APPLICATIONS OF EXPLORATORY Q-MATRIX DISCOVERY PROCEDURES IN DIAGNOSTIC CLASSIFICATION MODELS

BY

Emily Fall

Submitted to the graduate degree program in Psychology
and the Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Master's of Arts.

Chairperson Kristopher J. Preacher, PhD

William P. Skorupski, PhD

Jonathan L. Templin, PhD

Date defended: _____

The Thesis Committee for Emily Fall certifies

that this is the approved Version of the following thesis:

APPLICATIONS OF EXPLORATORY Q-MATRIX DISCOVERY PROCEDURES

IN DIAGNOSTIC CLASSIFICATION MODELS

Committee:

_____

Chairperson: Kristopher J. Preacher, PhD

_____

_____

Date approved: _____

APPLICATIONS OF EXPLORATORY Q-MATRIX DISCOVERY PROCEDURES IN DIAGNOSTIC CLASSIFICATION MODELS

BY

Emily Fall

Submitted to the graduate degree program in Psychology
and the Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Master's of Arts.

Chapter 1
Introduction

In 1918, Edward Thorndike observed that "Whatever exists at all exists in some quantity. To know it thoroughly involves knowing its quantity as well as its quality" (p. 16). From the time Thorndike made this simple statement, it has become a defining principle of educational measurement. The belief that traits and abilities could be examined quantitatively launched the field of psychometrics.

In 1923, Boring defined intelligence as "what the tests test." IQ tests were constructed, and the scores on those tests were translated directly as intelligence. The same has been done for many other traits, most notably, personality. The circular logic of Boring's definition confounds both intelligence researchers, and those who seek to further the field of measurement. However, this early definition propelled psychometric research into the dynamic domain it is today.

Since that time, in both psychology and education, one goal has become a major focus: diagnosis. For psychologists, this may mean characterizing personality traits or identifying psychological disorders. In the world of education, diagnosis means the identification of cognitive abilities or skills, particularly those relevant to large-scale assessment. Though the implementation of legislation such as No Child Left Behind (2001) has placed more emphasis on skill measurement and diagnosis in recent years, ability and intelligence have historically been focal points for psychologists and educational researchers.

Although the idea of psychometric assessment is not new, the world of educational testing has higher consequences than ever before. National educational

policy has made school funding dependent upon student performance on standardized tests, and poor performance can penalize needy students and schools even further. Furthermore, the material that is presented on these tests drives the curriculum, a practice often referred to as "teaching to the test." It is important to note that not only what is assessed, but how it is assessed has implications for what happens in education (Johnston & Costello, 2005). Therefore, it is paramount that educational tests be constructed to be as valid and efficient as possible.

Modern psychometrics has evolved over a series of stages beginning with simple models in classical test theory (CTT), to more complex latent variable models in item response theory (IRT). Increasing importance is placed not only on the types of general ability measured by large-scale assessments, but also on the individual skills that make up ability. Stakeholders want diagnostic measurements that reflect higher-order cognitive processes rather than test-specific strategies (Leighton & Gierl, 2007). To obtain this diagnostic information both the instruments used to assess individuals and the statistical procedures used to analyze the assessments must be designed with diagnostic properties in mind. Therefore, more recently, the latent framework of IRT has been applied to categorical latent variable models, which allow examinees to be categorized into diagnostic groups.

These Diagnostic Classification Models (DCM) are powerful analytic tools that provide information about student knowledge and ability by assessing skills demonstrated through a variety of testing frameworks. These models are particularly useful when diagnostic decisions about proficiency are required (Rupp & Templin,

2008a). In order for these procedures to be effective, the input for the models must be carefully considered. The set of skills an instrument measures as well as the patterns of these skills—which are reflected in the items on a test—determine the quality of the diagnostic information that can be obtained from these models.

For DCM, the skills represented by items are modeled using a Q-matrix. Tatsouka (1995) called the patterns of skills measured by these models *knowledge structures* and developed the Q-matrix to reflect examinee mastery of skills. Through a pattern of 0's and 1's, the Q-matrix establishes the relationships between latent variables representing knowledge structures (columns) and individual items on an assessment (rows) (Rupp & Templin, 2008b). Because the information contained within Q-matrix is the primary driver of the usefulness of a DCM, correct specification of the Q-matrix is essential.

When these assessments are written, experts in the field are often consulted to construct the items based on the skills deemed necessary to demonstrate proficiency in the content area. Therefore, incorporating expert opinion into the evaluation of skills on established instruments is important for accurate Q-matrix specification. However, because higher-order skills can be difficult to measure, experts often disagree on which of these skills are truly required to correctly respond to items. It is for this reason that probabilistic Q-matrix estimation, which allows for uncertainty, may lead to more accurate Q-matrices and, by extension, to better diagnostic information. The goal of the current work is to discover the most appropriate structure for the Q-matrix for each test and evaluate the impact of differing Q-matrix

structures on the quality of the models and their diagnostic information. The primary hypothesis of this investigation is that using probabilistic estimation methods for Q-matrix construction will yield DCMs with better fit, more stable parameter estimates, and more accurate classification rates.

To begin, it is important first to understand the emergence of psychometric theory and its impact on the current philosophy of cognitive diagnostic assessment. Further, the theory and applications of diagnostic classification, as well as how these models can be applied to reading comprehension tests, must be understood. Finally, probabilistic procedures will be applied to empirical data collected from adult education participants on a series of reading comprehension tests to demonstrate the usefulness of probabilistic estimation in discovering the structure of the Q-matrix.

Chapter 2

Background

Early Psychometric Theory

The most simplistic view of educational testing begins with classical test theory. Classical test theory (CTT) assumes that the test score (Y) of an individual consists of the true score (T), or ability of the examinee, plus measurement error (E), and is given as follows:

$$Y = T + E$$

(Lord, 1968). Using CTT, researchers began with the precept that traits or abilities can be known absolutely and that the true score represents the pure trait underlying an observed score (Osterlind, 2006). As with all measurement, the key to determining the true score is isolation and definition of the error around a measured score.

While CTT began the tradition of psychometric modeling of error, it has several shortcomings. Perhaps the most significant of these is that CTT does not allow the researcher to separate item characteristics from person characteristics. In CTT, the proportion of individuals who correctly answer the item is used to calculate item difficulty, and the measure of individual ability is dependent upon not only correct responses to items but also on item difficulty. Like Boring's (1923) definition, this logic is clearly circular.

Also, problems associated with test-retest reliability in CTT led researchers to develop parallel forms of tests. Parallel forms involve creating different items for each form, but items across forms are matched on difficulty such that the two tests have the same overall difficulty score. True parallel forms are difficult to create and even harder to verify (Hambleton, Swaminathan, & Rogers, 1991). In light of these drawbacks, a new psychometric procedure, item response theory (IRT), was developed.

IRT followed CTT as a means of measuring individual ability or traits. IRT models can be used to measure an individual's performance on a test as a function of some latent trait. In educational settings, this latent trait is often referred to as ability. IRT is used to describe not only the degree to which an individual possesses this latent trait, but also how the latent trait influences performance on a test (Marcoulides, 1999).

IRT holds several advantages over CTT. The latent framework allows a researcher to explain the relationship between individual ability and performance on a measurement device (Hambleton et al., 1991). Also, IRT models consider multiple facets of item responses. A single parameter model, commonly called the Rasch (1960) model, frames individual ability within the context of difficulty. However, in modern psychometrics both item and examinee characteristics hold importance, and models which allow the simultaneous evaluation of items and examinees are generally favored.

More complex IRT models were developed which predicted ability based not only on difficulty, but also on the power of items to discriminate between examinees with high versus low ability. As the technique advanced, models that allowed for the possibility of accurate guessing by examinees were also developed (Hambleton et al., 1991). These parameters allowed for more rigorous analysis, reducing the amount of noise in the models created by guessing and differing levels of ability.

Although IRT models contributed a great deal to understanding examinee ability and item and test characteristics, they left some questions unanswered. IRT, as a latent trait model, allowed researchers to understand individual ability and item functioning, but ability in the IRT framework represented a single construct rather than the underlying cognitive processes that make up ability.

Cognitive Diagnostic Assessment

The field of educational measurement is becoming increasingly focused on diagnostic measurements of cognitive ability as a means of large-scale assessment. Leighton and Gierl (2007) point out that this shift in focus requires a new philosophy that relies less on correlational evidence collected after tests have been administered and more on tests as an endeavor of scientific inquiry. This new philosophy is often labeled cognitive diagnostic assessment (CDA). CDA provides information about the cognitive processes underlying examinee performance and, by evaluating response patterns, provides richer information than traditional standardized assessment procedures, which focus on a general ability or response tendency in a single domain (Yang & Embretson, 2007).

Yang and Embretson (2007) outline three primary cognitive characteristics which are the focus of CDA. The first is the identification of skills profiles, which represent the most important skills of a given domain. The second is procedural knowledge and higher-order networks. Finally, a primary focus for CDA is cognitive processes or components. The focus of this work is the application of CDA theory and, more specifically, diagnostic classification models (DCM) to empirical data for determining skills profiles and the appropriate identification of skills on assessments.

While CTT and IRT provide frameworks for measuring ability, both techniques focus on a single measure. Once an assessment has been given and student ability calculated, comparisons of student performance fall along a continuum. Norms may be used to establish cutoffs which divide students into categories which represent levels of proficiency, but it is impossible to know what characteristics separated those students in one category from those in another. This leaves little information about how to improve the standing of those students in lower proficiency categories (Henson & Templin, under review). CDA seeks more informative categorizations, by which the specific skills of each class of examinees can be known, and Diagnostic Classification Modeling (DCM) is the type of latent class modeling by which this is accomplished.

Diagnostic Classification Models

Latent class models are extensions of latent trait models which focus on categorizing examinees into one of *K* latent classes based on their responses to a number of items on a given assessment (Maris, 1999). DCM is a latent class

modeling procedure in which an examinee's responses to dichotomous test items are modeled as a function of the latent class to which the examinee belongs (Templin & Henson, 2006). The use of these models to assess the psychometric properties of both learners and test items is wide-ranging in purpose, specification, and name. The same family of models has also been called cognitive diagnosis models (Templin & Henson, 2006), cognitive psychometric models (Rupp, 2007), multiple classification latent class models (Maris, 1999), and most recently, diagnostic classification models (Rupp & Templin, 2008b). Whatever the label, these models are special cases of latent class models that are used to draw connections between response data collected from participants and the properties of test items. Though DCM can be applied to many research disciplines, the focus here is on the applicability of these models for cognitive diagnostic assessment, where identifying classes of individuals based on skill patterns is useful, especially when a researcher is considering revising testing procedures or instructional strategies. The true advantage of this theoretical framework can only truly be seen when tests are designed using cognitive diagnostic assessment principles; when items on a test are designed with a specific set of skills in mind. Retrofitting these models to previously designed tests is less useful as coverage of each skill by items on the test may not be complete or it may not be possible to accurately identify which skills are reflected in each item.

In DCM, each latent class is defined as a set of examinees sharing a particular attribute pattern or profile. Each latent class represents a unique pattern of skills which an examinee has either mastered or not. Those examinees with the skills to

correctly answer an item are defined as 'masters' while those lacking a skill or skills are classified as 'non-masters'. The pattern of mastery and non-mastery for an examinee across all test-relevant skills determines his or her attribute profile. For a test with K relevant attributes, an attribute profile can be expressed as a K-dimensional row vector, $\boldsymbol{\alpha}_i$, for each examinee, i, comprising binary entries. An entry of '1' in position k indicates mastery of the k^{th} attribute in the profile and an entry of '0' indicates non-mastery. The number of latent classes, then, is 2^k. For example, if only two skills are measured, there are four latent classes: an examinee has mastered neither skill (*00*), the examinee has mastered only the first skill (*10*), the examinee has mastered only the second skill (*01*), and the examinee has mastered both skills (*11*). An individual's attribute profile, then, is one of these possible response patterns.

DCM differs from traditional IRT modeling in that the examinee characteristics which are being modeled are multiple, dichotomous skills rather than single, continuous latent traits. More specifically, whereas IRT models use a single ability measure to model an examinee's response, DCM uses a pattern of skills, which an examinee has either mastered or not, to model the probability of a correct response to a given item. The general goal of DCM is to use item responses to identify the correct latent class, and thereby the set of attributes, of each examinee.

Rupp and Templin (2008b) provide a definition of DCM which outlines eight criteria distinctly characterizing these models:

> Diagnostic classification models (DCM) are probabilistic, confirmatory, multidimensional latent-variable models with a simple or complex loading structure. They are suitable for modelling observable categorical response variables and contain unobservable (i.e., latent) categorical predictor variables.

> The predictor variables are combined in compensatory and noncompensatory ways to generate latent classes. DCM enable multiple criterion-referenced interpretations and associated feedback for diagnostic purposes, which is typically provided at a relatively fine-grain size. This feedback can be, but does not have to be, based on a theory of response processing grounded in applied cognitive psychology. Some DCM are further able to handle complex sampling designs for items and respondents, as well as heterogeneity due to strategy use (p. 226).

The first two criteria refer specifically to the nature of the models. DCM are multidimensional and confirmatory. A second set of criteria define how the models are specified: the complexity of their loading structures, the types of observed and latent variables that are used, and the lack of interactions among latent predictor variables. Finally, several criteria define how these models are used. They allow criterion-referenced interpretations that are diagnostic in nature, and they are flexible enough to model heterogeneity.

DCM are multidimensional because they include latent predictor variables, the number of which is determined by the number of skills assessed. Furthermore, hierarchical diagnostic classification models have been proposed (de la Torre & Douglas, 2004). The confirmatory nature of DCM arises out of the use of the Q-matrix as a loading structure. The 0's and 1's contained in the Q-matrix specify a hypothesis about response patterns. The Q-matrix is to DCM what the factor loading matrix is to factor analysis (Rupp & Templin, 2008b). A priori specification of this matrix, therefore, is a testable hypothesis about the structure of the model, making it a confirmatory procedure.

Unlike FA procedures that often have simple loading structures, DCM models specific, discrete latent variables which are the product of multiple component skills

and require more complex loading structures. Both observed response variables and latent predictor variables are categorical in nature, as opposed to the more familiar continuous framework of FA and IRT. Also, unlike FA and IRT, DCM does allow for interactions among latent variables (Rupp & Templin, 2008b). In fact, DCMs estimate all possible interactions between latent variables.

DCM also has the appeal of modeling many types of heterogeneity. These models are not restricted to modeling individual response patterns, but can be used to investigate between-subject differences in response patterns, as well as within-subject response patterns for individuals across tasks, and several other research questions. This wide-ranging applicability is due in part to the number of DCM variants that can be specified. The models can be either conjunctive or disjunctive. In conjunctive models, all attributes that are required must be mastered by an examinee in order to provide a correct response. Disjunctive models, however, require that only one of the necessary attributes for a given item be mastered in order for an examinee to provide a correct response.

Two specific models are considered for the current work: the DINA model and the DINO model. The DINA (Deterministic Inputs, Noisy "And" Gates) Model (Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977) is a model that predicts the probability of a correct response on an item given only the skill or skills determined to be necessary for that item. If the examinee possesses the skill necessary to answer the item, the probability of a correct response is high, but if the examinee does not possess the skill necessary to answer the item, the probability of a

correct response is low. The model is non-compensatory, which means the lack of a necessary skill cannot be compensated for by the possession of another skill.

The DINA model is conjunctive model and models the response of examinee *i* to item *j* through a dichotomous random variable, X_{ij}. A binary indicator of skill ξ_{ij} is first established for each examinee for each skill *k* where q_{jk} is the Q-matrix entry for item *j* on skill *k* and α_{ik} is the attribute vector for examinee *i*:

$$\xi_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$$

(2)

If a skill is not required for a given item, then $q_{ik} = 0$ and $\alpha_{ik} = 1$ regardless of the individual's attribute pattern; however if a particular skill for that item is required, then $q_{ik} = 1$, and the individual's mastery (or non-mastery) of that skill is now relevant. The product terms dictates the conjunctive property of the model. All necessary skills must mastered ($\alpha_{ik} = 1$) in order for $\xi_{ij} = 1$. If the examinee is missing even one relevant attribute $\xi_{ij} = 0$.

Much like IRT models, DCM accounts for the possibility that an examinee may correctly guess an item response. Therefore, included in the model is the probability even when $\xi_{ij} = 0$ an examinees gives a correct response. This is referred to as the guess parameter, g_j. In addition to the guess parameter, these models also consider that an examinee who has mastered all the relevant skills may 'slip' and answer incorrectly. Therefore, an additional parameter estimates the probability that even when $\xi_{ij} = 1$ an examinee fails to give a correct response. This is the slip parameter,

s_j. In the case that the examinee does possess all of the necessary skills, the probability that a correct response is given is one minus the "slip" parameter. In the case that an examinee does not possess the appropriate skills, the probability of a correct response is given as zero plus the "guess" parameter. Given both the slip and guess parameters, the conditional probability of examinee *i* responding correctly to item *j* is the following:

$$P(X_{ij} = 1 \left| \xi_{ij} \right.) = (1 - s_j)^{\xi_{ij}} g_j^{(1-\xi_{ij})}$$

(3)

Because many of the reading comprehension skills employed by examinees can be compensated for, or overlap with, other strategy options which the examinee may or may not possess, the conjunctive nature of the DINA model is somewhat limiting. This has particular consequences for the guess parameter which may be inflated because an examinee could lack the specified strategy for an item, but compensate by using another strategy. This compensation is not accounted for by the model and the difference between using a compensating strategy and simply guessing would be indistinguishable.

In contrast, the DINO (Deterministic Inputs, Noisy "Or" Gates) Model (Templin & Henson, 2006) does allow for compensating skills. For the DINO model, the binary indicator of skill ω_{ij} is defined as follows:

$$\omega_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}}$$

(4)

Because the model is compensatory $\omega_{ij} = 1$ if an examinee has mastered as least one relevant skill and $\omega_{ij} = 0$ only if none of the relevant skills are mastered. Here the slip parameter is interpreted as the probability that an examinee with mastery of at least one attributes fails to correctly respond to an item. In contrast to the DINA, where lack of mastery on only one required attribute defines the guess parameter, for the DINO the guess parameter is the probability that an examinee who is lacking all relevant attributes correctly responds to the item. Therefore, the conditional probability that an examinee correctly responds to an item is given as follows:

$$P(X_{ij} = 1 \left| \omega_{ij} \right.) = (1 - s_j)^{\omega_{ij}} g^{(1-\omega_{ij})}$$

(5)

Conditional upon an examinee's latent class membership, the probability of a positive response (i.e., a correct response or item endorsement, depending upon the item type) to a given item is determined by the match between the attributes represented by the examinee's latent class and the attributes thought to be relevant to the item and specified in the Q-matrix.

No matter what type of model is specified, they generally are fit under the assumption that the Q-matrix is correctly specified (Maris, 1999). The Q-matrix is "the core element that determines the quality of the diagnostic feedback for the instrument," and so the cost of misspecifying the Q-matrix can be extreme (Rupp & Templin, 2008a, p. 80). In most DCM models, the Q-matrix is established using

strictly binary values for each item based on whether or not a particular skill is required. Occasionally, determining the necessity of a given skill for each item can be relatively straightforward. For example, the skills required to answer an item on a math test can be easily established by visual examination of the symbols (addition, subtraction, etc.). However, on more subjective tests, such as reading comprehension tests, the underlying skills required by each item may be more difficult to determine.

Henson and Templin (under review) point out that, while constructing the Q-matrix is the most crucial and difficult step in DCM, it is often taken for granted. It is assumed that experts correctly identify exactly the skills needed; no more, no less. However, this assumption may not always be true, and the consequences of violating it can be seen in model parameter estimates, classification rates for examinees, and overall model fit. The appropriateness of the Q-matrix is often overlooked and as a result poorly fitting models due to Q-matrix misspecification cannot be identified as poorly fitting or corrected (de la Torre, 2008).

Because of these concerns, studies are now being conducted to determine the consequences of Q-matrix misspecification across a range of conditions. The impact of Q-matrix misspecification will be different for a conjunctive model than it is for a disjunctive model. Rupp and Templin (2008a) conducted a simulation study to examine the consequences of Q-matrix misspecification for the DINA model. The authors generated a known Q-matrix and then created several modified Q-matrices by deleting or adding extra items. Comparisons of models using the true Q-matrix were then compared to models using the modified Q-matrices. Their results showed

that when an extra skill is required, the slip parameter will be inflated while the guess parameter remains somewhat unaffected. However, when a required skill is omitted from the Q-matrix the guess parameter will be overestimated while the slip parameter remains somewhat unaffected (Rupp & Templin, 2008a). Items on an assessment, like individuals, have skill patterns or combinations. Each item may require from one to all of the skills in any combination. In order for individuals to be accurately classified, it is important that their skill pattern be reflected in at least one item on an assessment. In their simulation study, Rupp and Templin (2008a) also found that when an individual's skill pattern was not represented by any item on an assessment that individual was completely misclassified.

Chapter 3

Current Study Methods

Although some work has been done to determine the consequences of Q-matrix misspecification, less has been done to investigate methods which may improve the quality of Q-matrix construction. The purpose of the current study is to explore a probabilistic estimation procedure for Q-matrix construction.

Two key sources of data were utilized: expert opinions from which the Q-matrices were derived and participant data. Six experts were consulted to evaluate the possible skills that could be used on each item. The binary attributes in a cognitive diagnosis model represent the skills required to provide a correct response to each item. The participant data for the study come from literacy tests which were given to a set of examinees in adult basic education programs. Three literacy tests, the Comprehensive Adult Student Assessment System (CASAS), the National Assessment of Educational Progress (NAEP), and the Graduate Equivalency Degree (GED), were examined.

Skill Identification in Reading Comprehension

To investigate the skills needed for each item, the experts evaluated items on two tests on the basis of the type of reading strategy required to correctly answer each item. The experts included three Ph.D. scientists and one M.A. scientist, all specializing in learning disabilities. Additionally, one Ph.D. and one M.A. scientist specializing in reading were consulted. The reading specialists had a range of 8 to 24 years of teaching experience in grades K-12. Each of the experts had knowledge of

standardized achievement measures and assessment, reading instruction practices and patterns, reading skills, and strategy instruction.

The experts were trained in strategy identification using additional forms of the tests examined in this study. The panel worked through one practice test together to clearly define and identify the six skills. In identifying relevant skills for each item, the experts were asked to identify which single skill was most likely necessary for an examinee to correctly respond to each item. This forced choice framework limits somewhat the information that was ultimately used to construct the Q-matrices for each test. Each researcher then worked independently on a second practice form and the definitions were re-examined. From these practice sessions, the definitions of the skills were laid out.

In some instances, skills measured by assessments are relatively easy to identify. Consider that a student is given the following set of problems:

$$17 + 36 =$$

$$9 + 42 - 10 =$$

$$(58 + 103) + 14 =$$

If the student is able to successfully answer the entire set, it is easy to arrive at the conclusion that the student possesses the ability to add and possibly subtract. However, with more complex problems, such as reading comprehension, determining what skill is needed to correctly answer an item is much more difficult.

Reading is a complex task, and it is difficult to observe the underlying skills an examinee is applying when answering comprehension questions.

Using strategies in reading requires several simultaneous processes. Assuming an examinee utilizes multiple strategies, one correct strategy must be first selected, applied, and monitored (Wixson & Lipson, 1991). Given changes in text structure, context, or even difficulty of the passage, the appropriate strategy may change across, or even within, passages on a test. Additionally, it is not always the case that there is only a single appropriate strategy. A student asked to answer the math problems above could not substitute subtraction for addition and still arrive at the correct answer. However, a student asked to read and comprehend a passage of text may implement a strategy such as 'Determining the Main Idea', 'Summarizing' the passage, self-questioning, or any combination of these and arrive at a correct response.

The panel of experts was formed to determine which reading skills would be necessary to correctly answer each item. Specifically, the skills examined here are 'Determining the Main Idea', 'Summarizing', 'Drawing Inferences', 'Generating Questions', 'Creating Visual Images', and 'Looking for Clues' (Hock and Mellard, 2005). Relevant skills were determined based on several criteria: previous research on literacy, definitions which were established, and the experts' practice rating the items.

'Determining the Main Idea' is defined as selecting, deleting, condensing, and paraphrasing information to uncover what the author thinks is most important in the

paragraph. ‘Summarizing’ involves many of the same processes (selecting, deleting, and condensing) but for an entire passage rather than for a single paragraph. If the drawing inference strategy is required a reader needs to draw on prior knowledge, fill in details, and elaborate on what had been read to answer an item relating to a given selection. The ‘Generating Questions’ strategy is used when a reader begins to ask questions about setting, plot, characters, cause and effect, problem solutions, and other critical thinking requirements. ‘Creating Visual Images’ is most commonly used with very symbolic or highly descriptive selections. A reader transfers the words or clues into a mental image which can then be retained to help answer the questions. ‘Looking for Clues’ requires a student to examine a picture or caption, visually inspect a graph, or pull clues from the text such at titles, authors, headings, and key words. A reader may possess any combination of these skills, so correctly answering an item is a function of matching the strategy essential for a correct response to the item.

Participants

All subjects were selected from 14 adult basic education (ABE) programs in Kansas and Missouri. Participants in the study must have met the qualifications for enrollment in an adult education program. Qualifications included withdrawal from secondary education, age of 16 years or older, standardized tests scores below the maximum possible, and US citizenship or foreign national status.

The U.S. Department of Education’s National Reporting System (NRS) functional reading levels derived from the Comprehensive Adult Student Assessment

(CASAS) diagnostic test scores (CASAS, 2002) were used to categorize participants in the study. The NRS uses six adult educational levels where level one is Adult Basic Education (ABE) Beginning Literacy, level two is ABE Beginning Basic, level three is Low Intermediate ABE, level four is High Intermediate ABE, level five is Low Adult Secondary Education (ASE) and level six is High ASE. Participants in each of the six functional reading levels were pooled and a random sample was taken at each level from 713 willing participants recruited from ABE and ASE programs. A goal of sixty participants was set for each level, but due to the low incidence of level one and level two participants in the population, fewer participants were available for these levels. The final sample size was 312 (29 at level one, 44 at level two, 58 at level three, 61 at level four, 59 at level five, and 61 at level six).

It should also be noted that the reading tests examined here, with the exception of the CASAS, are designed primarily for students along a typical, or normally developing, educational trajectory. The world of adult education is a special population whose skills, presumably, map onto those of their younger counterparts who are on the typical trajectory. However, because the skills are obtained in a different time frame, and perhaps even a different sequence, the skills utilized by this population could differ greatly from the tests' intended population. This fact increases the need for understanding what the tests are measuring and what skills are required to successfully complete the items.

Assessments

The first test evaluated in this study was the Level C Reading Test of the CASAS. The CASAS is used to assess functional literacy in adults, and the Level C Reading Test evaluates examinees at approximately a sixth grade level. It consists of 39 items and requires skills relating to the interpretation of materials commonly encountered at home, at work, and in a community setting, such as charts, graphs, and advertisements. The test is timed; examinees are allowed 45 minutes to complete the items.

The format of the CASAS reading test involves a passage of text, a chart, or a graph followed by several questions pertaining to information provided within the material. A certain level of reading mastery is required to answer each item correctly. Because the items have varying formats, it is assumed that a range of reading skills could be employed in order to form a response to each item. Because of the nature of the CASAS as a test of functional literacy, many of the selections lend themselves to 'Looking for Clues' and 'Drawing Inferences', whereas the NAEP and GED had much greater variability in the utilization of the six skills.

The National Assessment of Education Progress (NAEP) was also examined. A 24-item reading comprehension sub-test was the focus of this analysis. Short paragraphs are provided and the examinee is asked to read the passage and then answer the multiple-choice questions that follow. Unlike the functional nature of the CASAS, the NAEP tests for reading comprehension skills within connected prose. The NAEP is designed for participants at higher skill levels, and therefore was only administered to those participants in Intermediate ABE and ASE levels (NRS levels

3-6). Both the nature of the test and the smaller sample make examination of the NAEP very different from the CASAS.

The Graduate Equivalency Degree (GED) test consists of 20 multiple-choice items drawn from short passages of connected prose. The GED was administered only to NRS levels five and six as it is intended for adult education participants nearing the end of their training. Because adult education programs are designed for adults who did not complete their high school education, the GED is the primary goal for many adults who enroll. Therefore, much of the training within these programs is designed toward passing this exam, so the skills necessary to correctly respond to items on the GED are of primary concern for adult education instruction.

Q-Matrix Estimation

To investigate the structure of Q-matrix, both the DINA and DINO models were fit with two Q-matrices: a deterministic Q-matrix created from expert ratings and a probabilistic Q-matrix created using an MCMC estimation algorithm. First, a traditional (deterministic) Q-matrix was established by aggregating the experts' ratings of skills on items. In order to receive a 1 on any strategy half of the raters had to endorse the use that strategy for the item; otherwise a 0 value was assigned. In most cases only one strategy was endorsed by four or more raters. Also, based on expert ratings, the probability of a strategy being required was estimated. The probability is given as the number of experts endorsing a strategy on an item out of the total number of experts. For the probabilistic Q-matrix, rather than rounding the

aggregated ratings to 0 or 1 this value was used as a prior probability for the estimation procedure.

The difference in format between tests, specifically the CASAS, meant that some skills would not be used on all tests; therefore the deterministic Q-matrices for each of the literacy tests contain different combinations of the possible six skills.

Finally, the models were fit using two estimation procedures: one for the probabilistic Q-matrix models, and another for the deterministic Q-matrix models. First, the parameters for the probabilistic Q-matrices were estimated using a Bayesian algorithm. The process is led by the prior probability distribution derived from the aggregation of expert opinion. At each iteration, each Q-matrix element is estimated as either 0 or 1. The final likelihood for each element of the Q-matrix is determined by the number of iterations in the algorithm where the element was equal to 1. For each test, the estimation was performed using the DINA model parameters and again using the DINO model parameters. Because of the complex nature of these models, a large number of iterations were necessary to ensure stability. Therefore, an MCMC chain length of 50,000 iterations, with a thinning interval, of 20 was used. The first 40,000 iterations were discarded as a burn-in period. Convergence of the MCMC estimation was checked using plots for the item parameters. As the algorithm progresses, each element in the probabilistic Q-matrix may change its value from 0 to 1 or vice versa. The results of this estimation were then used to form a new Q-matrix where elements with a value below 0.5 were set to zero and elements with a value at or above 0.5 were set to one. This algorithm-based Q-matrix is what was

used as input for the DCM. Both the DINA and DINO models were estimated using an EM algorithm and the goodness of fit was checked.

The process for the deterministic Q-matrix was much simpler. The original Q-matrix was used, and the respective DCM parameters were also estimated using an EM algorithm. The goodness of fit of the analysis was also noted.

Chapter 4

Results

For each of the three tests, item means, variances, and correlations were run (see Table 1-3).

Q-matrix Estimation

The final Q-matrices for each model differed based on the level of endorsement of each attribute. Tables 4-13 show the final item by attribute patterns for each model. Table 14 summarizes which attributes were relevant for each model. For the CASAS, experts endorsed only the 'Looking for Clues' strategy above the .5 threshold; therefore, the deterministic Q-matrix consisted of only one attribute. Using the estimation algorithm, the probabilistic Q-matrix for the CASAS under DINA initially had three attributes: 'Drawing Inferences', 'Generating Questions', and 'Looking for Clues'. Using this Q-matrix, however, the model failed to estimate parameters because of an inappropriate information matrix. Because only one skill was endorsed on a small number of items, the Q-matrix did not provide sufficient input information for each of the skill patterns. Therefore, the Q-matrix was collapsed into two attributes, combining the 'Drawing Inferences' and 'Generating Questions' attribute into using 'Looking for Clues' as the second attribute. Under the DINO condition, the initial probabilistic Q-matrix contained two attributes: 'Generating Questions' and 'Looking for Clues'. Again, with this Q-matrix as input, the model estimation failed due to an inappropriate information matrix, as only two

items required ‘Generating Questions’, so the Q-matrix was modified to contain only the ‘Looking for Clues’ attribute.

For the NAEP, the experts endorsed four attributes, and the deterministic Q-matrices included ‘Determining the Main Idea’, ‘Summarizing’, ‘Drawing Inferences’, and ‘Looking for Clues’. The probabilistic Q-matrix under DINA included three attributes: ‘Summarizing’, ‘Drawing Inferences’, and ‘Looking for Clues’. The three-attribute Q-matrix also had a inappropriate information matrix, as the ‘Drawing Inferences’ attribute was only required on two items; therefore, an alternative probabilistic Q-matrix was constructed with two attributes, dropping ‘Drawing Inferences’. Under the DINO condition, the probabilistic Q-matrix contained four attributes: ‘Determining the Main Idea’, ‘Summarizing’, ‘Drawing Inferences’, and ‘Looking for Clues’.

Finally, for the GED test, the experts again endorsed four of the available six skills: ‘Determining the Main Idea’, ‘Summarizing’, ‘Drawing Inferences’, and ‘Looking for Clues’. With the probabilistic estimation, five attributes emerged in the Q-matrices; however, some items had no attributes endorsed at or above the 0.5 cutoff. For each of the probabilistic conditions (DINA and DINO), two Q-matrices were constructed. The first Q-matrix relaxed the 0.5 cutoff to 0.2, which was the largest cutoff value that would allow every item to have at least one required skill. The second Q-matrix retained the strict 0.5 cutoff and contained items for which no skills were required. Under the DINA condition, the relaxed Q-matrix contained five attributes: ‘Determining the Main Idea’, ‘Summarizing’, ‘Drawing Inferences’,

‘Generating Questions’, and ‘Looking for Clues’. The strict Q-matrix contained only 4 attributes, leaving out ‘Generating Questions’. Under the DINO condition, both the relaxed and the strict Q-matrices mirrored the DINA condition.

Given the finding of Rupp and Templin (2008a) that accurate classification of individuals relies heavily on the representation of all possible skill patterns in the items on an assessment, it is also interesting to compare the patterns in the deterministic versus the probabilistic Q-matrices. For the CASAS models, under the DINA condition only one skill was required, so there is only a one possible skill pattern. Under the DINO condition, the deterministic Q-matrix required three skills, and four of the seven possible skill patterns were represented (see Table 5). The probabilistic Q-matrix, like those for the DINA model required only one skill.

For the NAEP models, the deterministic Q-matrix required 4 skills, which allows for 15 possible skill combinations. Of those, 9 were represented by the items (see Table 6). The probabilistic DINA Q-matrix for the NAEP required only 2 skills, and all 3 resulting skill combinations are represented by the items (see Table 7). The probabilistic DINO Q-matrix required the same 4 skills as the deterministic Q-matrix, but only covered 7 of the 15 possible skill combinations (see Table 8). Additionally, because the two Q-matrices use same skills, an item-by-item comparison of skill combinations is possible. A total of 17 of the 24 items have the same skill combinations for both the deterministic and the probabilistic Q-matrix conditions. The seven items with different skill combinations have 2 items where the probabilistic Q-matrix ‘deleted’ a skill, 3 items where the probabilistic Q-matrix

‘added’ a skill, and 2 items for which the skill combination was entirely different. The two skill patterns represented by the deterministic Q-matrix that were missing from the probabilistic Q-matrix are *0010* and *1100*. The impact of missing skill combinations as well as the addition or deletion of skills can be examined through the classification rates and slip and guess parameters.

For the GED models, 4 skills were required by the deterministic Q-matrix, for a total of 15 skill patterns of which 8 were represented (see Table 9). Under the relaxed probabilistic DINA estimation, the Q-matrix required 5 skills for 31 skill combinations (see Table 10). The GED’s 20 items cannot represent all of the possible skill combinations and only 9 were. Under the strict condition, the probabilistic Q-matrix with DINA estimation required 4 skills and 6 of the 15 skill combinations were represented (see Table 11). Five of these combinations matched those from the deterministic Q-matrix. One new combination (*1110*) was added and 3 combinations (*0101*, *0110*, and *1010*) were lost. Additionally, 11 items match skill combinations from deterministic to probabilistic, 7 items dropped skills, six of which resulted in having no required skills, 1 item had added skills, and 1 item had an entirely different attribute pattern.

Under the DINO condition, the relaxed probabilistic Q-matrix required 5 skills and 8 skill combinations were represented (see Table 12). The strict probabilistic Q-matrix required 4 skills and 5 skill combinations were represented (see Table 13). All 5 skill combinations match those in the deterministic Q-matrix; however 3 combinations were lost: *0011*, *0101*, and *1010*. Item-by-item comparisons show that

12 items match those of the deterministic Q-matrix, 7 have skills deleted, 5 of which resulted in having no skills required, and 1 items had an added skill.

Classification

The classification rates for all models showed high proportions of examinees lacking all relevant attributes. This is particularly true for the GED models, which ranged from 64% to 90% of examinees falling into this class. For the NAEP, the classification of individuals with *0000* attribute patterns was higher for the probabilistic DINO Q-matrix (37%) than for the deterministic DINO Q-matrix (30%); however, for the GED the probabilistic Q-matrices reduced the proportion of examinees classified under the *0000* pattern (see Table 15).

The class proportions for the models with a single-attribute Q-matrix matched across conditions; however some disagreement arose between models with two-attribute Q-matrices. Although the CASAS test with a probabilistic DINA estimated Q-matrix required a different first attribute ('Drawing Inferences and Generating Questions') than the NAEP test with a probabilistic DINA estimated Q-matrix ('Summarizing'), both Q-matrices endorsed 'Looking for Clues' as a required strategy. The results of the CASAS model suggest that 96% of examinees were lacking the 'Looking for Clues' attribute, whereas the NAEP model suggests that only 67% of examinees were lacking the skill. Table 17 shows the cross-classification of these individuals. Only the *01* skill pattern is of interest, as the first skill differs between the two tests. Of the 64 examinees identified as possessing only

the ‘Looking for Clues’ attribute by the NAEP model, only two were classified that way by the CASAS model.

For three sets of models, a comparison of classification rates for deterministic versus probabilistic Q-matrix models was possible. Because the true classification proportions cannot be known, concordance rates between the two types of models were examined. The NAEP test with DINO estimation had 100% correspondence between the deterministic and the probabilistic Q-matrix models, despite the fact that two skill combinations present in the deterministic Q-matrix were dropped in the probabilistic Q-matrix. For the GED test, however, the correspondence rates for both DINA and DINO models was much lower. Under the DINA condition, the deterministic and probabilistic Q-matrix models classified examinees the same only 61% of the time, and 25% of the discordant pairs were on attribute patterns that were present in one, but missing in the other Q-matrix condition. For the DINO model, the correspondence rate was only 70%, and almost 9% of the discordant pairs were due to the presence of a skill combination in only one of the two Q-matrix conditions.

Slip and Guess Parameters

Subtracting the deterministic Q-matrix parameters from the probabilistic Q-matrix parameters, general patterns for change can be observed. The high proportions of examinees classified as lacking all relevant skills led to large guessing parameters and often very low slip parameters; however, for many of the models, the probabilistic Q-matrix models reduced the overall magnitude of slip and guess parameters. For the CASAS test, using DINA estimation, the mean change for both

slip (Δ_s = -0.019) and guess (Δ_g = -0.119) parameters was negative, suggesting that overall classifications were more accurate with the probabilistic estimation. The largest change was -0.30 for the slip parameter and -0.22 for the guess parameter. The DINO estimation of the CASAS test was identical for the deterministic and probabilistic Q-matrix conditions because of the single-attribute Q-matrix.

Fitting the DINA model to the NAEP data showed that the probabilistic Q-matrix performed less well than the deterministic Q-matrix model with positive average change for both slip (Δ_s = 0.041) and guess (Δ_g = 0.051) parameters (see Table 23). The largest change was 0.542 for the slip parameter and 0.113 for the guess parameter. The probabilistic model used a two-attribute Q-matrix, and only 5% of examinees were classified as having all the necessary attributes, so the extremely large change in the slip parameter may be a result of this low proportion of masters. The DINO yielded somewhat mixed results for the NAEP data (see Table 24). Although there was only one item with a slip parameter greater than zero, the probabilistic Q-matrix had a lower magnitude of slip for that item (Δ_s = -0.004). Additionally, although some of the guessing parameters were a great deal lower for the probabilistic Q-matrix (Δ_{gi} = -0.342), and the mean change showed lower guessing overall for the deterministic Q-matrix (Δ_g = 0.012). Seven items had different skill combinations in the probabilistic condition. Though, the slip parameters provide little information about the impact of the addition or deletion of skills from items, the guess parameter for the two items for which a skill was deleted, the guess parameter was inflated.

Finally, for the GED, the results clearly indicate improvements in slip and guess parameters with the probabilistic Q-matrix. The extremely low proportion of examinees classified as masters of all specified attributes led to slip parameters equal to zero for all items on the GED test under all model conditions; however, the guess parameters were much larger and on average decreased with the probabilistic Q-matrix models. Under DINA conditions, the average change in guess parameters is low for both the relaxed and strict probabilistic Q-matrix conditions compared to the deterministic Q-matrix but favoring the probabilistic Q-matrix in both cases (Δ_{gr} = -0.013, Δ_{gs} = -0.015; see Table 25). The greatest degree of change for the guess parameter under the relaxed Q-matrix was -0.109, and for the guess parameter under the strict Q-matrix the largest change was -0.130. Nine items have different skill combinations from the deterministic to the probabilistic conditions, and 7 of those items had skills deleted. For these items, the general pattern showed lower guess parameters despite the fact that a missing skill should inflate the guess parameter; however, 6 of the 7 items required no skills.

Under DINO conditions, the average change for the guess parameters was even larger than for the DINA models (Δ_{gr} = -0.039, Δ_{gs} = -0.035; see Table 26). For the DINO model conditions, the largest absolute change for the guess parameter under the relaxed Q-matrix was -0.142, and the largest change for the guess parameter under the strict Q-matrix was -0.126. In the probabilistic Q-matrix, 8 items had different skill combinations from the deterministic condition and 7 of those items had skill deleted. The effect of deleted skills on the guess parameter was mixed.

Sometimes it was lower where for other items it was higher. Again, the majority of items with deleted skills required no skills.

DCM Fit Indices

Table 27 shows the model fit comparisons for each set of models. Because the Q-matrix for each of three CASAS models contained only one attribute the fit indices for these models are identical despite the differing model conditions. Under the DINA condition for the CASAS, however, the probabilistic Q-matrix, which contained two attributes, showed better fit (AIC = 13228.426) than the model in the deterministic Q-matrix condition (AIC = 14025.635). The DINA model for the NAEP also showed better fit with the probabilistic Q-matrix (AIC = 6267.040) compared to the deterministic Q-matrix (AIC = 6616.356). The same is true for the NAEP under DINO model conditions where the probabilistic model AIC = 6577.140 and the deterministic model AIC = 6621.739.

The GED is the only test that showed better model fit under the deterministic Q-matrix condition. For the DINA model, the deterministic model AIC = 3090.565, where the relaxed probabilistic model AIC = 3783.711. The strict probabilistic model fit was close to that of the deterministic model, but was still slightly higher (AIC = 3091.791). The strict probabilistic Q-matrix for GED with the DINO model showed better fit (AIC = 3105.529) than either the deterministic model (AIC = 3230.266) or the relaxed probabilistic model (AIC = 3803.109).

Additionally, most of the models show high entropy[1], suggesting good fit and high classification certainty, with the exception of the GED DINO model using the relaxed probabilistic Q-matrix (Entropy = 0.793).

[1] Entropy is an absolute fit statistic which is a measure of classification uncertainty, where 1.00 means that all individuals have been classified with absolute certainty so that values closer to 1.00 indicate better fitting models.

Chapter 5

Discussion

Overall, there is limited evidence to support the hypothesis that probabilistic estimation methods for Q-matrix construction will yield DCMs with better fit, more stable parameter estimates, and more accurate classification rates. While some models indicate that the probabilisitic estimation yielded all of these advantages, other models produced more ambiguous results. If the data for this study had been derived from tests designed for this purpose, the results might be more conclusive. Since these models were retrofit to existing standardized tests, only limited conclusions can be drawn.

Both the deterministic and the probabilistic Q-matrices tended toward overuse of a single attribute. The 'Looking for Clues' strategy was sometimes the only attribute endorsed, and even when the Q-matrix contained other attributes, 'Looking for Clues' was used for most of the items either alone or in addition to another strategy. Because of this over-endorsement several problems and questions arise. From a theoretical perspective, the overuse of a single strategy suggests that the tests examined do not adequately measure all of the underlying aspects of reading comprehension. The fact that experts failed to identify other skills as necessary for many of the items on these tests suggests that either these measurement tools may be limited in their ability to truly identify reading skills or that the skills underlying reading are poorly defined in this study. This is particularly true of the CASAS, which is expected, based on the nature of the test, to measure functional literacy

rather than a higher-order skill. The over-endorsement of skills could also be due in part to the directions experts were given with regard to rating the items. The experts were rating items in a forced-choice situation, where they were asked to rate which skill was most likely required rather than to identify any skills that might be required. If the experts had been allowed to select one or more skills that might be relevant to each item, many of these issues might have been resolved. Also, more skill patterns may have been represented, which would have implications for classification rates and slip and guess parameters as well.

From a modeling perspective, the over-endorsement of a single item is problematic for model estimation. First, it can cause identification problems which prevent accurate estimation of the model. This is demonstrated by the probabilistic Q-matrices for the CASAS test in both DINA and DINO conditions, and the probabilistic DINA Q-matrix for the NAEP model, which did not provide sufficient input for successful model estimation. Attributes in these Q-matrices had to be dropped or combined in order to estimate the DCM.

Additionally, the over-endorsement of a single attribute can lead to classification problems. Rupp and Templin (2008a) found that when all possible attribute combinations were not represented in the Q-matrix, examinees with those attribute patterns were misclassified. The classification rates may have been distributed more evenly if more skill patterns could have been represented, which may be a direct result of the forced choice framework given to the experts.

None of the Q-matrices examined here contain all the possible attribute patterns, and as a result, the classification rates for all models are suspect. Cross-classification of examinees points to this problem. For the two-attribute Q-matrix models, the NAEP and the CASAS were compared. Although the tests are different, as is the first attribute in the pattern, the pattern of possessing only the second skill, 'Looking for Clues', differs greatly for examinees in both conditions. In fact, the correspondence rate is only 3%. If the skill is universally defined, then examinees who possess the skill for one test should also possess it for another test. It is difficult, however, to conclude that this problem is solely an issue of poor model specification. It could be the result of poor definition of the skill or an inability of examinees to apply a skill in different contexts.

Cross-classification also showed that for those attribute patterns which were present in one Q-matrix but not in another, misclassifications did occur; however this is not the only cause of the misclassification. In the case of the NAEP with DINO estimation, the missing skill combination had no effect on classification rates, as the two models had 100% correspondence. For the GED with the DINA estimation 25% of the misclassification can be attributed to missing skill combinations, and just as Rupp and Templin (2008a) found, no individual with an unrepresented skill combination was classified the same across conditions. Since the true Q-matrix is unknown, it is difficult to determine if this is truly a misclassification, but the lack of correspondence points to the importance of skill combinations for consistent classification of individuals. For the GED under the DINO estimation, only about

9% of the disagreement in classification came as a result of missing skill patterns, but only one individual with a missing skill pattern was classified the same across conditions.

Some of these classifications problems could also be the result of the low class probabilities for many of the attribute profiles. Classification rates showed extremely high proportions of examinees in a single class: lacking all relevant skills. High rates of misclassification as a result of poor Q-matrix specification could cause these lopsided proportions. For the GED, probabilistic Q-matrix estimation lowered the proportion of examinees classified as having no relevant skills, suggesting that the probabilistic Q-matrix may be more appropriate; however, it is difficult to draw a definitive conclusion because the true classification is unknown. The high proportion of examinees lacking all relevant skills could also be indicative of inappropriate tests. The CASAS is the only one of the three tests designed specifically for this population. The NAEP and GED require higher level skills, and were not even administered to the students with lower skill levels. Additionally, the skills and strategies used by struggling adult readers may be different from those implemented by younger, more typically developing readers. Given that the high proportion of students classified as lacking relevant skills for the CASAS as well, however, problems with model specification cannot be ruled out as a cause for the classification rates.

In addition to classification rates, slip and guess parameters were examined to evaluate the quality of deterministic versus probabilistic Q-matrices. For several

models, the slip and guess parameters were reduced on average. Slip and guess parameters both indicate ‘mistakes’ made by examinees based on their skill set. If the Q-matrix is misspecified, the mistake can be attributed to the model rather than to the examinee. The CASAS test with the DINA model, and the GED DINA and DINO models all showed lower slip and guess parameters with the probabilistic Q-matrix, suggesting that it provides more accurate estimation that the deterministic Q-matrix.

Under the strict probabilistic conditions for GED models, even when skills were deleted for items, the guess parameters were lower. This is a little puzzling, as previous research has demonstrated that the omission of required skills inflates the guess parameter. Because the slip parameters were at or near zero for these models, and the true Q-matrix is unknown, it is difficult to say whether the probabilistic matrices omitted a necessary skill or the deterministic Q-matrices required an unnecessary skill. Therefore, the lower guess parameters in the probabilistic condition could indicate a more appropriate Q-matrix, but it is difficult to definitely conclude. Other models, showed lower slip and guess parameters with the deterministic Q-matrix making concrete conclusions about the effectiveness of probabilistic Q-matrix estimation even more difficult.

While classification rates and model parameters can provide feedback about the performance of the model, the fit indices give the most conclusive measure of quality by providing information about the appropriateness of the model for the given data. Table 27 shows the model fit for each set of models. Comparing deterministic Q-matrix models to their probabilistic counterparts, in almost every case the

probabilistic model is the better fitting model. The CASAS test with DINO estimation has equivalent deterministic and probabilistic models because of the single-attribute Q-matrices. Also, the GED test with DINA estimation is the one exception where the deterministic Q-matrix yields better model fit than the probabilistic Q-matrix. The relaxed and strict Q-matrices for the GED test were also examined. The strict models showed better fit in both DINA and DINO conditions, and the entropy for the DINA model was low. This is an important consideration for Q-matrix construction, suggesting that endorsing no skills for some items is better than endorsing skills when the probability of that skill being required for an item is less than 0.5. Overall, the fit indices indicate that the probabilistic estimation of Q-matrices for DCM is useful and can produce better fitting, more appropriate, models with higher levels of certainty in classification rates.

Chapter 6

Conclusions

On the surface, it seems the current study has raised more questions than provided answers. The inconsistency in the findings suggests that probabilistic estimation could be useful in the construction of appropriate Q-matrices for diagnostic classification models, but more work is needed. Several shortcomings in the design of the study could be easily remedied. First, the assessments which were chosen were perhaps not entirely appropriate for the examinees to which they were administered. The tests were designed for 'normally' developing readers, and the unique characteristic of adult readers may change response patterns as well as skill utilization. Additionally, the assessments examined here were not designed under a cognitive diagnostic assessment framework. The purpose of the tests is more general. Tests constructed using this framework and with items designed to represent skills more equitably might lead to improvements in discovering appropriate Q-matrices.

Finally, the forced-choice skill identification by experts limits the usefulness of the probabilistic estimation. If raters were allowed to select all possible skills instead of only the most likely, more skills may be endorsed by raters more often and provide a wider range of required skills in the Q-matrix. Also, experts were asked to simply select absolutely the necessity of a given skill for each item. If experts were allowed to provide the probability that a skill or skills were required for items, the prior probability matrix might be more informative and lead to even more appropriate posterior probabilities from which the Q-matrix can be constructed.

Overall, the results indicate that probabilistic estimation in the construction of Q-matrices could be useful. The importance of a correctly specified Q-matrix cannot be overstated, but is often underestimated. While DCM is an increasingly useful technique, the quality of these models and their ability to provide meaningful diagnostic information relies heavily on the appropriateness of the Q-matrix. Therefore, it is important to pursue methods that provide better information and more accurate Q-matrix construction. Probabilistic estimation procedures are a promising development.

References

Boring, E. (1923). Intelligence as the tests test it. *The New Republic, 6*, 35-37.

Comprehensive Adult Student Assessment System. (2002). *Employability competency system*. San Diego, CA: Foundation for Education Achievement.

de la Torre, J. & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333-353.

De la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*, 343-362.

Haertal, E. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333-352.

Hambleton, R. Swaminathan, H. & Rogers, H. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

Henson, R. & Templin, J. (under review). Implications of Q-matrix misspecification in estimation of the reparameterized unified model.

Hock, M. & Mellard, D. (2005). Reading comprehension strategies for adult literacy outcomes. *Journal of Adolescent and Adult Literacy, 49*, 192-200.

Johnston, P., & Costello, P. (2005). Principles for literacy assessment. *Reading Research Quarterly, 40*, 256–267.

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

Leighton, J. & Gierl, M. (2007). Why cognitive diagnostic assessment? In J. Leighton and M. Gierl, (Eds.), *Cognitive diagnostic assessment for education*. New York, NY: Cambridge University Press.

Lord, F., & Novick, M. (1968). Statistical theories of mental tests. Reading, MA: Addison-Wesley Publishing Company.

Macready, G. & Dayton, C. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*, 99-120.

Marcoulides, G. (1999). Generalizability theory: Picking up where the Rasch IRT model leaves off? In S. Embretson and S. Hershberger, (Eds.), *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ, US: Lawrence Erlbaum Associates, 129-152.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 197-212.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Osterlind, S. (2006). *Modern measurement: Theory, principles, and applications of mental appraisal.* Upper Saddle River, NJ: Pearson/Merrill Prentice Hall.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rupp, A. (2007). The answer is in the question: A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models. *International Journal of Testing, 7*, 95-125.

Rupp, A. & Templin, J. (2008a). The effects of misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*, 78-96.

Rupp, A. & Templin, J. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*, 219-262.

Rupp, A., Templin, J., & Henson R. (2010).Diagnostic measurement: Theory, methods, and applications.

Tatsuoka, K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. Nichols, S. Chipman, and R Brennan (Eds.) *Cognitively diagnostic assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Templin, J. & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287-305.

Thorndike, E. (1918). The nature, purposes, and general methods of measurements of educational products. In G.M. Whipple (Ed.), *Seventeenth yearbook of the national society for the study of education*. Bloomington, IN: Public School Publishing, 16-24.

Wixon, K. & Lipson, M. (1991). Perspectives on reading disability research. In R. Barr, M. Kamil, P. Mosenthtal, and P. Pearson (Eds.), *Handbook of reading research: Volume 2*. NY: Longman, 539-570.

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

Appendix

Table 1: Item Descriptives for CASAS Test

Item	*Mean*	*Variance*	*Item-Total Correlation*
1	0.450	0.500	0.238
2	0.700	0.459	0.260
3	0.500	0.502	0.304
4	0.560	0.498	0.363
5	0.800	0.400	0.356
6	0.640	0.481	0.415
7	0.500	0.502	0.325
8	0.600	0.492	0.519
9	0.560	0.498	0.595
10	0.730	0.445	0.442
11	0.410	0.494	0.242
12	0.500	0.502	0.526
13	0.560	0.498	0.237
14	0.570	0.497	0.400
15	0.480	0.502	0.434
16	0.400	0.492	0.362
17	0.510	0.502	0.145
18	0.450	0.500	0.513
19	0.360	0.481	0.299
20	0.350	0.478	0.349
21	0.380	0.487	0.446
22	0.580	0.495	0.554
23	0.250	0.434	0.231
24	0.680	0.468	0.534
25	0.180	0.385	0.096
26	0.430	0.497	0.476
27	0.320	0.468	0.279
28	0.440	0.498	0.547
29	0.440	0.498	0.521
30	0.280	0.450	0.098
31	0.470	0.501	0.341
32	0.540	0.500	0.358
33	0.530	0.501	0.403
34	0.250	0.434	0.018
35	0.490	0.502	0.558
36	0.300	0.459	0.031
37	0.200	0.400	0.223
38	0.590	0.494	0.174
39	0.450	0.500	0.373

Table 2: Item Descriptives for NAEP

Item	*Mean*	*Variance*	*Item-Total Correlation*
1	0.800	0.397	0.315
2	0.820	0.381	0.282
3	0.710	0.450	0.362
4	0.380	0.488	0.019
5	0.600	0.490	0.326
6	0.660	0.477	0.209
7	0.640	0.483	0.307
8	0.700	0.462	0.439
9	0.400	0.491	0.193
10	0.700	0.462	0.389
11	0.540	0.500	0.489
12	0.220	0.412	0.407
13	0.480	0.501	0.184
14	0.600	0.492	0.369
15	0.320	0.466	0.092
16	0.540	0.499	0.366
17	0.680	0.466	0.410
18	0.700	0.457	0.404
19	0.840	0.363	0.408
20	0.670	0.471	0.255
21	0.660	0.475	0.364
22	0.650	0.478	0.287
23	0.740	0.437	0.365
24	0.580	0.496	0.260

Table 3: Item Descriptives for GED

Item	*Mean*	*Variance*	*Item-Total Correlation*
1	0.330	0.474	0.231
2	0.350	0.480	0.010
3	0.190	0.395	-0.001
4	0.100	0.295	0.325
5	0.500	0.502	0.247
6	0.590	0.494	0.323
7	0.590	0.494	0.080
8	0.510	0.502	0.201
9	0.260	0.439	0.082
10	0.280	0.449	0.158
11	0.380	0.488	0.198
12	0.360	0.483	-0.002
13	0.630	0.486	0.269
14	0.410	0.494	0.139
15	0.710	0.454	0.186
16	0.590	0.494	0.338
17	0.530	0.501	0.175
18	0.170	0.379	0.214
19	0.280	0.449	0.067
20	0.680	0.470	0.212

Table 4: Q-matrix for CASAS Under Deterministic DINA and DINO and Probabilistic DINO Conditions

Attribute	*Determining the Main Idea*	*Summarizing*	*Drawing Inferences*	*Generating Questions*	*Visual Images*	*Looking for Clues*
1	0	0	0	0	0	1
2	0	0	0	1*	0	1
3	0	0	0	0	0	1
4	0	0	0	0	0	1
5	0	0	0	0	0	1
6	0	0	0	0	0	1
7	0	0	0	0	0	1
8	0	0	0	1*	0	1
9	0	0	0	1*	0	1
10	0	0	0	0	0	1
11	0	0	0	0	0	1
12	0	0	0	0	0	1
13	0	0	0	0	0	1
14	0	0	0	0	0	1
15	0	0	0	0	0	1
16	0	0	0	0	0	1
17	0	0	0	0	0	1
18	0	0	0	0	0	1
19	0	0	0	0	0	1
20	0	0	0	0	0	1
21	0	0	0	0	0	1
22	0	0	0	0	0	1
23	0	0	0	0	0	1
24	0	0	0	0	0	1
25	0	0	0	0	0	1
26	0	0	0	0	0	1
27	0	0	0	0	0	1
28	0	0	0	0	0	1
29	0	0	0	0	0	1
30	0	0	0	0	0	1
31	0	0	0	0	0	1
32	0	0	0	0	0	1
33	0	0	0	0	0	1
34	0	0	0	0	0	1
35	0	0	0	0	0	1
36	0	0	0	0	0	1
37	0	0	0	0	0	1
38	0	0	0	0	0	1
39	0	0	0	0	0	1

*Items in probabilistic DINO only. Deleted before estimation.

Table 5: Q-matrix for CASAS Under Probabilistic DINA Condition

Attribute	*Determining the Main Idea*	*Summarizing*	*Drawing Inferences*	*Generating Questions*	*Visual Images*	*Looking for Clues*
1	**0**	**0**	**0**	**0**	**0**	**1**
2	0	0	0	0	0	1
3	0	0	0	0	0	1
4	0	0	0	0	0	1
5	0	0	0	0	0	1
6	0	0	0	0	0	1
7	0	0	0	0	0	1
8	**0**	**0**	**0**	**1**	**0**	**1**
9	0	0	0	1	0	1
10	0	0	0	0	0	1
11	0	0	0	0	0	1
12	**0**	**0**	**1**	**0**	**0**	**0**
13	0	0	0	0	0	1
14	0	0	0	1	0	1
15	0	0	0	0	0	1
16	0	0	0	0	0	1
17	**0**	**0**	**1**	**0**	**0**	**1**
18	0	0	0	0	0	1
19	0	0	0	0	0	1
20	0	0	0	0	0	1
21	0	0	0	0	0	1
22	0	0	0	0	0	1
23	0	0	0	0	0	1
24	0	0	0	0	0	1
25	0	0	0	0	0	1
26	0	0	0	0	0	1
27	0	0	0	0	0	1
28	0	0	0	0	0	1
29	0	0	0	0	0	1
30	0	0	0	0	0	1
31	0	0	1	0	0	1
32	0	0	0	0	0	1
33	0	0	0	0	0	1
34	0	0	0	0	0	1
35	0	0	0	0	0	1
36	0	0	0	0	0	1
37	0	0	0	0	0	1
38	0	0	0	0	0	1
39	0	0	0	0	0	1

Bolded items represent unique skill combinations.

Table 6: Q-matrix for NAEP Under Deterministic Condition

Attribute	*Determining the Main Idea*	*Summarizing*	*Drawing Inferences*	*Generating Questions*	*Visual Images*	*Looking for Clues*
1	**0**	**0**	**0**	**0**	**0**	**1**
2	0	0	0	0	0	1
3	0	0	0	0	0	1
4	**0**	**1**	**0**	**0**	**0**	**0**
5	**1**	**0**	**0**	**0**	**0**	**0**
6	0	0	0	0	0	1
7	**1**	**1**	**0**	**0**	**0**	**0**
8	0	1	0	0	0	0
9	**1**	**0**	**0**	**0**	**0**	**1**
10	**0**	**0**	**1**	**0**	**0**	**0**
11	0	1	0	0	0	0
12	**0**	**0**	**1**	**0**	**0**	**1**
13	0	0	0	0	0	1
14	0	1	0	0	0	0
15	0	1	0	0	0	0
16	0	1	0	0	0	0
17	**0**	**1**	**1**	**0**	**0**	**0**
18	0	0	0	0	0	1
19	**0**	**1**	**0**	**0**	**0**	**1**
20	0	1	0	0	0	0
21	0	1	0	0	0	0
22	0	0	0	0	0	1
23	0	0	0	0	0	1
24	0	0	1	0	0	0

Bolded items represent unique skill combinations.

Table 7: Q-matrix for NAEP Under Probabilistic DINA Condition

Attribute	*Determining the Main Idea*	*Summarizing*	*Drawing Inferences*	*Generating Questions*	*Visual Images*	*Looking for Clues*
1	**0**	**0**	**0**	**0**	**0**	**1**
2	0	0	0	0	0	1
3	0	0	0	0	0	1
4	**0**	**1**	**0**	**0**	**0**	**0**
5	0	1	0	0	0	0
6	0	1	0	0	0	0
7	**0**	**1**	**0**	**0**	**0**	**1**
8	0	1	0	0	0	0
9	0	0	0	0	0	1
10	0	1	0	0	0	0
11	0	1	0	0	0	0
12	0	1	0	0	0	1
13	0	0	0	0	0	1
14	0	1	0	0	0	0
15	0	1	0	0	0	0
16	0	1	0	0	0	0
17	0	1	0	0	0	1
18	0	0	0	0	0	1
19	0	0	0	0	0	1
20	0	1	1*	0	0	0
21	0	1	1*	0	0	0
22	0	0	0	0	0	1
23	0	0	0	0	0	1
24	0	1	0	0	0	0

*Items deleted before estimation

Table 8: Q-matrix for NAEP Under Probabilistic DINO Condition

Attribute	*Determining the Main Idea*	*Summarizing*	*Drawing Inferences*	*Generating Questions*	*Visual Images*	*Looking for Clues*
1	**0**	**0**	**0**	**0**	**0**	**1**
2	0	0	0	0	0	1
3	0	0	0	0	0	1
4	**0**	**1**	**0**	**0**	**0**	**0**
5	**1**	**0**	**0**	**0**	**0**	**0**
6	1	0	0	0	0	0
7	0	0	0	0	0	1
8	0	1	0	0	0	0
9	0	0	0	0	0	1
10	0	1	0	0	0	0
11	0	1	0	0	0	0
12	0	0	0	0	0	1
13	0	0	0	0	0	1
14	0	1	0	0	0	0
15	0	1	0	0	0	0
16	0	1	0	0	0	0
17	0	1	0	0	0	0
18	**1**	**0**	**0**	**0**	**0**	**1**
19	**0**	**1**	**0**	**0**	**0**	**1**
20	0	1	0	0	0	0
21	0	1	0	0	0	0
22	0	0	0	0	0	1
23	**0**	**0**	**1**	**0**	**0**	**1**
24	**0**	**1**	**1**	**0**	**0**	**0**

Bolded items represent unique skill combinations.

Table 9: Q-matrix for GED Under Deterministic Condition

Attribute	*Determining the Main Idea*	*Summarizing*	*Drawing Inferences*	*Generating Questions*	*Visual Images*	*Looking for Clues*
1	**0**	**1**	**0**	**0**	**0**	**1**
2	**1**	**0**	**1**	**0**	**0**	**0**
3	**0**	**0**	**1**	**0**	**0**	**0**
4	**0**	**1**	**0**	**0**	**0**	**0**
5	0	0	1	0	0	0
6	**1**	**0**	**0**	**0**	**0**	**0**
7	**0**	**1**	**1**	**0**	**0**	**0**
8	1	0	0	0	0	0
9	0	1	0	0	0	0
10	1	0	1	0	0	0
11	0	1	0	0	0	0
12	0	0	1	0	0	0
13	0	0	1	0	0	0
14	**0**	**0**	**0**	**0**	**0**	**1**
15	0	0	1	0	0	0
16	0	0	0	0	0	1
17	0	0	0	0	0	1
18	1	0	0	0	0	0
19	**0**	**0**	**1**	**0**	**0**	**1**
20	0	1	0	0	0	0

Bolded items represent unique skill combinations.

Table 10: Q-matrix for GED Under Relaxed Probabilistic DINA Condition

Attribute	*Determining the Main Idea*	*Summarizing*	*Drawing Inferences*	*Generating Questions*	*Visual Images*	*Looking for Clues*
1	**0**	**0**	**0**	**0**	**0**	**1**
2	**1**	**1**	**1**	**0**	**0**	**0**
3	**1**	**1**	**1**	**1**	**0**	**1**
4	1	1	1	0	0	0
5	**0**	**0**	**1**	**0**	**0**	**0**
6	**1**	**0**	**0**	**0**	**0**	**0**
7	**0**	**1**	**1**	**0**	**0**	**0**
8	1	0	0	0	0	0
9	**0**	**1**	**0**	**0**	**0**	**0**
10	**1**	**0**	**1**	**1**	**0**	**1**
11	0	0	1	0	0	0
12	0	0	1	0	0	0
13	0	0	1	0	0	0
14	0	0	0	0	0	1
15	0	0	1	0	0	0
16	0	0	0	0	0	1
17	0	0	0	0	0	1
18	1	0	0	0	0	0
19	**0**	**0**	**1**	**0**	**0**	**1**
20	0	1	0	0	0	0

Bolded items represent unique skill combinations.

Table 11: Q-matrix for GED Under Strict Probabilistic DINA Condition

Attribute	*Determining the Main Idea*	*Summarizing*	*Drawing Inferences*	*Generating Questions*	*Visual Images*	*Looking for Clues*
1	**0**	**0**	**0**	**0**	**0**	**1**
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	**1**	**1**	**1**	**0**	**0**	**0**
5	**0**	**0**	**1**	**0**	**0**	**0**
6	**1**	**0**	**0**	**0**	**0**	**0**
7	0	0	0	0	0	0
8	1	0	0	0	0	0
9	**0**	**1**	**0**	**0**	**0**	**0**
10	0	0	0	0	0	0
11	0	0	1	0	0	0
12	0	0	0	0	0	0
13	0	0	0	0	0	0
14	0	0	0	0	0	1
15	0	0	1	0	0	0
16	0	0	0	0	0	1
17	0	0	0	0	0	1
18	1	0	0	0	0	0
19	**0**	**0**	**1**	**0**	**0**	**1**
20	0	1	0	0	0	0

Bolded items represent unique skill combinations.

Table 12: Q-matrix for GED Under Relaxed Probabilistic DINO Condition

Attribute	*Determining the Main Idea*	*Summarizing*	*Drawing Inferences*	*Generating Questions*	*Visual Images*	*Looking for Clues*
1	**0**	**0**	**0**	**0**	**0**	**1**
2	**1**	**0**	**1**	**0**	**0**	**0**
3	**0**	**0**	**1**	**0**	**0**	**0**
4	**0**	**1**	**0**	**0**	**0**	**0**
5	**0**	**1**	**1**	**0**	**0**	**0**
6	**1**	**0**	**0**	**0**	**0**	**0**
7	0	0	1	0	0	0
8	1	0	0	0	0	0
9	0	1	0	0	0	0
10	**1**	**0**	**1**	**1**	**0**	**1**
11	0	1	0	0	0	0
12	0	0	1	0	0	0
13	**0**	**1**	**1**	**1**	**0**	**1**
14	0	0	0	0	0	1
15	0	0	1	0	0	0
16	0	0	0	0	0	1
17	0	0	0	0	0	1
18	1	0	0	0	0	0
19	0	0	1	0	0	0
20	0	1	0	0	0	0

Bolded items represent unique skill combinations.

Table 13: Q-matrix for GED Under Strict Probabilistic DINO Condition

Attribute	*Determining the Main Idea*	*Summarizing*	*Drawing Inferences*	*Generating Questions*	*Visual Images*	*Looking for Clues*
1	**0**	**0**	**0**	**0**	**0**	**1**
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	**0**	**1**	**0**	**0**	**0**	**0**
5	**0**	**1**	**1**	**0**	**0**	**0**
6	**1**	**0**	**0**	**0**	**0**	**0**
7	**0**	**0**	**1**	**0**	**0**	**0**
8	1	0	0	0	0	0
9	0	1	0	0	0	0
10	0	0	0	0	0	0
11	0	1	0	0	0	0
12	0	0	0	0	0	0
13	0	0	0	0	0	0
14	0	0	0	0	0	1
15	0	0	1	0	0	0
16	0	0	0	0	0	1
17	0	0	0	0	0	1
18	1	0	0	0	0	0
19	0	0	1	0	0	0
20	0	1	0	0	0	0

Bolded items represent unique skill combinations.

Table 14: Q-matrix Attributes by Model

	Attribute	Determining the Main Idea	Summarizing	Drawing Inferences		Generating Questions		Visual Images	Looking for Clues
CASAS	Deterministic								
	DINA Prob			✓	*	✓	*		
	DINO Prob					✓	**		
NAEP	Deterministic								
	DINA Prob			✓	**				
	DINO Prob								
GED	Deterministic								
	DINA Relaxed Prob								
	DINA Strict Prob								
	DINO Relaxed Prob								
	DINO Strict Prob								

*Combined into one strategy before DCM estimation
**Deleted before DCM estimation

Table 15: Class Membership in Proportions for Models with Four Attributes

Model	NAEP DINA Deterministic	NAEP DINO Deterministic	NAEP DINO Probabilistic	GED DINA Deterministic	GED DINA Strict Probabilistic	GED DINO Deterministic	GED DINO Strict Probabilistic
Pattern							
0000	0.399	0.298	0.373	0.717	0.637	0.902	0.690
0001	0.039	0.031	0.035	0.053	0.018	0.035	0.035
0010	0.237	0.386	0.127	0.018	0.089	0.000	0.115
0011	0.022	0.031	0.031	0.018	0.018	0.009	0.009
0100	0.009	0.075	0.013	0.035	0.097	0.009	0.027
0101	0.000	0.000	0.018	0.009	0.009	0.009	0.009
0110	0.018	0.018	0.022	0.009	0.027	0.000	0.009
0111	0.009	0.000	0.000	0.000	0.009	0.009	0.009
1000	0.083	0.039	0.154	0.062	0.035	0.009	0.044
1001	0.018	0.018	0.013	0.027	0.018	0.009	0.018
1010	0.083	0.031	0.101	0.018	0.018	0.000	0.009
1011	0.026	0.031	0.035	0.009	0.009	0.000	0.009
1100	0.018	0.009	0.026	0.009	0.009	0.009	0.009
1101	0.013	0.018	0.013	0.009	0.000	0.000	0.009
1110	0.013	0.004	0.039	0.000	0.009	0.000	0.000
1111	0.013	0.013	0.000	0.009	0.000	0.000	0.000

Table 16: Class Membership in Proportions for Models with One or Two Attributes

Model	CASAS DINA Deterministic	CASAS DINA Probabilistic	CASAS DINO Deterministic	CASAS DINA Probabilistic	NAEP DINA Probabilistic
Pattern					
0	0.944		0.944	0.944	
1	0.056		0.056	0.056	
00		0.585			0.605
01		0.010			0.281
10		0.375			0.061
11		0.030			0.053

Table 17: Cross-Classification Probabilistic DINA Estimation for CASAS by NAEP Models

Pattern	00	01	10	11
00	71	33	6	5
01	1	2	0	0
10	60	27	8	7
11	6	2	0	0

*Row categories represent CASAS classifications while column categories represent NAEP classifications.

Table 18: Cross-Classification Deterministic by Probabilistic Q-Matrix for NAEP with DINO Estimation

	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
0000	68	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0001	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0010	0	0	88	0	0	0	0	0	0	0	0	0	0	0	0	0
0011	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0
0100	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0
0101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0110	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0
0111	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1000	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
1001	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0
1010	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0
1011	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0
1100	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
1101	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0
1110	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
1111	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3

*Row categories represent deterministic Q-matrix classifications while column categories represent probabilistic Q-matrix classifications.

Table 19: Cross-Classification Fixed by Probabilistic Q-Matrix for GED with DINA Estimation

	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
0000	63	0	4	0	9	1	3	1	0	0	0	0	0	0	0	0
0001	5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0010	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0011	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0100	0	0	4	0	0	0	0	0	1	0	0	0	0	0	0	0
0101	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0110	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0111	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1000	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
1001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1110	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1111	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

*Row categories represent deterministic Q-matrix classifications while column categories represent probabilistic Q-matrix classifications.

Table 20: Cross-Classification Fixed by Probabilistic Q-Matrix for GED with DINO Estimation

	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
0000	78	**5**	**13**	0	**3**	0	**1**	0	**3**	0	**1**	0	**1**	0	0	0
0001	0	1	0	0	0	0	0	0	0	**2**	0	**1**	0	0	0	0
0010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0011	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0100	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0
0101	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0
0110	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0111	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
1000	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
1001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1110	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1111	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

*Row categories represent deterministic Q-matrix classifications while column categories represent probabilistic Q-matrix classifications.

Table 21: Slip and Guess Parameters for CASAS Test with DINA Model

Item	Deterministic Q-matrix		Probabilistic Q-matrix	
	Slip	Guess	Slip	Guess
1	0.124	0.372	0.111	0.297
2	0.062	0.684	0.000	0.642
3	0.000	0.490	0.000	0.373
4	0.000	0.573	0.000	0.447
5	0.000	0.855	0.000	0.800
6	0.000	0.698	0.000	0.552
7	0.187	0.465	0.112	0.334
8	0.000	0.658	0.000	0.501
9	0.000	0.642	0.000	0.453
10	0.130	0.738	0.221	0.645
11	0.337	0.497	0.224	0.376
12	0.000	0.551	0.000	0.339
13	0.180	0.533	0.000	0.416
14	0.000	0.611	0.000	0.421
15	0.121	0.544	0.221	0.403
16	0.000	0.410	0.000	0.282
17	0.175	0.509	0.000	0.393
18	0.119	0.441	0.110	0.259
19	0.000	0.326	0.000	0.248
20	0.302	0.413	0.000	0.336
21	0.181	0.366	0.221	0.220
22	0.000	0.625	0.000	0.447
23	0.319	0.245	0.445	0.215
24	0.000	0.712	0.000	0.574
25	0.226	0.190	0.219	0.142
26	0.000	0.434	0.000	0.331
27	0.190	0.312	0.218	0.191
28	0.065	0.416	0.000	0.210
29	0.000	0.392	0.000	0.225
30	0.169	0.288	0.000	0.216
31	0.160	0.514	0.000	0.384
32	0.000	0.490	0.000	0.384
33	0.000	0.476	0.000	0.355
34	0.535	0.280	0.498	0.255
35	0.128	0.553	0.251	0.335
36	0.498	0.335	0.499	0.316
37	0.499	0.213	0.625	0.125
38	0.000	0.612	0.000	0.499
39	0.000	0.442	0.000	0.313

Table 22: Slip and Guess Parameters for CASAS Test with DINO Model

Item	Deterministic Q-matrix		Probabilistic Q-matrix	
	Slip	Guess	Slip	Guess
1	0.124	0.372	0.124	0.372
2	0.062	0.684	0.062	0.684
3	0.000	0.490	0.000	0.490
4	0.000	0.573	0.000	0.573
5	0.000	0.855	0.000	0.855
6	0.000	0.698	0.000	0.698
7	0.187	0.465	0.187	0.465
8	0.000	0.658	0.000	0.658
9	0.000	0.642	0.000	0.642
10	0.130	0.738	0.130	0.738
11	0.337	0.497	0.337	0.497
12	0.000	0.551	0.000	0.551
13	0.180	0.533	0.180	0.533
14	0.000	0.611	0.000	0.611
15	0.121	0.544	0.121	0.544
16	0.000	0.410	0.000	0.410
17	0.175	0.509	0.175	0.509
18	0.119	0.441	0.119	0.441
19	0.000	0.326	0.000	0.326
20	0.302	0.413	0.302	0.413
21	0.181	0.366	0.181	0.366
22	0.000	0.625	0.000	0.625
23	0.319	0.245	0.319	0.245
24	0.000	0.712	0.000	0.712
25	0.226	0.190	0.226	0.190
26	0.000	0.434	0.000	0.434
27	0.190	0.312	0.190	0.312
28	0.065	0.416	0.065	0.416
29	0.000	0.392	0.000	0.392
30	0.169	0.288	0.169	0.288
31	0.160	0.514	0.160	0.514
32	0.000	0.490	0.000	0.490
33	0.000	0.476	0.000	0.476
34	0.535	0.280	0.535	0.280
35	0.128	0.553	0.128	0.553
36	0.498	0.335	0.498	0.335
37	0.499	0.213	0.499	0.213
38	0.000	0.612	0.000	0.612
39	0.000	0.442	0.000	0.442

Table 23: Slip and Guess Parameters for NAEP Test with DINA Model

Item	Deterministic Q-matrix		Probabilistic Q-matrix	
	Slip	Guess	Slip	Guess
1	0.000	0.570	0.000	0.675
2	0.000	0.714	0.000	0.695
3	0.000	0.513	0.000	0.596
4	0.000	0.314	0.542	0.370
5	0.000	0.408	0.000	0.464
6	0.000	0.544	0.000	0.575
7	0.000	0.460	0.000	0.495
8	0.366	0.490	0.085	0.529
9	0.000	0.330	0.000	0.315
10	0.000	0.525	0.082	0.564
11	0.000	0.251	0.000	0.364
12	0.000	0.074	0.000	0.079
13	0.000	0.395	0.000	0.392
14	0.000	0.341	0.000	0.435
15	0.000	0.218	0.490	0.262
16	0.000	0.322	0.159	0.376
17	0.000	0.452	0.000	0.518
18	0.000	0.421	0.000	0.527
19	0.000	0.599	0.000	0.706
20	0.000	0.498	0.000	0.525
21	0.000	0.466	0.000	0.483
22	0.000	0.482	0.000	0.519
23	0.000	0.543	0.000	0.631
24	0.000	0.394	0.000	0.458

Table 24: Slip and Guess Parameters for NAEP Test with DINO Model

Item	Deterministic Q-matrix		Probabilistic Q-matrix	
	Slip	Guess	Slip	Guess
1	0.000	0.517	0.000	0.563
2	0.000	0.642	0.000	0.300
3	0.000	0.409	0.000	0.484
4	0.000	0.303	0.000	0.321
5	0.000	0.385	0.000	0.401
6*	0.000	0.546	0.000	0.556
7*	0.000	0.360	0.000	0.436
8	0.422	0.385	0.333	0.471
9	0.000	0.333	0.000	0.334
10	0.000	0.487	0.000	0.539
11	0.000	0.212	0.000	0.277
12*	0.000	0.047	0.000	0.082
13	0.000	0.382	0.000	0.401
14	0.000	0.334	0.000	0.370
15	0.000	0.203	0.000	0.214
16	0.000	0.266	0.000	0.277
17*	0.000	0.356	0.000	0.400
18*	0.000	0.395	0.000	0.401
19	0.000	0.575	0.000	0.606
20	0.000	0.489	0.000	0.457
21	0.000	0.426	0.000	0.412
22	0.000	0.425	0.000	0.488
23*	0.000	0.508	0.000	0.520
24*	0.000	0.457	0.000	0.428

* Skill combinations differ between deterministic and probabilistic Q-matrices: Items 6 and 7 have entirely different combinations, items 12 and 17 have one skill deleted in the probabilistic Q-matrix, items 18, 23, and 24 have one skill added in the probabilistic Q-matrix

Table 25: Slip and Guess Parameters for GED Test with DINA Model

Item	Deterministic Q-matrix		Relaxed Probabilistic Q-matrix		Strict Probabilistic Q-matrix	
	Slip	Guess	Slip	Guess	Slip	Guess
1*	0.000	0.357	0.000	0.259	0.000	0.269
2*^	0.000	0.374	0.000	0.356	0.000	0.371
3*^	0.000	0.123	0.000	0.197	0.000	0.220
4*	0.000	0.012	0.000	0.037	0.000	0.014
5	0.000	0.382	0.000	0.341	0.000	0.272
6	0.000	0.448	0.000	0.477	0.000	0.424
7*^	0.000	0.567	0.000	0.477	0.000	0.491
8	0.000	0.423	0.000	0.457	0.000	0.425
9	0.000	0.271	0.000	0.162	0.000	0.141
10*^	0.000	0.247	0.000	0.180	0.000	0.198
11*	0.000	0.286	0.000	0.309	0.000	0.285
12*^	0.000	0.382	0.000	0.346	0.000	0.390
13*^	0.000	0.564	0.000	0.563	0.000	0.539
14	0.000	0.289	0.000	0.326	0.000	0.365
15	0.000	0.651	0.000	0.651	0.000	0.615
16	0.000	0.522	0.000	0.475	0.000	0.471
17	0.000	0.476	0.000	0.493	0.000	0.550
18	0.000	0.083	0.000	0.119	0.000	0.134
19	0.000	0.254	0.000	0.214	0.000	0.210
20	0.000	0.590	0.000	0.608	0.000	0.624

* Skill combinations differ between deterministic and strict probabilistic Q-matrices: Item 11 has an entirely different combination, items 1, 2, 3, 7, 10, 12 and 13 have at least one skill deleted in the strict probabilistic Q-matrix, item 4 has 2 skills added in the strict probabilistic Q-matrix.
^ Items for which no skills were required.

Table 26: Slip and Guess Parameters for GED Test with DINO Model

Item	Deterministic Q-matrix		Relaxed Probabilistic Q-matrix		Strict Probabilistic Q-matrix	
	Slip	Guess	Slip	Guess	Slip	Guess
1*	0.000	0.284	0.000	0.221	0.000	0.293
2*^	0.000	0.337	0.000	0.347	0.000	0.386
3*^	0.000	0.157	0.000	0.154	0.000	0.204
4	0.000	0.039	0.000	0.040	0.000	0.013
5*	0.000	0.441	0.000	0.406	0.000	0.315
6	0.000	0.530	0.000	0.485	0.000	0.443
7*	0.000	0.549	0.000	0.515	0.000	0.482
8	0.000	0.471	0.000	0.495	0.000	0.450
9	0.000	0.265	0.000	0.124	0.000	0.172
10*^	0.000	0.226	0.000	0.176	0.000	0.193
11	0.000	0.353	0.000	0.358	0.000	0.284
12*^	0.000	0.402	0.000	0.281	0.000	0.382
13*^	0.000	0.614	0.000	0.472	0.000	0.554
14	0.000	0.357	0.000	0.370	0.000	0.345
15	0.000	0.693	0.000	0.660	0.000	0.630
16	0.000	0.542	0.000	0.461	0.000	0.491
17	0.000	0.485	0.000	0.429	0.000	0.516
18	0.000	0.144	0.000	0.182	0.000	0.123
19	0.000	0.250	0.000	0.257	0.000	0.195
20	0.000	0.646	0.000	0.575	0.000	0.614

***** Skill combinations differ between deterministic and strict probabilistic Q-matrices: Items 1, 2, 3, 7, 10, 12 and 13 have at least one skill deleted in the strict probabilistic Q-matrix, item 5 has one skill added in the strict probabilistic Q-matrix.

^ Items for which no skills were required.

Table 27: Model Fit Comparison

Model			Fit Results			
Test	Model	Q-matrix	AIC	BIC	Adj. BIC	Entropy
CASAS	DINA	Deterministic	14025.635	14318.497	14067.954	0.984
CASAS	*DINA*	*Probabilistic*	*13228.428*	*13817.858*	*13313.601*	
CASAS	DINO	Deterministic	14025.635	14318.497	14067.954	0.984
CASAS	DINO	Probabilistic	14025.635	14318.497	14067.954	0.984
NAEP	DINA	Deterministic	6616.356	7984.665	6720.106	0.934
NAEP	*DINA*	*Probabilistic*	*6267.040*	*6606.545*	*6292.783*	
NAEP	DINO	Deterministic	6621.739	7990.048	6725.489	0.917
NAEP	*DINO*	*Probabilistic*	*6577.140*	*7945.449*	*6680.890*	*0.929*
GED	*DINA*	*Deterministic*	*3090.565*	*4004.240*	*2945.470*	*0.988*
GED	DINA	Probabilistic 1	3783.711	5613.788	3493.086	0.793
GED	DINA	Probabilistic 2	3091.791	4005.465	2946.695	0.981
GED	DINO	Deterministic	3230.266	4143.941	3085.170	0.994
GED	DINO	Probabilistic 1	3803.109	5633.186	3512.484	0.944
GED	*DINO*	*Probabilistic 2*	*3105.529*	*4019.204*	*2960.433*	*0.984*