0.84	0.81
A13a4a59	11111111000001111110111010101101111	1.25	1.23
A13a4a4b9	10000100000001111110111010110000011	0.29	0.26
A13a3b59	11111101000000101111011110101100111	0.94	0.92
A13a3b4a9	10000100000001111111111110100000011	0.41	0.38
A123a59	11111101011110101110011010101110111	1.25	1.23
A123a4a9	10000100000011111110111010100010011	0.41	0.38
A123a3b9	10000100000010101111011110100010011	0.29	0.26
A123a3b4a9	10000100000011111111111110100010011	0.63	0.60
A123a3b59	11111101011110101111011110101110111	1.46	1.45
A123a4a4b9	10000100000011111110111010110010011	0.52	0.49
A123a4a59	11111111011111111110111010101111111	1.80	1.82
A123a589	11111101011110101110011011101110111	1.35	1.34
A13a3b4a4b9	10000100000001111111111110110000011	0.52	0.49
A13a3b4a59	11111111000001111111111110101101111	1.46	1.45
A13a3b589	11111101000000101111011111101100111	1.05	1.02
A13a4a589	11111111000000101110011011101101111	1.05	1.02
A13a4a4b59	11111111100001111110111010111101111	1.46	1.45
A123a3b4a4b9	10000100000011111111111110110010011	0.74	0.71
A123a3b4a59	11111111011111111111111110101111111	2.08	2.15
A123a3b589	11111101011110101111011111101110111	1.57	1.57
A123a4a4b59	11111111111111111110111010111111111	2.08	2.15
A123a4a589	11111111011111111110111011101111111	1.93	1.97
A13a3b4a4b59	11111111100001111111111110111101111	1.68	1.69
A13a3b4a589	11111111000001111111111111101101111	1.57	1.57
A13a4a4b589	11111111000001111110111011111101111	1.46	1.45
A13a3b4a4b589	11111111100001111111111111111101111	1.80	1.82
A123a4a4b589	11111111111111111110111011111111111	2.26	2.37
A123a3b4a589	11111111011111111111111111101111111	2.26	2.37
A123a3b4a4b59	11111111111111111111111110111111111	2.49	2.64
A123a3b4a4b589	11111111111111111111111111111111111	2.78	2.73

Table 24

Ability Estimates for Form B Hierarchy One

Attributes Mastered	Expected Response Pattern	Ability Estimate	
		Group 1	Group 2
A0	00000000000000000000000000000000000	-2.10	-2.63
A13a	00100000000000001110000000000000001	-1.42	-2.15
A123a	00100000000001001110000000000000001	-1.25	-2.04
A13a4a	00100000000000001110000110000000001	-1.08	-1.91
A13a9	00100001000000001110000000000010001	-1.08	-1.91
A123a4a	00100000000001001110000110000000001	-0.91	-1.79
A123a9	00100001000001001110000000000011001	-0.75	-1.65
A13a3b9	00100001001000001110010000000110101	-0.45	-1.37
A13a4a9	00100001000000001111100110100010001	-0.31	-1.23
A13a59	00100001000010001110000000001010011	-0.60	-1.51
A13a589	01110001000010001110000000001010011	-0.31	-1.23
A13a4a59	00101011000010001111100111101010011	0.41	-0.41
A13a4a4b9	00100001000000001111110011010010001	-0.18	-1.09
A13a3b59	00100101001010001110010000001110111	0.07	-0.81
A123a59	00100001000111001110000000001011011	-0.18	-1.09
A123a4a9	00100001010001001111101110100011001	0.19	-0.68
A123a3b9	00100001001001001110010000000111101	-0.18	-1.09
A13a3b4a9	00100001001000001111110110100110101	0.19	-0.68
A123a3b4a9	00100001011001001111111110100111101	0.63	-0.15
A123a3b59	00100101001111001110010000001111111	0.41	-0.41
A123a4a4b9	00100001010001011111101110100011001	0.30	-0.54
A123a4a59	00101011010111001111101111101011011	0.96	0.24
A123a589	01110001000111001110000000001011011	0.07	-0.81
A13a3b4a4b9	00100001001000001111111110100110101	0.30	-0.54
A13a3b4a59	00101111001010001111110111101110111	0.96	0.24
A13a3b589	01110101001010001110010000001110111	0.30	-0.54
A13a4a4b59	00101011000010111111100111111010011	0.74	-0.02
A13a4a589	01111011000010001111100111101010011	0.63	-0.15
A123a3b4a4b9	00100001011001011111111110100111101	0.74	-0.02
A123a3b4a59	00101111011111001111111111101111111	1.55	0.94
A123a3b589	01110101001111001110010000001111111	0.63	-0.15
A123a4a4b59	00101011010111111111101111111011011	1.30	0.64
A123a4a589	01111011010111001111101111101011011	1.18	0.50
A13a3b4a4b59	00101111001010111111110111111110111	1.30	0.64
A13a3b4a589	01111111001010001111110111101110111	1.18	0.50
A13a4a4b589	01111011100010111111100111111010011	1.07	0.37
A13a3b4a4b589	01111111101010111111110111111110111	1.69	1.10
A123a4a4b589	11111011110111111111101111111011011	1.83	1.27
A123a3b4a589	01111111011111001111111111101111111	1.83	1.27
A123a3b4a4b59	00101111011111111111111111111111111	2.00	1.45

A123a3b4a4b589	11111111111111111111111111111111111	2.76	2.45

Table 25

Ability Estimates for Form C Hierarchy One

Attributes Mastered	Expected Response Pattern	Ability Estimate	
		Group 1	Group 2
A0	00000000000000000000000000000000000	-2.21	-2.17
A13a	00000000101000010011001000100000000	-1.12	-1.10
A13a88	00000000101000010111001000100000000	-0.94	-0.92
A13a9	00000000101000010011001001101000000	-0.78	-0.76
A13a5	10000000101000010011001000100000000	-0.94	-0.92
A13a4a	00001000101000010011001100100000100	-0.62	-0.60
A13a89	00000000101000010111001001101001000	-0.47	-0.45
A13a58	10000000101000010111001000100000000	-0.78	-0.76
A13a59	10000000101000010011001001101000000	-0.62	-0.60
A13a4a8	00001000111000010111001100100000100	-0.33	-0.31
A13a4a9	00001100101010010011111101101000100	0.19	0.21
A13a4a5	10001000101000010011001100100000100	-0.47	-0.45
A13a3b9	00000000101100011011001001101000010	-0.33	-0.31
A123a9	00000011101001110011001011111000000	0.07	0.08
A123a3b9	00000011101101111011001011111000010	0.43	0.44
A123a4a9	00001111101011110011111111111000100	0.89	0.89
A123a59	10000011101001110011001011111000000	0.19	0.21
A123a89	00000011101001110111001011111001000	0.31	0.32
A13a3b4a9	00001100101110011011111101101000110	0.56	0.55
A13a3b59	10000000101100011011001001101000010	-0.19	-0.17
A13a3b89	00000000101100011111001001101001010	-0.06	-0.04
A13a4a4b9	00011100101010010011111101101110101	0.66	0.67
A13a4a58	10001000111000010111001100100000100	-0.19	-0.17
A13a4a59	11101100101010010011111101101000100	0.55	0.55
A13a4a89	00001100111010010111111101101001100	0.55	0.55
A13a589	10000000101000010111001001101001000	-0.33	-0.31
A123a3b4a9	00001111101111111011111111111000110	1.23	1.23
A123a3b59	10000011101101111011001011111000010	0.55	0.55
A123a3b89	00000011101101111111001011111001010	0.66	0.67
A123a4a4b9	00011111101011110011111111111110101	1.35	1.35
A123a4a59	11101111101011110011111111111000100	1.23	1.23
A123a4a89	00001111111011110111111111111001100	1.23	1.23
A123a589	10000011101001110111001011111001000	0.43	0.44
A13a3b4a4b9	00011100101110011011111101101110111	1.00	1.00
A13a3b4a59	11101100101110011011111101101000110	0.89	0.89
A13a3b4a89	00001100111110011111111101101001110	0.89	0.89
A13a3b589	10000000101100011111001001101001010	0.07	0.08

A13a4a4b59	11111100101010010011111101101110101	1.00	1.00
A13a4a4b89	01111100111010010111111101101111101	1.23	1.23
A13a4a589	11101100111010010111111101101001100	0.89	0.89
A123a3b4a4b9	00011111101111111011111111111110111	1.76	1.76
A123a3b4a59	11101111101111110111111111111000110	1.62	1.61
A123a3b4a89	00001111111111111111111111111001110	1.62	1.61
A123a3b589	10000011101101111111001011111001010	0.77	0.78
A123a4a4b59	11111111101011110011111111111110101	1.76	1.76
A123a4a4b89	00011111111011110111111111111111101	1.76	1.76
A123a4a589	11101111111011110111111111111001100	1.62	1.61
A13a3b4a4b59	11111100101110011011111101101110111	1.35	1.35
A13a3b4a4b89	00011100111110011111111101101111111	1.35	1.35
A13a3b4a589	11101100111110011111111101101001110	1.23	1.23
A13a4a4b589	11111100111010010111111101101111101	1.35	1.35
A13a3b4a4b589	11111100111110011111111101101111111	1.76	1.76
A123a4a4b589	11111111111011110111111111111111101	2.30	2.29
A123a3b4a589	11101111111111111111111111111001110	2.10	2.09
A123a3b4a4b89	00011111111111111111111111111111111	2.30	2.29
A123a3b4a4b59	11111111101111110111111111111110111	2.30	2.29
A123a3b4a4b589	11111111111111111111111111111111111	2.88	2.89

Table 26

Ability Estimates for Form A Hierarchy Two

Attributes Mastered	Expected Response Pattern	Ability Estimates	
		Group 1	Group 2
A0	00000000000000000000000000000000000	-2.00	-2.02
A13a	10000100000000000010000000100000000	-1.51	-1.53
A13a9	10000100000000001010011010100000001	-0.67	-0.70
A13a5	11100100000000000010000000101100000	-0.84	-0.87
A123a	10000100000010000010000000100010000	-1.18	-1.21
A13a79	10000100000000001110011010100000011	-0.36	-0.39
A13a69	10000100000000101010011010100000001	-0.51	-0.54
A13a59	11111100000000001010011010101100101	0.29	0.26
A13a4a9	10000100000001011010111010100000001	-0.22	-0.25
A13a3b9	10000100000000001011011010100000001	-0.51	-0.54
A123a9	10000100000010001010011010100010001	-0.36	-0.39
A123a5	11100100000010000010000000101110000	-0.51	-0.54
A13a589	11111100000000001010011011101100101	0.41	0.38
A13a679	10000100000000101110011010100000011	-0.22	-0.25
A13a569	11111101000000101010011010101100101	0.52	0.49
A13a579	11100100000000001110011010101100011	0.17	0.14
A13a4a79	10000100000001011110111010100000011	0.05	0.02
A13a4a59	11111100000001011010111010101100101	0.63	0.60
A13a4a69	10000100000001111010111010100000001	-0.08	-0.11
A13a3b69	10000100000000101011011010100000001	-0.36	-0.39
A13a3b79	10000100000000001111011010100000011	-0.22	-0.25
A13a3b59	11111100000000001011011010101100101	0.41	0.38
A13a3b4a9	10000100000001011011111010100000001	-0.08	-0.11
A123a79	10000100000010001110011010100010011	-0.08	-0.11
A123a69	10000100000010101010011010100010001	-0.22	-0.25
A123a4a9	10000100000011011010111010100010001	0.05	0.02
A123a59	11111100000010001010011010101110101	0.52	0.49
A123a3b9	10000100000010001011011010100010001	-0.22	-0.25
A13a5789	11111100000000001110011011101100111	0.63	0.60
A13a5689	11111101000000101010011011101100101	0.63	0.60
A13a5679	11111101000000101110011010101100111	0.74	0.71
A13a4a679	10000100000001111110111010100000011	0.17	0.14
A13a4a589	11111110000001011010111011101101101	0.94	0.92
A13a4a579	11111110000001011110111010101101111	1.05	1.02
A13a4a569	11111111000001111010111010101101101	1.05	1.02
A13a4a4b59	11111100100001011010111010101100101	0.74	0.71
A13a4a4b69	10000100000001111010111010110000001	0.05	0.02
A13a3b679	10000100000000101111011110100000011	0.05	0.02
A13a3b589	11111100000000001011011011101100101	0.52	0.49
A13a3b579	11111100000000001111011110101100111	0.74	0.71

A13a3b569	11111100000000101011011010101100101	0.52	0.49
A13a3b4a79	10000100000001011111111110100000011	0.29	0.26
A13a3b4a69	10000100000001111011111010100000001	0.05	0.02
A13a3b4a59	11111110000001011011111010101101101	0.94	0.92
A123a679	10000100000010101110011010100010011	0.05	0.02
A123a4a79	10000100000011011110111010100010011	0.29	0.26
A123a589	11111100011010001010011011101110101	0.84	0.81
A123a579	11111100011010001110011010101110111	0.94	0.92
A123a569	11111101011110101010011010101110101	1.05	1.02
A123a4a69	10000100000011111010111010100010001	0.17	0.14
A123a4a59	11111110011011011010111010101111101	1.25	1.23
A123a3b79	10000100000010001111011110100010011	0.17	0.14
A123a3b69	10000100000010101011011010100010001	-0.08	-0.11
A123a3b59	11111100011010001011011010101110101	0.84	0.81
A123a3b4a9	10000100000011011011111010100010001	0.17	0.14
A13a56789	11111101000000101110011011101100111	0.84	0.81
A13a4a5789	11111110000001011110111011101101111	1.15	1.13
A13a4a5689	11111111000001111010111011101101101	1.15	1.13
A13a4a5679	11111111000001111110111010101101111	1.25	1.23
A13a4a4b679	10000100000001111110111010110000011	0.29	0.26
A13a4a4b589	11111110100001011010111011101101101	1.05	1.02
A13a4a4b579	11111110100001011110111010101101111	1.15	1.13
A13a4a4b569	11111111000001111010111010111101101	1.25	1.23
A13a3b5789	11111100000000001111011111101100111	0.84	0.81
A13a3b5689	11111101000000101011011011101100101	0.74	0.71
A13a3b5679	11111101000000101111011110101100111	0.94	0.92
A13a3b4a679	10000100000001111111111110100000011	0.41	0.38
A13a3b4a589	11111110000001011011111011101101101	1.05	1.02
A13a3b4a579	11111110000001011111111110101101111	1.25	1.23
A13a3b4a569	11111111000001111011111010101101101	1.15	1.13
A13a3b4a4b69	10000100000001111011111010110000001	0.17	0.14
A13a3b4a4b59	11111110100001011011111010101101101	1.05	1.02
A123a5789	11111100011010001110011011101110111	1.05	1.02
A123a5689	11111101011110101010011011101110101	1.15	1.13
A123a5679	11111101011110101110011010101110111	1.25	1.23
A123a4a679	10000100000011111110111010100010011	0.41	0.38
A123a4a589	11111110011011011010111011101111101	1.35	1.34
A123a4a579	11111110011011011110111010101111111	1.48	1.45
A123a4a569	11111111011111111010111010101111101	1.57	1.57
A123a4a4b69	10000100000011111010111010110010001	0.29	0.26
A123a4a4b59	11111110111011011010111010101111101	1.35	1.34
A123a3b679	10000100000010101111011110100010011	0.29	0.26
A123a3b589	11111100011010001011011011101110101	0.94	0.92
A123a3b579	11111100011010001111011110101110111	1.15	1.13
A123a3b569	11111101011110101011011010101110101	1.15	1.13
A123a3b4a79	10000100000011011111111110100010011	0.52	0.49

A123a3b4a69	10000100000011111011111010100010001	0.29	0.26
A123a3b4a59	11111110011011011011111010101111101	1.35	1.34
A13a4a56789	11111111000001111110111011101101111	1.35	1.34
A13a4a4b5789	11111110100001011110111011101101111	1.25	1.23
A13a4a4b5689	11111111100001111010111011111101101	1.35	1.34
A13a4a4b5679	11111111100001111110111010111101111	1.46	1.45
A13a3b56789	11111101000000101111011111101100111	1.05	1.02
A13a3b4a5789	11111110000001011111111111101101111	1.35	1.34
A13a3b4a5689	11111111000001111011111011101101101	1.25	1.23
A13a3b4a5679	11111111000001111111111110101101111	1.46	1.45
A13a3b4a4b679	10000100000001111111111110110000011	0.52	0.49
A13a3b4a4b589	11111110100001011011111011101101101	1.15	1.13
A13a3b4a4b579	11111110100001011111111110101101111	1.35	1.34
A13a3b4a4b569	11111111100001111011111010111101101	1.35	1.34
A123a56789	11111101011110101110011011101110111	1.35	1.34
A123a4a4b679	10000100000011111110111010110010011	0.52	0.49
A123a4a4b579	11111110111011011110111010101111111	1.57	1.57
A123a4a5789	11111110011011011110111011101111111	1.57	1.57
A123a4a5689	11111111011111111010111011101111101	1.68	1.69
A123a4a5679	11111111011111111110111010101111111	1.80	1.82
A123a4a4b589	11111110111011011010111011101111101	1.46	1.45
A123a4a4b569	11111111111111111010111010111111101	1.80	1.82
A123a3b5789	11111100011010001111011111101110111	1.25	1.23
A123a3b5689	11111101011010101011011011101110101	1.15	1.13
A123a3b5679	11111101011110101111011110101110111	1.46	1.45
A123a3b4a679	10000100000011111111111110100010011	0.63	0.60
A123a3b4a589	11111110011011011011111011101111101	1.46	1.45
A123a3b4a579	11111110011011011111111110101111111	1.68	1.69
A123a3b4a569	11111111011111111011111010101111101	1.68	1.69
A123a3b4a4b69	10000100000011111011111010110010001	0.41	0.38
A123a3b4a4b59	11111110111011011011111010101111101	1.46	1.45
A13a4a4b56789	11111111000001111110111011111101111	1.46	1.45
A13a3b4a56789	11111111000001111111111111101101111	1.57	1.57
A13a3b4a4b5789	11111110100001011111111111101101111	1.46	1.45
A13a3b4a4b5689	11111111100001111011111011111101101	1.46	1.45
A13a3b4a4b5679	11111111100001111111111110111101111	1.68	1.69
A123a4a4b5789	11111110111011011110111011101111111	1.68	1.69
A123a4a4b5679	11111111111111111110111010111111111	2.08	2.15
A123a4a56789	11111111011111111110111011101111111	1.93	1.97
A123a4a4b5689	11111111111111111010111011111111101	1.93	1.97
A123a3b56789	11111101011110101111011111101110111	1.57	1.57
A123a3b4a5789	11111110011011011111111111101111111	1.80	1.82
A123a3b4a5689	11111111011111111011111011101111101	1.80	1.82
A123a3b4a5679	11111111011111111111111110101111111	2.08	2.15
A123a3b4a4b679	10000100000011111111111110110010011	0.74	0.71
A123a3b4a4b589	11111110111011011011111011101111101	1.57	1.57

A123a3b4a4b579	11111110111011011111111110101111111	1.80	1.82
A123a3b4a4b569	11111111111111111011111010111111101	1.93	1.97
A13a3b4a4b56789	11111111100001111111111111111101111	1.80	1.82
A123a4a4b56789	11111111111111111110111011111111111	2.26	2.37
A123a3b4a56789	11111111011111111111111111101111111	2.26	2.37
A123a3b4a4b5789	11111110111011011111111111101111111	1.93	1.97
A123a3b4a4b5689	11111111111111111011111011111111101	2.08	2.15
A123a3b4a4b5679	11111111111111111111111110111111111	2.49	2.64
A123a3b4a4b56789	11111111111111111111111111111111111	2.78	2.73

Table 27

Ability Estimates for Form B Hierarchy Two

Attributes Mastered	Expected Response Pattern	Ability Estimate	
		Group 1	Group 2
A0	00000000000000000000000000000000000	-2.10	-2.63
A13a	00100000000000001110000000000000001	-1.42	-2.15
A13a9	00100001000000001110000000000000001	-1.25	-2.04
A13a4a	00100000000000001110000110000000001	-1.08	-1.91
A123a	00100000000001001110000000000000001	-1.25	-2.04
A13a4a9	00100001000000001110000110000000001	-0.91	-1.79
A123a9	00100001000001001110000000000001001	-0.91	-1.79
A123a4a	00100000000001001110000110000000001	-0.91	-1.79
A13a69	00100001000000001110000000000010001	-1.08	-1.91
A13a59	00100001000010001110000000001000011	-0.75	-1.65
A13a3b9	00100001001000001110010000000100101	-0.60	-1.51
A123a3b9	00100001001001001110010000000101101	-0.31	-1.23
A123a4a9	00100001010001001110101110100001001	-0.06	-0.95
A123a59	00100001000111001110000000001001011	-0.31	-1.23
A123a69	00100001000001001110000000000011001	-0.75	-1.65
A13a3b4a9	00100001001000001110110110100100101	-0.06	-0.95
A13a3b59	00100101001010001110010000001100111	-0.06	-0.95
A13a3b69	00100001001000001110010000000110101	-0.45	-1.37
A13a4a59	00100001000010001110100110101000011	-0.18	-1.09
A13a4a69	00100001000000001110100110100010001	-0.45	-1.37
A13a569	00100001000010001110000000001010011	-0.60	-1.51
A13a4a79	00100001000000001111100110100000001	-0.45	-1.37
A13a4a4b9	00100001000000001110100110100000001	-0.60	-1.51
A13a589	01110001000010001110000000001000011	-0.45	-1.37
A123a3b4a9	00100001011001001110111110100101101	0.41	-0.41
A123a3b59	00100101001111001110010000001101111	0.30	-0.54
A123a3b69	00100001001001001110010000000111101	-0.18	-1.09
A123a4a4b9	00100001010001011110101110100001001	0.07	-0.81
A123a4a59	00100001010111001110101111101001011	0.52	-0.28
A123a4a69	00100001010001001110101110100011001	0.07	-0.81
A123a4a79	00100001010001001111101110100001001	0.07	-0.81
A123a569	00100001000111001110000000001011011	-0.18	-1.09
A13a3b4a4b9	00100001001000011110110110100100101	0.07	-0.81
A13a3b4a59	00100101001010001110110111101100111	0.52	-0.28
A13a3b4a69	00100001001000001110110110100110101	0.07	-0.81
A13a3b4a79	00100001001000001111111110100100101	0.07	-0.81
A13a3b569	00100101000010001110000000001010011	-0.45	-1.37

A13a4a4b59	00100001000010111110100111111000011	0.30	-0.54
A13a4a4b69	00100001000000011110100110100010001	-0.31	-1.23
A13a4a4b79	00100001000000011111100110100000001	-0.31	-1.23
A13a4a569	00100001000010001110100111101010011	0.07	-0.81
A13a4a579	00100001000010001111100111101000011	0.07	-0.81
A13a4a679	00101011000000000111110011010010001	-0.06	-0.95
A123a589	01110001000111001110000000001001011	-0.06	-0.95
A13a3b589	01110101001010001110010000001100111	0.19	-0.68
A13a4a589	01110001000010001110100111101000011	0.19	-0.68
A13a5689	01110001000010001110000000001010011	-0.31	-1.23
A123a3b4a4b9	00100001011001011110111110100101101	0.52	-0.28
A123a3b4a59	00100101011111001110111111101101111	1.07	0.37
A123a3b4a69	00100001011001001110111110100111101	0.52	-0.28
A123a3b4a79	00100001011001001111111110100101101	0.52	-0.28
A123a3b569	00100101001111001110010000001111111	0.41	-0.41
A123a3b589	01110101001111001110010000001101111	0.52	-0.28
A123a4a4b59	00100001010111111110101111111001011	0.85	0.11
A123a4a4b69	00100001010001011110101110100011001	0.19	-0.68
A123a4a4b79	00100001010001011111101110100001001	0.19	-0.68
A123a4a569	00100001010111001110101111101011011	0.63	-0.15
A123a4a579	00101011010111001111101111101001011	0.85	0.11
A123a4a589	01110001010111001110101111101001011	0.74	-0.02
A123a4a679	00100001010001001111101110100011001	0.19	-0.68
A123a5689	01110001000111001110000000001011011	0.07	-0.81
A13a3b4a4b59	00100101001010111110110111111100111	0.85	0.11
A13a3b4a4b69	00100001001000011110110110100110101	0.19	-0.68
A13a3b4a4b79	00100001001000011111111110100100101	0.19	-0.68
A13a3b4a569	00100101001010001110110111101110111	0.63	-0.15
A13a3b4a579	00101111001010001111110111101100111	0.85	0.11
A13a3b4a589	01110101001010001110110111101100111	0.74	-0.02
A13a3b4a679	00100001001000001111110110100110101	0.19	-0.68
A13a3b5689	01110101001010001110010000001110111	0.30	-0.54
A13a4a4b569	00100001000010111110100111111010011	0.41	-0.41
A13a4a4b579	00101011000010111111100111111000011	0.63	-0.15
A13a4a4b589	01110001100010111110100111111000011	0.63	-0.15
A13a4a4b679	00100001000000011111100110100010001	-0.18	-1.09
A13a4a5679	00101011000010001111100111101010011	0.41	-0.41
A13a4a5689	01110001000010001110100111101010011	0.30	-0.54
A123a3b4a4b59	00100101011111111111011111111101111	1.42	0.79
A123a3b4a4b69	00100001011001011110111110100111101	0.63	-0.15
A123a3b4a4b79	00100001011001011111111110100101101	0.63	-0.15
A123a3b4a569	00100101011111001110111111101111111	1.18	0.50
A123a3b4a579	00101111011111001111111111101101111	1.42	0.79

A123a3b4a589	01110101011111001110111111101101111	1.30	0.64
A123a3b4a679	00100001011001001111111110100111101	0.63	-0.15
A123a3b5689	01110101001111001110010000001111111	0.63	-0.15
A123a4a4b569	00100001010111111110101111111011011	0.96	0.24
A123a4a4b579	00101011010111111111101111111001011	1.18	0.50
A123a4a4b589	11110001110111111110101111111001011	1.30	0.64
A123a4a4b679	00100001010001011111101110100011001	0.30	-0.54
A123a4a5679	00101011010111001111101111101011011	0.60	0.24
A123a4a5689	01110001010111001110101111101011011	0.85	0.11
A123a4a5789	01111011010111001111101111101001011	1.07	0.37
A13a3b4a4b569	00100101001010111110110111111110111	0.96	0.24
A13a3b4a4b579	00101111001010111111110111111100111	1.18	0.50
A13a3b4a4b589	01110101101010111110110111111100111	1.18	0.50
A13a3b4a4b679	00100001001000011111110110100110101	0.30	-0.54
A13a3b4a5679	00101111001010001111110111101110111	0.96	0.24
A13a3b4a5689	01110101001010001110110111101110111	0.85	0.11
A13a3b4a5789	01111111001010001111110111101100111	1.07	0.37
A13a4a4b5679	00101011000010111111100111111010011	0.74	-0.02
A13a4a4b5689	01110001000010111110100111111010011	0.63	-0.15
A13a4a4b5789	01111011000010111111100111111000011	0.85	0.11
A13a4a56789	01111011000010001111100111101000011	0.52	-0.28
A123a3b4a4b569	00100101011111111110111111111111111	1.55	0.94
A123a3b4a4b579	00101111011111111111111111111101111	1.83	1.27
A123a3b4a4b589	11110101111111111110111111111101111	2.00	1.45
A123a3b4a4b679	00100001011001011111111110100111101	0.74	-0.02
A123a3b4a5679	00101111011111001111111111101111111	1.55	0.94
A123a3b4a5689	01110101011111001110111111101111111	1.42	0.79
A123a3b4a5789	01111111011111001111111111101101111	1.68	1.10
A123a4a4b5679	00101011010111111111101111111011011	1.30	0.64
A123a4a4b5689	11110001110111111110101111111011011	1.42	0.79
A123a4a4b5789	11111011110111111111101111111001011	1.69	1.10
A123a4a56789	01111011010111001111101111101011011	1.18	0.50
A13a3b4a4b5679	00101111001010111111110111111110111	1.30	0.64
A13a3b4a4b5689	01110101101010111110110111111110111	1.30	0.64
A13a3b4a4b5789	01111111101010111111110111111100111	1.55	0.94
A13a3b4a56789	01111111001010001111110111101110111	1.18	0.50
A13a4a4b56789	01111011100010011111100111101010011	0.85	0.11
A13a3b4a4b56789	01111111101010111111110111111110111	1.69	1.10
A123a4a4b56789	11111011110111111111101111111011011	1.83	1.27
A123a3b4a56789	01111111011111001111111111101111111	1.83	1.27
A123a3b4a4b5789	11111111111111111111111111111101111	2.67	2.15
A123a3b4a4b5689	11110101111111111110111111111111111	2.18	1.66
A123a3b4a4b5679	00101111011111111111111111111111111	2.00	1.45

A123a3b4a4b56789	1111111111111111111111111111111111	2.76	2.45

Table 28

Ability Estimates for Form C Hierarchy Two

Attributes Mastered	Expected Response Pattern	Ability Estimate	
		Group 1	Group 2
A0	00000000000000000000000000000000	-2.21	-2.17
A13a	00000000001000010011001000100000000	-1.29	-1.27
A13a8	00000000001000010111001000100000000	-1.12	-1.10
A13a9	00000000001000010011001001100000000	-1.12	-1.10
A13a5	10000000001000010011001000100000000	-1.12	-1.10
A13a6	00000000101000010011001000100000000	-1.12	-1.10
A13a4a	00001000001000010011001100100000100	-0.78	-0.76
A123a9	00000011001001110011001001110000000	-0.33	-0.31
A13a3b9	00000000001100011011001001100000010	-0.62	-0.60
A13a4a5	10001000001000010011001100100000100	-0.62	-0.60
A13a4a6	00001000101000010011001100100000100	-0.62	-0.60
A13a4a8	00001000011000010111001100100000100	-0.47	-0.45
A13a4a9	00001100001010010011111101100000100	-0.06	-0.04
A13a56	10000000101000010011001000100000000	-0.94	-0.92
A13a58	10000000001000010111001000100000000	-0.94	-0.92
A13a59	10000000001000010011001001100000000	-0.94	-0.92
A13a68	00000000101000010111001000100000000	-0.94	-0.92
A13a69	00000000101000010011001001101000000	-0.78	-0.76
A13a89	00000000001000010111001001100000000	-0.94	-0.92
A123a3b9	00000011001101111011001001110000010	0.07	0.08
A123a4a9	00000111001011110011111001110000000	0.19	0.21
A123a59	10000011001001110011001001110000000	-0.19	-0.17
A123a69	00000011101001110011001011111000000	0.07	0.08
A123a89	00000011001001110111001001110001000	-0.06	-0.04
A13a3b4a9	00001100001110011011111101100000110	0.31	0.32
A13a3b59	10000000001100011011001001100000010	-0.47	-0.45
A13a3b69	00000000101100011011001001101000010	-0.33	-0.31
A13a3b89	00000000001100011111001001100001010	-0.33	-0.31
A13a4a4b9	00011100001010010011111101100010101	0.31	0.32
A13a4a56	10001000101000010011001100100000100	-0.47	-0.45
A13a4a58	10001000011000010111001100100000100	-0.33	-0.31
A13a4a59	11101100001010010011111101100000100	0.31	0.32
A13a4a68	00001000101000010111001100100000100	-0.47	-0.45
A13a4a69	00001100101010010011111101101000100	0.19	0.21
A13a4a89	00001100011010010111111101100001100	0.31	0.32
A13a568	10000000101000010111001000100000000	-0.78	-0.76
A13a569	10000000101000010011001001101000000	-0.62	-0.60

A13a589	10000000001000010111001001100001000	-0.62	-0.60
A13a689	00000000101000010111001001100001000	-0.62	-0.60
A123a3b4a9	00001111001111111011111101110000110	0.89	0.89
A123a3b59	10000011001101111011001001110000010	0.19	0.21
A123a3b69	00000011101101111011001011111000010	0.43	0.44
A123a3b89	00000011001101111111001001110001010	0.31	0.32
A123a4a4b9	00011111001011110011111101110010101	0.89	0.89
A123a4a59	11101111001011110011111101110000100	0.89	0.89
A123a4a69	00001111101011110011111111111000100	0.89	0.89
A123a4a89	00001111011011110111111101110001100	0.89	0.89
A123a569	10000011101001110011001011111000000	0.19	0.22
A123a589	10000011001001110111001001110001000	0.07	0.08
A123a689	00000011101001110111001011110001000	0.19	0.21
A13a3b4a4b9	00011100001110011011111101100010111	0.66	0.67
A13a3b4a59	11101100001110011011111101100000110	0.66	0.67
A13a3b4a69	00001100101110011011111101101000110	0.55	0.55
A13a3b4a89	00001100011110011111111101100001110	0.66	0.67
A13a3b569	10000000101100011011001001101000010	-0.19	-0.17
A13a3b589	10000000001100011111001001100001010	-0.19	-0.17
A13a3b689	00000000101100011111001001101001010	-0.06	-0.04
A13a4a4b68	00001000111000010111001100100000100	-0.33	-0.45
A13a4a4b69	00011100101010010011111101101110101	0.66	-0.31
A13a4a4b89	00011100011010010111111101100011101	0.66	-0.31
A13a4a568	10001000111000010111001100100000100	-0.19	0.67
A13a4a569	01101100101010010011111101101000100	0.43	0.67
A13a4a589	11101100011010010111111101100001100	0.66	-0.17
A13a4a689	00001100111010010111111101101001100	0.55	0.44
A13a5689	10000000101000010111001001101001000	-0.33	0.67
A123a3b4a4b9	00011111001111111011111101110010111	1.23	0.55
A123a3b4a59	11101111001111111011111101110000110	1.23	-0.31
A123a3b4a69	00001111101111111011111111111000110	1.23	1.23
A123a3b4a89	00001111011111111111111101110001110	1.23	1.23
A123a3b569	10000011101101111011001011111000010	0.55	1.23
A123a3b589	10000011001101111111001001110001010	0.43	1.23
A123a3b689	00000011101101111111001011111001010	0.66	0.55
A123a4a4b59	11111111001011110011111101110010101	1.23	0.44
A123a4a4b69	00011111101011110011111111111110101	1.35	0.67
A123a4a4b89	00011111011011110111111101110011101	1.23	1.23
A123a4a569	11101111101011110011111111111000100	1.23	1.35
A123a4a589	11101111011011110111111101110001100	1.23	1.23
A123a4a689	00001111111011110111111111111001100	1.23	1.23
A13a3b4a4b59	11111100001110011011111101100010111	1.00	1.23
A13a3b4a4b69	00011100101110011011111101101110111	1.00	1.23

A13a3b4a4b89	00011100011110011111111101100011111	1.00	1.00
A13a3b4a569	11101100101110011011111101101000110	0.88	1.00
A13a3b4a589	11101100011110011111111101100001110	1.00	1.00
A13a3b4a689	00001100111110011111111101101001110	0.89	0.89
A13a3b5689	10000000101100011111001001101001010	0.07	1.00
A13a4a5689	11101100111010010111111101101001100	0.89	0.89
A13a4a4b569	11111100101010010011111101101110101	1.00	0.08
A13a4a4b589	11111100011010010111111101100011101	1.00	0.89
A13a4a4b689	00011100111010010111111101101111101	1.00	-0.17
A123a5689	10000011101001110111001011111000000	0.31	1.00
A123a3b4a4b59	11111111001111111011111101110010111	1.62	1.00
A123a3b4a4b69	00011111101111111011111111111110111	1.76	1.00
A123a3b4a4b89	00011111011111111111111101110011111	1.62	0.32
A123a3b4a569	11101111101111111011111111111000110	1.62	1.61
A123a3b4a589	11101111011111111111111101110001110	1.62	1.76
A123a3b4a689	00001111111111111111111111111001110	1.62	1.61
A123a3b5689	10000011101101111111001011111001010	0.77	1.61
A123a4a4b568	10001011110011101110011101100000100	0.55	1.61
A123a4a4b569	11111111101011100111111111111110101	1.76	1.61
A123a4a4b589	11111111011011101111111101110011101	1.62	0.78
A123a4a4b689	00011111111011101111111111111111101	1.76	0.55
A123a4a5689	11101111111011101111111111111001100	1.62	1.76
A13a3b4a4b568	10001000111100011111001100100000110	0.19	1.61
A13a3b4a4b569	11111100101110011011111101101110111	1.35	1.76
A13a3b4a4b589	11111100011110011111111101100011111	1.35	1.61
A13a3b4a4b689	00011100111110011111111101101111111	1.35	0.21
A13a3b4a5689	11101100111110011111111101101001110	1.23	1.35
A13a4a4b5689	11111100111010010111111101101111101	1.35	1.35
A13a3b4a4b5689	11111100111110011111111101101111111	1.76	1.35
A123a4a4b5689	11111111111011101111111111111111101	2.30	1.23
A123a3b4a5689	11101111111111111111111111111001110	2.10	1.35
A123a3b4a4b689	00011111111111111111111111111111111	2.30	1.76
A123a3b4a4b589	11111111011111111111111101110011111	2.10	2.29
A123a3b4a4b569	11111111101111111011111111111110111	2.30	2.09
A123a3b4a4b5689	11111111111111111111111111111111111	2.88	2.29

Following the assignment of ability estimates to knowledge states, the attribute hierarchy method was applied to the data from the three test forms using the Fortran program. The Fortran program successfully applied the attribute hierarchy method to the data. All output files were successfully created and saved. The probability file was then used to evaluate model-data fit.

Analysis of model-data fit. Having fit a diagnostic model to the data using the attribute hierarchy method, the final step in the Stage Two process was to examine model-data fit in response to research questions one and two. Results from the two groups of examinees were compared to cross-validate the findings. Table 29 that follows displays classification consistency and HCI fit values for each hierarchy, group, and form combination. Across the two groups, similar values were obtained, indicating consistency of findings. Classification consistency values fell between .92 and .98, indicating that for all form, hierarchy, and group combinations, over 90% of students were successfully classified into a single knowledge state. Similarly, all HCI values fell between .87 and .90 for all form, hierarchy, and group combinations, indicating that both hierarchies had excellent model-data fit, according to the specifications put for by Leighton et al. (2009). Because an exceptional level of fit was observed across all three forms, it was determined that a single model of reading comprehension could be applied to multiple forms of a critical reading assessment.

Table 29

Model-Data Fit Indices by Hierarchy, Form, and Group

Hierarchy, Form, & Group	Classification Consistency	HCI
Hierarchy 1, Form A, Group 1	0.93	0.87
Hierarchy 1, Form A, Group 2	0.95	0.88
Hierarchy 1, Form B, Group 1	0.96	0.88
Hierarchy 1, Form B, Group 2	0.98	0.90
Hierarchy 1, Form C, Group 1	0.92	0.90
Hierarchy 1, Form C, Group 2	0.92	0.90
Hierarchy 2, Form A, Group 1	0.95	0.87
Hierarchy 2, Form A, Group 2	0.96	0.87
Hierarchy 2, Form B, Group 1	0.96	0.88
Hierarchy 2, Form B, Group 2	0.98	0.89
Hierarchy 2, Form C, Group 1	0.94	0.88
Hierarchy 2, Form C, Group 2	0.98	0.88

While both the models obtained from Hierarchy One and Hierarchy Two demonstrated excellent fit to the data, one model was selected to move forward with the study. To make the decision, model-data fit was examined. Both models provided roughly equivalent fit to the data, as evidenced by overlapping values for both the classification consistency and HCI. Because of this finding, the model based on Hierarchy One was selected for retention, as it was the more parsimonious model and provided roughly equivalent fit and classification consistency as did the more complicated model.

In response to research question one regarding which of the two models provided better fit to the data, the researcher concluded that while both models provided excellent fit, the more parsimonious model was retained because the slightly more complex model did not add noticeable difference in model-data fit. In response to research question two regarding whether a single model could be fit to multiple forms of a reading comprehension assessment, the researcher determined that because both models provided excellent fit across the three forms, it could be concluded that a single model could in fact be retrofit to multiple forms of an assessment.

Chapter 5 – Discussion

Cognitive diagnostic modeling has become increasingly implemented in the areas of educational measurement and cognitive psychology. Various stakeholders, including test takers, educators, and state vendors desire greater information when reporting results from assessments, beyond a simple total score summary. While previous research has identified various challenges associated with reporting at the subscore level, the use of diagnostic models has arisen as one way to provide the more fine-grained score reporting desired by stakeholders.

Although diagnostic models are often implemented through the construction of a complete cognitive diagnostic assessment system, the practice of retrofitting diagnostic models to assessments already in use has become more and more prevalent in response to these reporting demands. The practice of retrofitting diagnostic models to assessments already in use has many limitations, including limitations to the effectiveness of the model due to the pre-specified set of items measuring various skills. However, despite these shortcomings retrofitting is increasingly being implemented to meet the reporting demands of stakeholders due to the ability to report at the attribute level. Because of this implementation, it is imperative that models used for retrofitting accurately describe the scope of the skills required for the assessment to which it is being fit.

The current practice of retrofitting a diagnostic model to an assessment already in use typically begins by examining the items on a single form and specifying a model to encompass the skills measured by that form of the assessment. While this approach allows researchers to ensure that the attributes included in the model accurately reflect the content of the assessment, the model is typically only fit to that single test form alone. In

operational practice, it is desirable to be able to use a single diagnostic model across multiple forms of an assessment, whether administered within or across years, to ensure consistency when reporting at the attribute level and allow for the possibility of future construction of a cognitive diagnostic assessment using the same model.

The current study sought to address this gap in the research by examining the extent that a single diagnostic model of reading comprehension could be fit across three parallel forms of a passage-based critical reading assessment. In addition, rather than specifying a unique model based on the items contained on each form, two different diagnostic models were specified based on previous research in the area of passage-based reading comprehension assessments, and their fit compared across the forms to determine if one provided better fit over and above the other.

Towards this aim, the current study sought to address two research questions, as follows:

1. Which hierarchy of cognitive skills related to responding to passage-based reading comprehension provides the best fit to the data, a more parsimonious or slightly more complex model?
2. Can a single cognitive diagnostic model of reading comprehension be fit to multiple forms of a critical reading assessment using the attribute hierarchy method?

To respond to these research questions, the research presented in the current study followed two specific areas of inquiry. In order to determine whether a single model of passage-based reading comprehension could be fit across multiple forms of data, the models had to first be specified. Based on a review of previous research in diagnostic

modeling for reading comprehension assessments, it was determined that two models should be compared. Thus, research conducted to respond to the first research question specifically focused on determining whether a more parsimonious diagnostic model fit the three forms of data better than a slightly more complex diagnostic model.

The research conducted to address the second research question in this study pertained to determining whether a single diagnostic model of reading comprehension could be fit across multiple forms of data. Fit to multiple forms of data required evidence of model-data fit across all three forms. Once it had been determined which diagnostic model would be retained, the three forms of data were evaluated to determine whether approximately equivalent model-data fit was obtained across all three administrations, to support the use of a single model across multiple forms of data.

In order to obtain results for each of the two aforementioned research questions, the study was organized into two distinct stages of research. Stage One included the specification of attribute hierarchies and coding of items for the attributes necessary to provide a correct response. Stage Two consisted of data analysis and model fit. Conclusions drawn based on the results of the two research questions are discussed in the sections that follow, organized by the two stages of research outlined in the methods and results sections.

Stage One

The first stage of research laid the framework upon which the remainder of the study was conducted, and thus could be argued to be the most important aspect of the research presented here. The research began by first specifying the relevant attributes to be included in the hierarchy. Without proper identification of the relevant attributes for the

model, the following stages of research would suffer, as evidence of strong model-data fit would not be possible.

Specification of attributes. In order to ensure support for the use of attributes in the model, relevant research was reviewed for attributes included in previous models of passage-based reading comprehension. This approach allowed for justification of the inclusion of each of the attributes in the model. However, by only including those attributes previously specified in the literature it is possible that some necessary attributes were not included had they not been previously identified in other models. As such, future research should be conducted in order to continue to explore the necessary attributes to be included in a model of passage-based reading comprehension. This will ensure that all components that should be included in the model are represented. Furthermore, future research may need to be conducted to determine the extent that the model varies by grade or type of text the student encounters (e.g. narrative versus expository).

If upon further exploration it is determined that the model includes the necessary attributes, future research might be conducted to expand on the use of the hierarchies specified by Wang and Gierl (2011). Researchers might determine the extent that the hierarchies can be applied to additional measures of passage-based reading comprehension, such as with other assessments administered to high schools students for statewide accountability or other decision-making purposes. Should the hierarchies be found to be representative of the domain, they could potentially serve as the basis for the construction of new cognitive diagnostic assessments specifically designed to assess passage-based reading comprehension.

Coding process. In the present study, content experts collaborated to code the items for each of the attributes included in the model. While obtaining a consensus decision regarding the coding of items to attributes ensured that a single Q-matrix could be obtained, using a consensus approach might have altered the scope of the matrix that would have been obtained had a different coding approach been used, for example had the content experts coded the items for the attributes independently. During the consensus process, the experts might have been swayed by the views of their peers, particularly those with more dominant personalities, seniority, or other factors, which could have impacted the ultimate coding of the attributes to the items. Future research could be conducted to determine the extent that the Q matrix varies based on the coding approach used, and which of these matrices provides the best fit to the data.

A potential limitation resulting from using a set of attributes defined in previous studies was a lessened understanding of the scope of each attribute by those conducting the coding. During the coding process, the attributes in the hierarchy were at times interpreted in different ways among the three content experts. Without having actually specified the attributes, the researchers relied on descriptions in the literature to clarify any confusion rather than being able to describe the attributes from the perspective of having created them. This limitation could potentially be avoided in the future by ensuring that content experts thoroughly discuss the meaning of the attributes, ensuring the inclusion of examples of items that measure each attribute, and ensuring that the rationale and meaning of the attributes included in the diagnostic model are thoroughly documented in the literature going forward.

To address this constraint in the current study, the content experts engaged in conversation regarding the meaning of the attributes and the structure of the model. The content experts determined that in order to code attribute 9 (evaluate response options) as required for a correct response to an item, the student would have to consult each of the response options prior to selecting their response. In contrast, an item would not require Attribute 9 if the student could read the stem and provide the answer without having to read through the options (e.g. if the options had not presented, the student would still be able to provide the correct response). Similarly, the content experts determined that for Attributes A4a and A4b, which required understanding of "larger sections of text," larger sections included text spanning multiple paragraphs. Attributes 7 and 8, complex vocabulary and difficulty syntactic structure respectively, were subjectively evaluated based on the content experts' shared experiences working with students at the high school level. However, future iterations would likely benefit from the use of a specific framework for evaluating the complexity of the text.

Upon review of the final set of codes, several interesting findings were observed. For example, after reviewing the codes for Form A it became evident that all occurrences of Attribute 8 were always coded with Attribute 9. This indicates that for this form, items that required students to use rhetorical knowledge also required them to evaluate the response options in order to select the current answer. Test developers could potentially use this information to evaluate whether that combination of skills was actually intended by the item or if in fact they had intended students to be able to use rhetorical knowledge without needing to first see what the possible responses were. Such analyses would inform subsequent item development as well as future uses of the model.

An additional interesting finding resulting from examination of the final set of codes provided by the content experts was the sheer number of times Attribute 9 was required to provide a correct response to an item (coded as a required skill for between 22 and 27 items on each form). This finding demonstrates that in order to be successful on the assessment, a student must largely be able to discern between a set of provided responses rather than being able read the question and know the answer without consulting the options for the best choice. This finding has interesting implications for the construct that the assessment actually seems to be measuring and the types of skills a student would need to have mastered in order to be successful.

Stage Two

The second stage of research involved analysis of the data, which was conducted over many steps. In order to conduct the data analysis necessary for fitting the model, expected response patterns had to first be generate for each of the forms.

Expected response generation. All possible expected response patterns were hand-generated by the researcher by listing all the possible combinations of attributes that could have been mastered. This involved consideration of both the prerequisites required by each attribute as well as actual observations of the ways the attributes could be combined. For example, for Form A, Attribute 8 always occurred with Attribute 9. Thus, combinations that included Attribute 8 but not Attribute 9 were not included in the set of expected response patterns for Form A.

One limitation of having a single researcher list the expected response patterns was that by listing the knowledge states by hand, the potential for error increased. An expected response pattern that should have been included could potentially have been omitted.

While the researcher ensured that all expected response patterns were double checked for accuracy, there was no process for checking reliability with a second researcher. In future instances requiring the retrofitting a diagnostic model, researchers should be mindful of this potential for error. Although time-consuming, researchers might consider cross validating the list of expected response patterns with at least one other person to ensure that all knowledge states are properly identified.

Fortran programming. To prepare the data for the Fortran program fitting the model to the data using the attribute hierarchy method, files had to first be prepared for each test form. As there were multiple forms, groups, and hierarchies to consider, this resulted in a total of twelve instances where the model was fit to the data. Each of the twelve instances required the pre-specification of the requisite input files in order to fit the model. The preparation of these input files constituted one of the more time-consuming aspects of the project, as the data files had to be entered in a specific format, item response theory parameter estimates obtained for the three-parameter logistic model, and ability estimates for each of the knowledge states. In addition to the time requirement to prepare these files, the process also required organization of the necessary files by the researcher. This process is highlighted here to alert researchers wishing to apply the model in future iterations of the planning and time requirements necessary for this portion of the project.

Once the data files had been prepared for analysis, the Fortran program was used to fit the model to the data using the attribute hierarchy method. While the use of a Fortran program successfully allowed for completion of the current study, the approach did have several limitations, including the need to run the program twelve times for each hierarchy-by-form-and-group combination. Future research may want to explore more time efficient

approaches that can be used in operational testing programs. In addition, future research should likely be conducted to create more manageable output files. In the current format, the files were quite large due to the sheer volume of information being recorded. Specifically, the combination of a large sample size, 35 items, and a large number of possible knowledge states created data files that were cumbersome to manage. In some cases the files were too large to even open, and could only be accessed by opening a portion of the data file. Perhaps the output files could be better structured to only include the necessary information to facilitate interpretation of the data (e.g. only including the information relevant to the knowledge state into which the individual is classified).

Research question 1. It was during the second stage of research that Research Question 1 was addressed. In order to determine whether a more parsimonious or more complex model better fit the data, the fit of the two models to the three forms of data were evaluated for fit using classification consistency values and the hierarchy consistency index. Based on these indices, it was determined that the more complex model did not provide additional information over and above the more parsimonious model. As a result, the simpler model was retained for subsequent analyses in Stage Two of the research.

The decision to retain the more parsimonious model was based on the values obtained for the fit indices. The two models explored in the current study had approximately equivalent fit as evidenced by the classification consistency and HCI values. In the original study conducted by Wang and Gierl (2011), in which the hierarchies were originally specified, the two models were also found to have nearly equivalent fit to the data. In both studies, the more parsimonious model was retained because the more complex model did not appear to add any additional information with the inclusion of the

two additional attributes to the model. However, according to theory and other studies conducted in the area of passage-based reading comprehension, these two attributes represent important concepts that should be included when considering the skills necessary to provide correct responses to items measuring reading comprehension. Thus, for theoretical reasons, it may be desirable to instead select the more complex model as a better representation of the construct. Future research should also be conducted to determine the extent that an increase in items, particularly those measuring the additional attributes in Hierarchy Two, necessitates a positive change in model-data fit. Similarly, future research may be conducted to determine whether the additional knowledge states allowed by the more complex model better represents students' skill sets than the fewer knowledge states obtained by using the more parsimonious model.

As previously stated, comparison of the fit of the two models based on the attribute hierarchies was contingent on proper model specification and coding of the attributes to the items. To the extent that the codes used in the current study deviate from what the codes should actually be, model-data fit would be impacted.

In addition, the current study included classification consistency and the HCI as fit indices based on previous research conducted in the area of diagnostic modeling. However, additional fit indices could have been chosen, for example, including the recent item consistency index (Lai, Gierl, & Cui, 2012). The use of different fit indices could potentially provide support for differences in the ultimate conclusions made in response to Research Question 1.

Research question 2. Research Question 2 was also addressed during the second stage of research. In response to the second research question, addressing whether a single

model could be fit across multiple forms, the classification consistency and HCI values were determined to be quite similar over the three forms. Because the values were so similar it was concluded that a single model of passage-based reading comprehension could be fit across multiple years of data.

As previously stated, both Hierarchy One and Hierarchy Two provided evidence of excellent model-data fit. While the more parsimonious model was retained for the present study, the conclusions reached for the present research question regarding whether a single model could be fit across multiple forms of data would be retained regardless of the hierarchy selected in response to research question one. Both models demonstrated excellent fit across all three forms, providing support for the use of a single diagnostic model being fit across multiple consecutive years of assessment data.

While future research is needed to support this finding, the implications resulting from the essentially equivalent fit across multiple forms indicates that a single model could be used by a testing company to provide a consistent diagnostic modeling approach over multiple administrations. This finding provides support for the possibility of retrofitting test forms to a single model both within and across years, as would benefit those programs that desire increased reporting capabilities but are not yet ready to transition to a diagnostic assessment platform for various reasons (e.g. funding, time and resource constraints). However, those implementing such an approach should be mindful of the limitations associated with retrofitting diagnostic models and ensure that the items used in future iterations of the assessment assess the skills represented in the diagnostic model. To the extent that the test plan or construct are modified, the diagnostic model used should

also be revisited and fit indices explored to ensure that the model accurately represents the skills test takers engage in during the assessment.

Study Limitations

In addition to the limitations previously specified for each of the two stages of research, one overarching limitation in the current study involved the challenge of retrofitting a model to previously administered test forms. As in the present study, attributes can only be coded for those items that are present on the form. As such, it can be challenging to anticipate whether a sufficient number of items are present to warrant the construction of diagnostic score reports a priori. Rather, the coding process must first be conducted in order to determine if a sufficient number of items are present to reliably measure each attribute in the model. Similarly, a retrofitted model may not be as encompassing as a model that would be used as the foundation for development of a cognitive diagnostic assessment, impacting model specification and fit, particularly if there are not an adequate number of items associated with the identified cognitive processes (see Gierl & Cui, 2008; Rupp & Templin, 2008). However, basing the model on sound cognitive theory and incorporating a large number of items may mitigate this effect.

Summary

The findings of the current study suggest the retention of the more parsimonious model of passage-based reading comprehension. In addition, the findings indicate that a single hierarchy of skills can be fit to multiple parallel forms of an assessment. However, this study also demonstrates a need for continued research in the area of diagnostic modeling, particularly with regard to the process of retrofitting diagnostic models to assessments already in use.

Summary of potential future research. The research conducted in the present study affords many opportunities for future research. These opportunities are summarized below:

1. Exploration of additional attributes to be included in the attribute hierarchies
2. Analysis to determine if attribute hierarchies should vary by grade or type of text administered (e.g. narrative or expository)
3. Application of hierarchies to additional test forms, grades, and assessments
4. Evaluate extent that Q matrix differs based on coding approach
5. Comparison of model-data fit based on Q matrices to determine optimal fit to the data
6. Inclusion of specific framework for evaluating text complexity attributes
7. Exploration of more time-efficient approaches for applying model across multiple form-hierarchy-group combinations
8. Determination of more manageable output files for the Fortran program
9. Exploration of the impact of including additional items for attributes in hierarchy two to determine if a change is obtained in model-data fit
10. Analysis to determine if the additional knowledge states included for Hierarchy Two better captures student skill sets
11. Replication of finding that a single model can be applied across multiple forms of an assessment

References

Birenbaum, M., & Tatsuoka, K. K. (1993). Applying an IRT-based cognitive diagnostic model to diagnose students' knowledge states in multiplication and division with exponents. *Applied Measurement in Education, 6*, 255-268. doi: 10.1207/s15324818ame0604_1

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 47*, 423-466. doi: 10.1111/0023-8333.00016

Buck, G., & Tatsuoka, K. K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*, 119-157. doi: 10.1177/026553229801500201

Buck, G., VanEssen, T., Tatsuoka, K. K., Kostin, I., Lutz, D., & Phelps, M. (1998). Development, selection, and validation of a set of cognitive and linguistic attributes for the SAT I verbal, sentence completion section. Princeton, NJ: Educational Testing Service.

Cai, L., Thissen, D., & du Toit, S. (2011). IRTPRO for Windows. Lincolnwood, IL: Scientific Software International.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*, 595-624. doi: 10.1007/s11336-008-9063-2

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, pp. 979-1030). Amsterdam: Elsevier.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11*, 175-193. doi: 10.1177/014662168701100207

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta psychologica, 37*, 359-374. doi: 10.1016/0001-6918(73)90003-6

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of Educational Measurement, 44*, 325-340. doi: 10.1111/j.1745-3984.2007.00042.x

Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: Interdisciplinary Research & Perspective, 6*, 263-268. doi: 10.1080/15366360802497762

Gierl, M. J., Cui, Y., & Hunka, S. (2007, April). *Using connectionist models to evaluate examinees' response patterns on tests: An application of the attribute hierarchy method to assessment engineering.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Gierl, M. J., Leighton, J. P., Wang, C., Zhou, J., Gokiert, R., & Tan, A. (2009). Validating cognitive models of task performance in algebra on the SAT®. New York: The College Board.

Gierl, M. J., Wang, C., & Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT. *The Journal of Technology, Learning and Assessment, 6*(6), 4-49.

Gierl, M. J., Zheng, Y., & Cui, Y. (2008). Using the attribute hierarchy method to identify and interpret cognitive skills that produce group differences. *Journal of Educational Measurement, 45*, 65-89. doi: 10.1111/j.1745-3984.2007.00052.x

Gorin, J. S., & Embretson, S. E. (2006). Item diffficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30*, 394-411. doi: 10.1177/0146621606288554

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60).

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL.* (Ph.D. 3182288), University of Illinois at Urbana-Champaign, United States -- Illinois. ProQuest Dissertations & Theses (PQDT) database.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing, 26*, 31-73. doi: 10.1177/0265532208097336

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258-272. doi: 10.1177/01466210122032064

Lai, H., Gierl, M. J., & Cui, Y. (2012, April). *Item consistency index: An item-fit index for cognitive diagnostic assessment*. Paper presented at the annual meeting of the National Council on Measurment in Education, Vancouver, BC.

Leighton, J. P., Cui, Y., & Cor, M. K. (2009). Testing expert-based and student-based cognitive models: An application of the attribute hierarchy method and hierarchy consistency index. *Applied Measurement in Education, 22*(3), 229-254. doi: 10.1080/08957340902984018

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26*(2), 3-16. doi: 10.1111/j.1745-3992.2007.00090.x

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*, 205-237. doi: 10.1111/j.1745-3984.2004.tb01163.x

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement, 33*, 579-598. doi: 10.1177/0146621609331960

Marini, J. P., Mattern, K. D., & Shaw, E. J. (2011). Examining the linearity of the PSAT/NMSQT-FYGPA relationship. New York: College Board.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379-416. doi: 10.1111/j.1745-3984.1996.tb00498.x

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research & Perspective, 6*, 219-262. doi: 10.1080/15366360802490866

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*, 333-352. doi: 10.1111/j.1745-3984.1997.tb00522.x

Sinharay, S., Puhan, G., & Haberman, S. J. (2009, April). *Reporting diagnostic scores: Temptations, pitfalls, and some solutions*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: American Council on Education/Macmillan.

Svetina, D., Gorin, J. S., & Tatsuoka, K. K. (2011). Defining and comparing the reading comprehension construct: A cognitive-psychometric modeling approach. *International Journal of Testing, 11*, 1-23. doi: 10.1080/15305058.2010.518261

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354. doi: 10.1111/j.1745-3984.1983.tb00212.x

Tatsuoka, K. K. (1991). Boolean algebra applied to determination of universal set of knowledge states: DTIC Document.

Tatsuoka, K. K. (1993). Item construction and psychometric models appropriate for constructed responses. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cogntive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 107-133). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Routledge.

Trout, D. L., & Hyde, B. (2006, April). *Developing score reports for statewide assessments that are valued and used: Feedback from K-12 stakeholders.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

VanderVeen, A., Huff, K., Gierl, M., McNamara, D. S., Louwerse, M., & Graesser, A. (2007). Developing and validating instructionally relevant reading competency profiles measured by the critical reading section of the SAT Reasoning Test™. In D. S. McNamara (Ed.), *Reading comprehension strategies* (pp. 137-171). Mahwah, NJ: Erlbaum.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287-307. doi: 10.1348/000711007x193957

Wang, C., & Gierl, M. J. (2007, April). *Investigating the cognitive attributes underlying student performance on the SAT® critical reading subtest: An application of the Attribute Hierarchy Method.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Wang, C., & Gierl, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement, 48*, 165-187. doi: 10.1111/j.1745-3984.2011.00142.x

Wang, C., Gierl, M. J., & Leighton, J. P. (2006, April). *Investigating the cognitive attributes underlying student performance on a foreign language reading test: An application of the attribute hierarchy method.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Yang, X., & Embretson, S. E. (2007). Construct validity and cognitive diagnostic assessment. In J. P. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 119-145). New York: Cambridge University Press.

Appendix A

Human Subjects Approval

1/2/2013
HSCL #20589

Amy Clark
CETE
730 JRP

The Human Subjects Committee Lawrence Campus (HSCL) has reviewed your research project application

20589 Clark - Kingston (CETE) Validation of a Cognitive Diagnostic Model of Reading Comprehension

and approved this project under the expedited procedure provided in 45 CFR 46.110 (f) (5) Research involving materials (data, documents, records, or specimens) that have been collected, or will be collected solely for non-research purposes. As described, the project complies with all the requirements and policies established by the University for protection of human subjects in research. Unless renewed, approval lapses one year after approval date.

Since your research presents no risk to participants and involves no procedures for which written consent is normally required outside of the research context HSCL has waived the requirement for a signed consent form (45 CFR 46.117 (c) (2).

1. At designated intervals until the project is completed, a Project Status Report must be returned to the HSCL office.
2. Any significant change in the experimental procedure as described should be reviewed by this Committee prior to altering the project.
3. Notify HSCL about any new investigators not named in original application. Note that new investigators must take the online tutorial at http://www.rcr.ku.edu/hscl/hsp_tutorial/000.shtml.
4. Any injury to a subject because of the research procedure must be reported to the Committee immediately.
5. When signed consent documents are required, the primary investigator must retain the signed consent documents for at least three years past completion of the research activity. If you use a signed consent form, provide a copy of the consent form to subjects at the time of consent.
6. If this is a funded project, keep a copy of this approval letter with your proposal/grant file.

Please inform HSCL when this project is terminated. You must also provide HSCL with an annual status report to maintain HSCL approval. Unless renewed, approval lapses one year after approval date. If your project receives funding which requests an annual update approval, you must request this from HSCL one month prior to the annual update. Thanks for your cooperation. If you have any questions, please contact me.

Sincerely,

Christopher Griffith, J.D.
Assistant Coordinator
Human Subjects Committee - Lawrence

cc: Neal Kingston

Human Subjects Committee Lawrence
Youngberg Hall I 2385 Irving Hill Road I Lawrence, KS 66045 I (785) 864-7429 I HSCL@ku.edu I www.rcr.ku.edu/hscl

Appendix B

Fortran Code

```
PROGRAM PROBABILITY

INTEGER,ALLOCATABLE::ER(:,:),OR(:,:),ITEMA(:),ITEMB(:),DA(:,:,:),DB(:,:,:).
REAL,ALLOCATABLE::THETA(:),AA(:),BA(:),CA(:),PA(:,:,:),PPA(:,:),AB(:),BB(:),CB(:),PB(:,:,:),
PPB(:,:)
CHARACTER(8)::ID(45),SID(10000),PPP

NE=10000
NI=36
NER=20

ALLOCATE
(ER(NER,NI),OR(NE,NI),THETA(NER),A(NI),B(NI),C(NI),ITEM(NI),D(NE,NI,NER),P(NE,NI,N
ER),PP(NE,NER))

OPEN (10,FILE="ERV.OUT")
100 FORMAT (A2,1X,20I1,1X,F9.6)

OPEN (20,FILE="DATA.TXT")
110 FORMAT (A8,1X,20I1)

OPEN (30,FILE="BILOG_PAR.OUT")
120 FORMAT (T7,I2,1X,3F9.5)

DO i=1,NER
 READ (10,100) ID(i),(ER(i,j),j=1,NI),THETA(i)
END DO

DO i=1,NE
 READ (20,110) SID(i),(OR(i,j),j=1,NI)
END DO

DO i=1,NI
 READ (30,120) ITEM(i),A(i),B(i),C(i)
END DO

OPEN (50,FILE="DjA.OUT")
150 FORMAT (100I3)

DO i=1,NE
 DO j=1,NI
  DO k=1,NER
```

```
   D(i,j,k)=ER(k,j)-OR(i,j)
   END DO
  END DO
END DO

DO i=1,NE
  DO k=1,NER
   WRITE (50,150) (D(i,j,k),j=1,NI)
  END DO
END DO

DO i=1,NE
  DO j=1,NI
    DO k=1,NER
    IF (D(i,j,k)==-1) THEN
     P(i,j,k)=C(j)+((1-C(j))/(1+EXP(-1.702*A(j)*(THETA(k)-B(j)))))
    ELSE IF (D(i,j,k)==1) THEN
     P(i,j,k)=1-(C(j)+((1-C(j))/(1+EXP(-1.702*A(j)*(THETA(k)-B(j))))))
    ELSE IF (D(i,j,k)==0) THEN
     P(i,j,k)=1
    END IF
    END DO
  END DO
END DO

OPEN (60,FILE="Prob.OUT")
160 FORMAT (2i5,2x,100F10.6)
DO i=1,NE
  DO k=1,NER
   WRITE (60,160) i,k,(P(i,j,k),j=1,NI)
  END DO
END DO

OPEN (70,FILE="FINALTHETAA.OUT")
170 FORMAT (F10.6,1X,F10.6,1X,20I2)
180 FORMAT ("OBSERVED RESPONSE PATTERN -- ",100I2)
190 FORMAT (//)
200 FORMAT ("    THETA    P(THETA)         EXPECTED RESPONSE MATRIX")
210 FORMAT ("EXAMINEE # ",A8)
220 FORMAT ("#",I5)
PP=1
DO i=1,NE
  DO j=1,NI
   DO k=1,NER
      PP(i,k)=(PP(i,k)*P(i,j,k))
   END DO
```

```
  END DO
END DO
DO i=1,NE
  WRITE (70,220) i
  WRITE (70,180) (OR(i,j),j=1,NI)
  WRITE (70,210) SID (i)
  WRITE (70,190)
  WRITE (70,200)
  DO k=1,NER
  WRITE (70,170) THETA (k),PP(i,k),(ER(k,j),j=1,NI)
  END DO
WRITE (70,190)
END DO

OPEN (80,FILE="PP.OUT")
250 FORMAT
(A8,1X,F10.6,1X,F10.6,1X,F10.6,1X,F10.6,1X,F10.6,1X,F10.6,1X,F10.6,1X,F10.6,1X,F10.6,1X,
F10.6)
DO i=1,NE
  WRITE (80,250) SID(i),(PP(i,k),k=1,NER)
END DO
END PROGRAM
```