The Application of a Cognitive Diagnosis Model via an Analysis of a Large-Scale Assessment and a Computerized Adaptive Testing Administration

by

Meghan Kathleen McGlohen, B.S., M. A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

In Partial Fulfillment

Of the Requirements

For the Degree of

Doctor of Philosophy

The University of Texas Austin

May 2004

The Application of a Cognitive Diagnosis Model via an Analysis of a Large-Scale Assessment and a Computerized Adaptive Testing Administration

Publication No. _____

Meghan Kathleen McGlohen, Ph.D.
The University of Texas at Austin, 2004

Supervisor: Hua Hua Chang

Our society currently relies heavily on test scores to measure individual progress, but typical scores can only provide a limited amount of information. For instance, a test score does not reveal which of the assessed topics were mastered and which were not well understood. According to the U.S. government, this is no longer sufficient.

The *No Child Left Behind Act of 2001* calls for diagnostic information to be provided for each individual student, along with information for the parents, teachers, and principals to use in addressing individual student needs. This opens the door for a new

area of psychometrics that focuses on the inclusion of diagnostic feedback in traditional standardized testing. This diagnostic assessment could even be combined with techniques already developed in the arena of computer adaptive testing to individualize the assessment process and provide immediate feedback to individual students.

This dissertation is comprised of two major components. First, a cognitive diagnosis-based model, namely the fusion model, is applied to two large-scale mandated tests administered by the Texas Education Agency; and secondly, computer adaptive testing technology is incorporated into the diagnostic assessment process as a way to develop a method of providing interactive assessment and feedback for individual examinees' mastery levels of the cognitive skills of interest. The first part requires attribute assignment of the standardized test items and the simultaneous IRT-based estimation of both the item parameters and the examinee variables under the fusion model. Examinees are classified dichotomously into mastery and non-mastery categories for the assigned attributes. Given this information, it is possible to identify the attributes with which a given student needs additional help. The second part focuses on applying CAT-based methodology, and in particular item selection, to the diagnostic testing process to form a dynamic test that is sensitive to individual response patterns while the examinee is being administered the test. This combination of computer adaptive testing with diagnostic testing will contribute to the research field by enhancing the results that students and their parents and teachers receive from educational measurement.

CHAPTER ONE: INTRODUCTION

Typically, large-scale standardized assessments provide a single summary score to reflect the overall performance level of the examinee in a certain content area. The utility of large-scale standardized assessment would be enhanced if the assessment also provided students and their teachers with useful diagnostic information in addition to the single overall score. Currently, smaller-scale assessments, such as teacher-made tests, are the means of providing such helpful feedback to students throughout the school year. Negligible concern is expressed about the considerable classroom time that is taken by the administration of these formative teacher-made tests because they are viewed as integral parts of instruction. Conversely, educators view standardized testing of any kind as lost instruction time (Linn, 1990). Some advantages of standardized tests over teacher-made tests are that they allow for the comparison of individuals across various educational settings, they are more reliable, and they are objective and equitable (Linn, 1990). The advantage of teacher-made tests, on the other hand, is that they provide very specific information to the students regarding their strengths and weaknesses in the tested material. Large-scale standardized testing would be even more beneficial if it could also contribute to the educational process in a role beyond that of evaluation while maintaining these existing advantages, such as the reporting of diagnostic feedback. Then the students could use this information to target improvement in areas where they are deficient.

A new approach to educational research has begun to effloresce in order to provide the best of both worlds. This research area, dealing with the application of

cognitive diagnosis in the assessment process, aims to provide helpful information to parents, teachers, and students, which can be used to direct additional instruction and study to the areas needed most by the individual student. This beneficial information provided by diagnostic assessment deals with the fundamental elements or building-blocks of the content area. The combination of these elements or attributes comprises the content domain of interest. This form of diagnostic assessment is an appropriate approach to conducting formative assessment, because it provides specific information regarding each measured attribute or content element to every examinee, rather than a single score result.

An ideal assessment would not only be able to meet the meticulous psychometric standards of current large-scale assessments, but would also be able to provide specific diagnostic information regarding the individual examinee's educational needs. In fact, the provision of this additional diagnostic information by large-scale state assessments has recently become a requirement; the *No Child Left Behind Act of 2001* mandates that such feedback be provided to parents, teachers and students.

Despite this requirement, constructing diagnostic assessment from scratch is expensive and impractical. A more affordable solution is to incorporate diagnostic measurement into existing assessments that state and local governments are already administering to public school students. So, in order to incorporate the benefits of diagnostic testing with the current assessment situation, cognitively diagnostic approaches would need to be applied to an existing test.

Diagnostic assessment is a very advantageous approach to measurement. In traditional testing, different students may get the same score for different reasons

(Tatsuoka M. M. and Tatsuoka, K. K., 1989), but in diagnostic testing, these differences can be discovered and shared with the examinee and his/her teacher. Diagnostic assessment allows the testing process to serve an additional instructional purpose in addition to the traditional purposes of assessment (Linn, 1990), and can be used to integrate instruction and assessment (Campione and Brown, 1990). Furthermore, diagnostic testing offers a means of selecting instructional material according to an individual's needs (Embretson, 1990).

While traditional tests can accomplish assessment goals, such as a ranked comparison of examinees or grade assignments based on certain criteria, they do not provide individualized information to teachers or test-takers regarding specific content in the domain of interest (Chipman, Nichols, and Brennan, 1995). Traditional assessment determines what an individual has learned, but not what s/he has the capacity to learn (Embretson, 1990). Diagnostic assessment can be applied to areas involving the identification of individuals who are likely to experience difficulties in a given content domain, and it can help provide specific information regarding the kinds of help an individual needs. Furthermore, the cognitive diagnosis "can be used to gauge an individual's readiness to move on to higher levels of understanding and skill" in the given content domain. (Gott, 1990, p. 174).

Current approaches dealing with cognitive diagnosis focus solely on the estimation of the knowledge state, or attribute vector, of the examinees. This dissertation proposes the combination of the estimation of item response theory (IRT)-based individual ability levels ($\hat{\theta}$) along with an emphasis on the diagnostic feedback provided

by individual attribute vectors ($\hat{\underline{\alpha}}$), thus bridging the current standard in testing technology with a new area of research aimed at helping students benefit from the testing process through diagnostic feedback.

The goal of this research is to not only simultaneously measure individuals' knowledge states and conventional unidimensional IRT ability levels, but to do so in an efficient way. This research will apply the advantages of computerized adaptive testing to the new measurement area of cognitive diagnosis. The goal of computerized adaptive testing is to tailor a test to each individual examinee by allowing the test to hone in on the examinees' ability levels in an interactive manner. Accordingly, examinees are relieved from answering many items that are not representative of their abilities. To accomplish this goal, this research study relies on the technology of computer adaptive testing.

The aim of this research is to combine the advantages of computerized adaptive testing with the helpful feedback provided by cognitively diagnostic assessment, to enhance the existing testing process. This dissertation proposes a customized diagnostic testing procedure that provides both conventional unidimesional ability estimates as well as a report of attribute mastery status to the examinees, instead of just one or the other. The key idea of this work is to utilize the shadow test technique to optimize the estimation of the traditional IRT-based ability level, $\hat{\theta}$, and then select an item from this shadow test that is optimal for the cognitive attribute vector, $\hat{\underline{\alpha}}$, for each examinee. But first, the founding concepts of this research must be elucidated.

CHAPTER TWO: LITERATURE REVIEW

Traditional IRT-Based Testing

Item response theory (IRT) is a common foundation for wide-scale testing. IRT is based on the idea of test homogeneity (Loevinger, 1947) and logistic statistical modeling (Birnbaum, 1968), and uses these probabilistic models to describe the relationship between item response patterns and underlying parameters. IRT uses the item as the unit of measure (rather than the entire test) to obtain ability scores that are on the same scale despite the differences in item administration across examinees (Wainer and Mislevy, 2000). As outlines in Rogers, Swaminathan and Hambleton's (1991) text, two main axioms are assumed when employing IRT:

(1) The performance of an individual on a set of test items can be rationalized by an underlying construct, latent trait, or set thereof. The context of educational testing uses individual ability levels as the trait, which accounts for correct/incorrect response patterns to the test items.

(2) The interconnection between this performance and the intrinsic trait or set of traits can be represented by a monotonically increasing function. That is to say, as the level of ability or trait of interest increases, the probability of a response reflecting this increase (specifically for this context, a correct response) also increases (Rogers, Swaminathan and Hambleton, 1991).

A plethora of possible IRT probability models are available, and a few of the most common will be briefly discussed. Each of these three models maps out a probabilistic

association between the items and the ability level of the examinee (*j*), denoted as θ_j. These three models are respectively referred to as the one-, two-, and three-parameter logistic models. The one-parameter logistic model considers the difficulty level of the items on the test, denoted as b_i for each item *i*. As the name suggests, item with a higher degree of difficult are harder to respond correctly to, and hence call on a higher level of the ability trait θ to do so. The probability of obtaining a correct response to item *i* given a specific ability level θ_j is shown in Equation 1:

$$P_{ij}\left(Y_{ij}=1\middle|\theta_j\right)=\frac{e^{\left(\theta_j-b_i\right)}}{1+e^{\left(\theta_j-b_i\right)}} \tag{1}$$

for $i = 1, 2, \ldots, n$, where *n* is the total number of items, and Y_i denotes the response to item *i* (Rogers, Swaminathan and Hambleton, 1991). The one-parameter logistic model is also referred to as the Rasch model.

Next, the two-parameter logistic model also involves this item difficulty parameter, b_i, but includes another item parameter which deals with the item discrimination, denoted as a_i. Item discrimination reflects an item's facility in *discriminating* between examinees of differing ability levels. The value of the item discrimination parameter is proportional to the slope of the probability function at the location of b_i on the ability axis (Rogers, Swaminathan and Hambleton, 1991). The probability of obtaining a correct response to item *i* given an ability level θ_j, is shown in Equation 2, as described by the two-parameter logistic model:

$$P_i\left(Y_{ij}=1\middle|\theta_j\right)=\frac{e^{Da_i\left(\theta_j-b_i\right)}}{1+e^{Da_i\left(\theta_j-b_i\right)}} \tag{2}$$

for $i = 1, 2, \ldots, n$, and where D is a scaling constant, and n and Y_i hold the same meaning as in the previous model (Rogers, Swaminathan and Hambleton, 1991).

Finally, the three-parameter logistic model (3PL) includes both the difficulty level b_i and discrimination parameter a_i, but also adds a third item parameter, called the pseudo-chance level, denoted as c_i (Rogers, Swaminathan and Hambleton, 1991). The pseudo-chance level allows for the instance where the lower asymptote of the probability function is greater than zero; that is to say, the examinees have some non-zero probability of responding correctly to the item regardless of ability level. The 3PL probability of obtaining a correct response to item i given a θ_j level of ability is shown in Equation 3:

$$P_{ij}\left(Y_{ij} = 1 | \theta_j\right) = c_i + (1 - c_i)\frac{e^{Da_i\left(\theta_j - b_i\right)}}{1 + e^{Da_i\left(\theta_j - b_i\right)}} \tag{3}$$

for $i = 1, 2, \ldots, n$, and all variables are defined as previously noted (Rogers, Swaminathan and Hambleton, 1991). Notice the similarities between these models, and in particular, how each is based on the previous. In fact, the latter two can simplify to the other one(s) by setting c_i to zero (for the two-parameter logistic model) or by setting a_i equal to one and c_i to zero (for the one-parameter logistic model).

These are just a few of the possible model available for describing the relationship between item response patterns and the latent individual ability levels and item parameters. The next portion of this work will discuss a different approach to testing called diagnostic assessment, and will eventually lead to a discussion about how to combine these long-established IRT approaches with new diagnostic assessment techniques.

Diagnostic Assessment Models

The focus of cognitive diagnosis is to provide individual feedback to examinees regarding each of the attributes measured by the assessment. An *attribute* is defined as a "task, subtask, cognitive process, or skill" somehow involved in the measure (Tatsuoka, 1995, p.330). But measuring individuals with respect to the attributes is not the only requirement of a good cognitive diagnosis model. A model must also allow the *items* to be examined in the context of the cognitive diagnosis, or else the results from the assessment cannot be standardized or understood in a larger testing framework (Hartz, 2002). In fact, Hartz, Roussos, and Stout (2002) describe three desirable characteristics of an ideal cognitive diagnosis model as the ability to

(1) Allow the attributes to be appraised with respect to individual examinees,

(2) Allow the relationship between the items and the attributes to be evaluated,

(3) Statistically estimate the parameters involved in the model.

There are at least fourteen models for diagnostic assessment, most of which fall under two major branches. The first deals with models based on Fischer's (1973) Linear Logistic Test Model (LLTM). The second branch of the cognitive diagnosis literature deals with models that employ Tatsuoka's and Tatsuoka's (1982) Rule Space methodology as a foundation. Each model has its strengths and weaknesses. A handful of the models will be described, followed by a justification of the model of choice for this research study.

Fischer's LLTM

One of the first cognitive diagnosis models was the LLTM, developed by Fischer in 1973. The LLTM is an expansion of the basic Rasch model to take into account the cognitive operations required for correct item responses. To do this, the Rasch item difficulty is partitioned into discrete cognitive attribute-based difficulties. Hence, the Rasch item difficulty (denoted as σ_i) equals the weighted sum of these attribute-based difficulties, as shown in Equation (4):

$$\sigma_i = \sum_k f_{ik}\eta_k + c \tag{4}$$

where f_{ik} is the weight of factor k in item i, η_k is the difficulty parameter (or "effect") of factor k across the entire exam, and c is a normalizing constant (Fischer, 1973). The weight of factor k in item i, denoted as f_{ik}, indicates the extent to which factor k is required by item i. (Please note that this concept is analogous to a Q-matrix entry, which is discussed in further detail subsequently.) Substituting this for the item difficulty parameter in the Rasch model yields the LLTM, as presented in Equation (5):

$$P\left(X_{ij}=1\middle|\theta_j\right)=\frac{1}{1+e^{-\left(\theta-\left(\sum_k f_{ik}\eta_k+c\right)\right)}} \tag{5}$$

where X_{ij} equals one when examinee j responds correctly to item i and equals zero otherwise.

The key idea in the LLTM that makes it applicable in the context of diagnostic assessment is that the item difficulty is comprised of the composite of the influences of the basic cognitive operations, or "factors," necessary for correctly solving an item. These operations can be thought of as cognitive building blocks or attributes.

The person parameter, however, remains a single unidimensional ability parameter (θ_j), without any sort of attribute-specific estimate for individual examinees. The LLTM lays the foundation for exploring the cognitively diagnostic relationship between items and their underlying attributes, but it does not incorporate a measure to identify the presence or absence of such attributes in individual examinees.

Rule Space Methodology

The rule space methodology was developed by Kikumi Tatsuoka and her associates (1982), and is comprised of two parts. The first part involves determining the relationship between the items on a test and the attributes that they are measuring. Each examinee may or may not hold a mastery-level understanding of each attribute, and in fact, may hold a mastery-level understanding of any combination thereof. The combinations of attributes which are mastered and not mastered by an individual examinee are depicted in an attribute vector, which is also referred to as a "knowledge state."

The description of which items measure which attributes is illustrated in a Q-matrix. A Q-matrix is sometimes referred to as an incidence matrix, but not in this text. The Q-matrix is a [K x n] matrix of ones and zeros, where K is the number of attributers to be measured and n is the number of items on the exam. For a given element of the Q-matrix in the k^{th} row and the i^{th} column, a one indicates that item i does indeed measure attribute k and a zero indicates it does not. For example, notice the following [3 x 4] Q-matrix:

$$\mathbf{Q} = \begin{array}{c|cccc} & i1 & i2 & i3 & i4 \\ A1 & 0 & 1 & 0 & 1 \\ A2 & 1 & 0 & 0 & 1 \\ A3 & 1 & 0 & 1 & 0 \end{array}$$

The first item measures the second and third attributes, while the second item measures the first attribute only. The third item only measures the third attribute, while the fourth item measures the first two attributes. The consultation of experts in the measured content domain is a good approach to constructing the Q-matrix to determine if an item measures a particular attribute. Other possible approaches to constructing Q-matrices include borrowing from the test blueprint or intuitively evaluating each item to infer which attributes are being assessed. Each element in the Q-matrix is denoted as q_{ik} where the subscripts i and k denote the item and attribute of interest, respectively. This q_{ik} representation mirrors the f_{ik}, parameter in Fischer's LLTM (symbolizing the weight of factor k in item I, as discussed on page 9 of this chapter). They are not exactly the same, however, because Fischer's f_{ik}, weight can take on values greater than unity, while the Q-matrix entry q_{ik} cannot. They are synonymous, however, when all of the f_{ik} weights are dichotomized.

Next, the information provided by the Q-matrix needs to be translated into a form that can be compared with individual observed item response patterns. An observed item response pattern is a vector of ones and zeros that represents how an individual performed on a test. For example, a response pattern of [0 1 0 1 1 1] indicates that the individual responded incorrectly to the first and the third items on the test, but answered each of the other four items correctly. This comparison between the Q-matrix and

individual observed item response patterns is accomplished by establishing a series of ideal response patterns. An ideal response pattern is a response pattern that is obtained through a particular hypothetical combination of mastery and non-mastery levels of the attributes (Tatsuoka, 1995). Notice that the word "ideal" does not indicate a perfect response pattern, but rather "suggests perfect fit with an underlying theory," (Tatsuoka, 1995, p. 339). These ideal response patterns are ascertained systematically by using rules.

A rule is defined as "a description of a set of procedures or operations that one can use in solving a problem in some well-defined procedural domain," (Tatsuoka & Tatsuoka, 1987, p.194). Rules are determined through logical task analysis (Tatsuoka, 1990) and deterministic methods used in artificial intelligence (Tatsuoka and Tatsuoka, 1987). A computer program is used to generate the possible ideal response patterns that would be obtained from the application of a variety of rules. Indubitably, both correct rules and erroneous rules exist for any given assessment. The consistent application of all of the correct rules would result in correct answers for the entire test. The consistent application of some correct rules with other incorrect rules would result in a specific response pattern including both ones and zeros. If a student consistently applies a specific combination of rules to all of the items in a test, then his/her response pattern would match exactly the ideal response pattern produced by the computer program for that exact combination of rules (Tatsuoka, 1995). This provides a finite number of ideal response patterns with which the observed response patterns can be compared.

In order for an examinee to get an item right, s/he must possess all of the attributes that the item measures. This is analogous to electrical circuitry: a current can only flow when all switches are closed. In this analogy, a closed switch symbolizes a mastered attribute and an electrical current represents a correct response to an item. A correct response to an item can only be obtained by an examinee if all attributes involved in a give item are mastered (Tatsuoka and Tatsuoka, 1997; Tatsuoka, 1995). Boolean algebra can be used to explain the cognitive requirements for item response patterns (Birenbaum and Tatsuoka, 1993). Boolean algebra is a mathematical discipline dealing with sets or collections of objects, such as attributes, and is commonly used in dealing with circuits. (For a detailed description of Boolean algebra, please see Whitesitt, 1995). Furthermore, one specific feature of Boolean algebra, called Boolean descriptive functions, can be used to map the relationship between the attributes and item response patterns (Tatsuoka, 1995). Consequently, the various attribute vectors, also known as a knowledge states, can be determined from the set of erroneous rules used to find the ideal response pattern that corresponds with an individual's observed response pattern.

For instance, if an individual taking a math test knows and follows all of the applicable rules correctly except that s/he does not know how to borrow in subtraction, then s/he would get every item right except those that involve borrowing. Thus, his/her attribute vector would include a value of unity for all the attributes except the one(s) that involve borrowing, which would be represented in the attribute vector by a zero. Then diagnostic information would be provided that explains this individual needs more instruction in the area of borrowing in subtraction.

Unfortunately, this scenario does not adequately reflect the reality of a testing situation in its entirety. When an examinee applies an erroneous rule, s/he most likely does not apply the same erroneous rule consistently over the entire test (Birenbaum and Tatsuoka, 1993). Consequently, this inconsistency results in deviations of the observed response pattern from the ideal response pattern. Moreover, careless errors, uncertainty, fatigue, and/or a temporary lapse in thinking can result in even more deviations of an observed response pattern from the ideal response patterns. Tatsuoka (1990) states, "Even if a student possesses some systematic error [i.e. is following an erroneous rule], it is very rare to have the response pattern perfectly matched with the patterns theoretically generated by its algorithm," (p. 462).

In fact, the total number of possible response patterns is an exponential function of the number of items. For n items, 2^n possible response patterns exist. Fortunately, Boolean algebra can be utilized to help reduce the overwhelming number of possible item response patterns to a quantity more computationally manageable.

These random errors made by examinees, referred to as "slips," can be statistically thought of in terms of the probability of a slip occurring or even the probability of having any number of slips occur. Tatsuoka and Tatsuoka (1987) derive a theoretical distribution of the slips, and call it the "bug distribution." The bug distribution calculates the probability of making up to s number of slips by multiplying the probabilities of a slip occurring for all items where one did occur and the probabilities of a slip not occurring for all items where one did not occur, and summing across all s

slips for all possible combinations of *s* slips. This probability distribution for a given rule *R* is presented in Equation (6):

$$\sum_{s=0}^{S}\left[\sum_{\Sigma u_i=s}\prod_{i=1}^{n} p_i^{u_i}(1-p_i)^{u_i}\right] \tag{6}$$

with a mean of $\mu_R = \sum_{i=1}^{t} p_i + \sum_{i=t+1}^{n} q_i$

and a variance of $\sigma_R^2 = \sum_{i=1}^{n} p_i q_i$

where the number of slips, denoted as *s*, ranges from one to *S,* and u_i equals unity when a slip occurs on item *i* and is assigned a zero when it does not. Also, *n* is the number of items and *t* denotes the total score (Tatsuoka and Tatsuoka, 1987; Tatsuoka, 1995).

Conveniently, response patterns that result from inconsistencies in rule applications cluster around the corresponding ideal response patterns; however, the response variance that makes up these clusters complicates classification of the individual examinees because it can no longer be achieved by matching up observed response vectors exactly with the ideal response vectors.

Due to the variability of possible item response patterns, a method is needed for classifying individuals into knowledge states (i.e. identifying their attribute vector) when the response patterns do not reflect exactly an ideal response pattern. To address this, the second part of the rule space methodology deals with the construction of an organized space for classifying the examinees' response patterns into the knowledge state categories that are established in the first part of the methodology.

Once all possible knowledge states are generated from the Q-matrix, item response theory is used to construct this classification space for identifying each of the examinees into one of the predetermined knowledge states. Unfortunately, the attribute variables are unobservable constructs and cannot be represented in such a space directly. Instead, the item response theory functions are used in combination with a new parameter, ζ (zeta), developed by Tatsuoka (1984) to measure the "atypicality" of the response patterns. Values of ζ are calculated as the standardized product of two residual matrices (where residuals are calculated as the difference between an observed and expected value and standardization is achieved by dividing the product by the standard deviation of its distribution).

This parameter ζ can be depicted with the item response theory person parameter θ, which symbolizes ability level, as axes in a Cartesian space in which the knowledge states for each of the ideal response patterns can be graphically represented (Tatsuoka, 1990), as demonstrated by Figure 1.

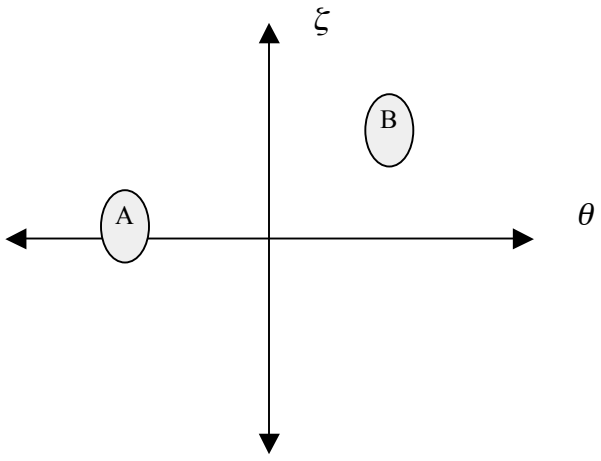

Figure 1: *Two knowledge states in the two-dimensional rule space.*

Notice that knowledge state A is farther left on the θ (ability) scale than knowledge state B. This means that knowledge state B requires a higher level of ability to acquire than knowledge state A. Also notice that knowledge state A is much closer to the θ axis (where the value of ζ, the atypicality index, is zero), which means that this knowledge state occurs more frequently than B. Conversely, knowledge state B is farther away from the θ axis (i.e. the magnitude of ζ is greater), which means that it is a more atypical knowledge state.

Likewise, the various observed response patterns can be represented in terms of each pattern's estimated ζ and θ levels. Furthermore, this Cartesian space can be used to determine which ideal response pattern an observed response pattern is closest to (Birenbaum and Tatsuoka, 1993). Closeness between ideal and observed item response patterns can be approximated by a distance measure, such as the Mahalanobis distance; this metric describes the distance between an observation and the centroid of all observations (Stevens, 1996). Also, Bayes' decision rules may be used to minimize misclassifications (Birenbaum and Tatsuoka, 1993). Bayes' decision rules simply use probabilities to determine which category is most likely when uncertainty is present. The categories are then used to determine an individual's attribute mastery pattern. "Once this classification has been carried out, one can indicate with a specified probability level which attributes a given examinee is likely to have mastered," (Birenbaum and Tatsuoka, 1993, p. 258).

Attribute mastery patterns are represented as a vector of zeros and ones, where a one signifies mastery of a given attribute and zeros signify non-mastery. An attribute

vector, containing *k* (the total number of attributes) elements, is estimated for each individual and denoted as $\underline{\alpha}$. For example, if a test measures three attributes, the estimated attribute vector for a particular examinee may be α_j = [0 1 1], meaning s/he has mastered the second and third attributes, but not the first. Attribute vectors can be used to provide helpful diagnostic information that specifies the area(s) each examinee needs help in. For this example examinee, the diagnostic information would include an indication that the student needs to work on the educational construct corresponding to the first attribute.

The rule space methodology allows for graphical representation of more concepts than just the rule space (as depicted in Figure 1). Other important features of this model can be illustrated visually, such as probability curves, influence diagrams, and knowledge state trees. An influence diagram displays which items measure which attributes and which attributes influence other attributes. In Figure 2, the influence diagram represents the item attribute relationships as described by the Q-matrix as well as the relationship between the attributes.

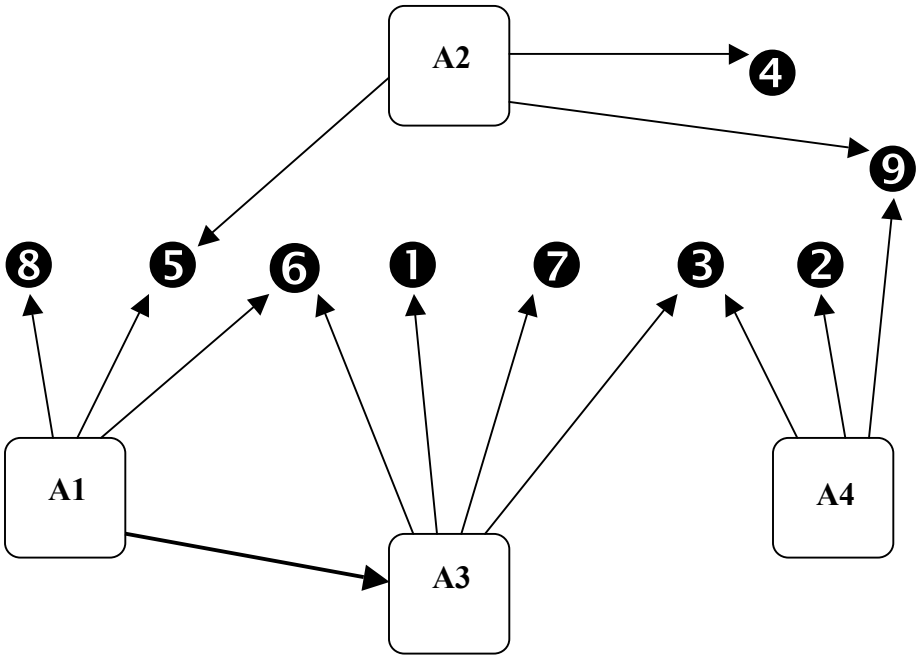

Figure 2: *Nine items measuring four attributes.*

In this figure, an arrow indicates an item measures a certain attribute, with the arrowhead facing the item. In this example, four items, numbers 3, 5, 6, and 9, measure two attributes each, while the remaining items measure one attribute each. Also, the influence of attribute 1 on attribute 3 is represented by an arrow as well. This indicates that attribute 1 is a prerequisite for attribute 3.

The rule space methodology can also be used to construct a knowledge state tree. A knowledge state tree is a very important graphical representation because it portrays the incremental relationships between the knowledge states. This is particularly useful

because it draws out how to improve from one knowledge state to a more mastered knowledge state (Tatsuoka and Tatsuoka, 1997). For example, see Figure 3.

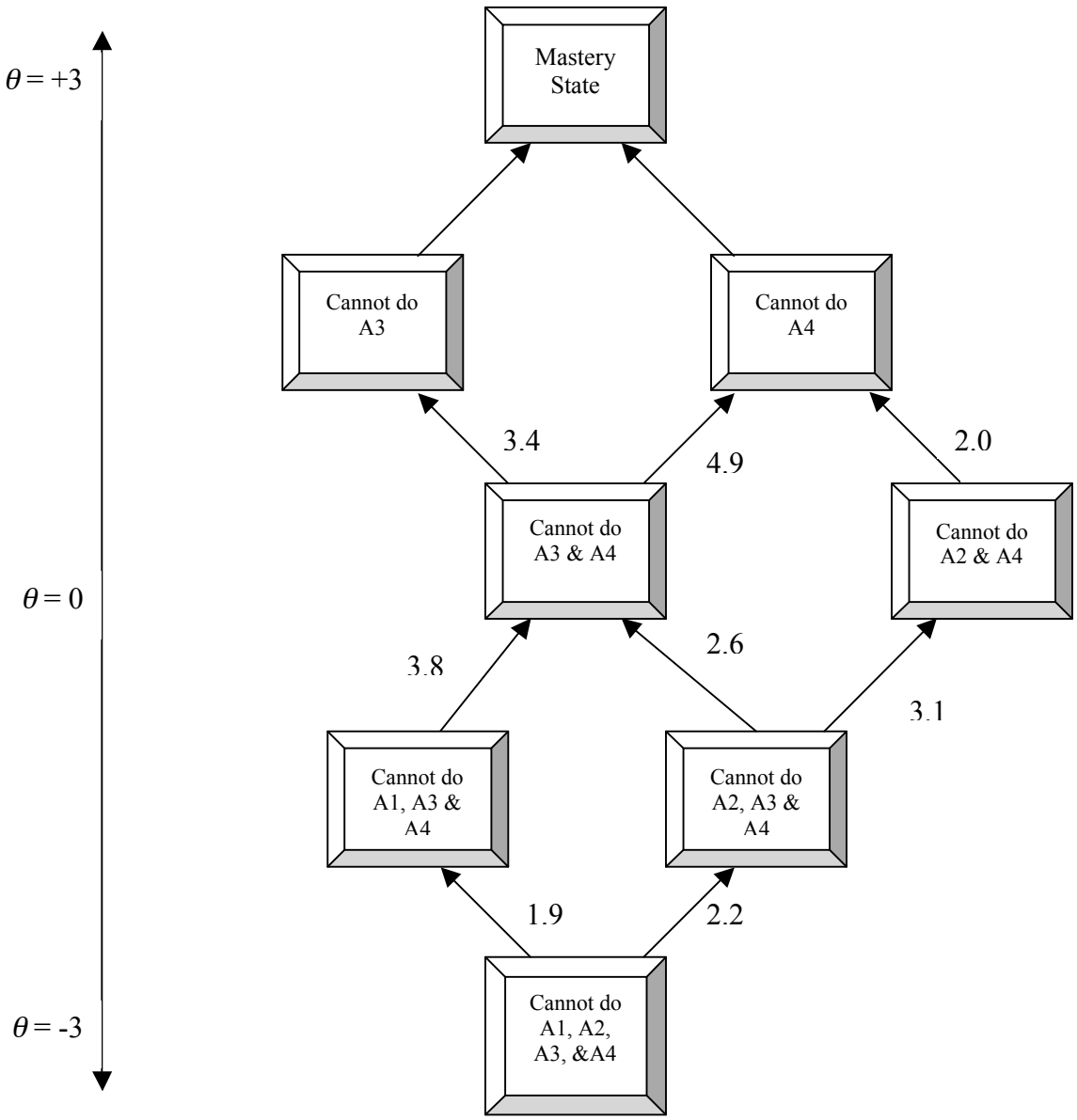

Figure 3: *A tree representing eight knowledge states.*

In traditional testing, higher scores would be assigned to individuals in the knowledge states that are higher in the figure. Knowledge states lower in the figure represent lower traditional scores. Notice the knowledge states "Cannot do A3" and "Cannot do A4" are next to each other. In a traditional testing situation, individuals within both of these knowledge states would receive the same score. Alternatively, the rule space method allowed these examinees to be provided more specific information about their abilities.

Obviously, it is possible for an examinee to lack more than one attribute. This tree is useful because it shows the order in which the non-mastered attributes need to be learned. For instance, if a student cannot do attribute 1, 3, or 4 (i.e. his/her attribute vector is [0, 1, 0, 0]), then the diagram dictates that s/he should next move to the knowledge state "Cannot do A3 & A4." Therefore, the student should learn attribute 1 next. Then there is a choice to move to either state "Cannot do A3" or "cannot do A4." In such a situation, the next appropriate knowledge state is the one with the shortest Mahalanobis distance (Tatsuoka and Tatsuoka, 1997). In this case, the next knowledge state to achieve would be "cannot do A3" with a Mahalanobis distance 3.4; so the student needs to learn attribute 4. Then, once attribute 3 is mastered, the student will accomplish the knowledge state total mastery. Tatsuoka and Tatsuoka (1997) suggest programming this method of determining an appropriate path for instruction in the form of a computer adaptive tutor for remedial instruction. This is revolutionary because it embarks on the combination of the rule space method of diagnostic assessment with computer adaptive

assessment to provide immediate feedback and immediate individualized instruction in the area(s) the examinee needs most.

The rule space methodology is revolutionary and advantageous in many ways, but there is always room for improvement. One drawback is that the rule space methodology does not take into account a way to evaluate the relationships between the items and the attributes.

The Unified Model

DiBello, Stout, and Rousses (1995) based their new model, named the unified model, on the rule space method. They attempted to improve on one of the underlying ideas of the rule space approach. In the unified model, the source of random error is broken down into different four types of systematic error. In the rule space model, all of these would be considered random slips.

They examined the possible sources of random errors and categorized them into four groups. Hence, while there is only one type of random error in the rule space model (slips) there are four sources of aberrant response variance in the unified model. First, they explain that strategy selection is a source of response variation, because an examinee may answer an item using a different strategy than the one assigned in the Q-matrix. Second, completeness of the Q-matrix is considered an important issue. An item may measure an attribute that is not listed in the Q-matrix. For example, a worded math problem includes a verbal component; if the Q-matrix does not contain a verbal attribute, then the Q-matrix would be considered incomplete. Third, the unified model takes

"positivity" into account. Positivity addresses inconsistencies that arise when students who do not possess a certain attribute happen to respond correctly to an item that measures the attribute, or when students who do possess a certain attribute do not apply it correctly and respond erringly to an item measuring the possessed attribute. Positivity takes on a high value when individuals who possess an attribute use it correctly, while students who lack an attribute miss the items that measure it. The less this is the case, the lower the value of positivity. Lastly, a category remains for random errors that are not caused by any of these other three issues. These are called slips and include mental glitches such as finding the correct solution to a problem and then bubbling in a different multiple-choice option. Notice the term "slips" is used more generally for the rule space approach than the unified model.

DiBello, Stout, and Rousses (1995) introduce a new parameter for dealing with the issues of strategy choice and incompleteness of the Q-matrix called the latent residual ability (confusingly, this is denoted as θ_j, but is different than the IRT ability level θ_j). The unified model is the first to include such a parameter. The latent ability space consists of $\underline{\alpha}_Q$, which is the part addressed in the Q-matrix, and α_b, which is the remaining latent ability not included in $\underline{\alpha}_Q$. This parameter θ_j is intended to measure underlying construct α_b, while $\underline{\alpha}_Q$ is measured by the parameter $\underline{\alpha}_j$.

One might ask, why not simply add more attributes to the Q-matrix to account for these issues? More attributes mean more attribute parameters. While additional parameters may allow for enhanced distinctions and alleviate certain classification problems caused by strategy choice and incompleteness of the Q-matrix, these added

parameters would most likely complicate the measurement process more than they would benefit it. More parameters require a greater degree of complexity in the estimation procedure. Also, an increase in the number of attribute parameters to be estimated requires an increase in the number of items on the test to obtain acceptable reliability (DiBello, Stout, and Rousses, 1995). But this may not be practical when so many educators feel test administration is already too long. For the sake of parsimony, including additional parameters is only advantageous when there is a real measurement interest in assessing the mastery/non-mastery of those attributes. Hence, the inclusion of these issues in a model without having to add more attribute parameters is optimal, and this is what DiBello, Stout, and Roussos (1995) have accomplished. The unified model is illustrated in Equation (7):

$$P\left(X_i = 1 \middle| \underline{\alpha}_j, \theta_j\right) = d_i \prod_{k=1}^{K} \pi_{ik}^{\alpha_{jk} \cdot q_{ik}} r_{ik}^{*(1-\alpha_{jk}) \cdot q_{ik}} P_{c_i}(\theta_j) + (1 - d_i) P_{b_i}(\theta_j) \qquad (7)$$

where α_{jk} denotes examinee j's mastery of attribute k, with a one indicating mastery and a zero denoting non-mastery. Also, q_{ik} is the Q-matrix entry for item i and attribute k, and θ_j is the latent residual ability and $P(\theta_j)$ is the Rasch model with the item difficulty parameter specified by the subscript of P. The parameter π_{ik} is the probability that person j will correctly apply attribute k to item i given that person j does indeed posses attribute k; mathematically, this is written as $\pi_{ik} = P(Y_{ijk} = 1 \mid \alpha_{jk} = 1)$ with Y_{ijk} equaling unity when correct application of the attribute is present. Lastly d_i is the probability of selecting the Q-based strategy over all other possible strategies.

The unified model includes a large number of parameters to deal with a plethora of psychometric elements. Having such an amplified set of parameters in the model is both a blessing and a curse. It is a precarious balance that must be met between improving accuracy through the inclusion of more parameters and the issue of statistical identifiability of the many parameters in the model. Jiang (1996) demonstrated that, in fact, the $2K_i+3$ item parameters contained in the unified model are just too many to be uniquely identified.

This model is named the unified model because it uses a deterministic approach to estimating knowledge state classification, and yet it also takes into account random errors. Hence, it unifies both deterministic and stochastic approaches. The unified model is advantageous in that it takes in to account the necessity for assessing examinees with respect to underlying attributes, as well as the requirement for examining the relationship between the items and the attributes rather than just one or the other. Some other advantages of the unified model include the innovative use of the latent residual ability θ_j to help avoid the problems associated with too many latent classes being incorporated into the assessment process (a suitable alternative to the addition of superfluous attribute parameters) and the ability for the model to handle the use of multiple solution strategies by examinees (DiBello, Stout, and Rousses, 1995). In order to apply the unified model in diagnostic assessment, the item parameters must be estimated, which is to say, the model must be calibrated. However, the model lacks practicality in this sense because the parameters involved are not uniquely statistically estimable (Jiang, 1996).

The Fusion Model

The fusion model was based on the unified model (Hartz, Roussos, and Stout, 2002) which, in turn, was based on Tatsuoka's rule space methodology (DiBello, Stout, and Roussos, 1995). The fusion model retains the advantages of the unified model while reducing the number of parameters involved so that they are statistically identifiable. The unified model has $2K_i+3$ parameters for each item, while the fusion model only has $K+1$ (where K is the number of attributes).

Fischer's LLTM does not estimate individuals' mastery level of each attribute. The Rule Space model does not evaluate the relationship between the items and the attributes. The Unified Model satisfies both of these needs, but does not have statistically estimable parameters. The fusion model simplifies the unified model so that the parameters may be estimated. The fusion model was selected as the cognitive analysis model of choice for this study due to the fact that it includes all three features described by Hartz, et al. (2002) as crucial for a successful cognitive diagnosis model, including (1) the estimation of examinees' attribute mastery levels, (2) the ability to relate the items to the attributes, and (3) statistical identifiability of the model's parameters. The item response function for the fusion model is illustrated below in Equation (8), as described by Hartz, et al. (2002):

$$P\left(X_{ij}=1|\underline{\alpha}_j,\theta_j\right)=\pi_i^*\prod_{k=1}^{K}r_{ik}^{*\left(1-\alpha_{jk}\right)\times q_{ik}}P_{c_i}\left(\theta_j\right) \tag{8}$$

where

$P_{c_i}\left(\theta_j\right)$ = The Rasch model with difficulty parameter c_i.

$$\pi_i^* = \prod_{k=1}^{K} P(Y_{ijk} = 1 | \alpha_{jk} = 1)^{q_{ik}}$$

$$r_{ik}^* = \frac{P(Y_{ijk} = 1 | \alpha_{jk} = 0)}{P(Y_{ijk} = 1 | \alpha_{jk} = 1)}$$

Y_{ijk} = 1 when examinee *j* correctly applies attribute *k* to item *i,* and 0 otherwise.

α_{jk} = 1 when examinee *j* has mastered attribute *k*, and 0 otherwise.

c_i = the amount the item response function relies on θ_j after accounting for the attribute assignments in the Q-matrix.

Also, the attribute vector for individual *j* is denoted as $\underline{\alpha}_j$, and θ_j is the residual ability parameter, which deals with content measured by the test that is not included in the Q-matrix.

Parameters

The fusion model introduces many new parameters not present in any other model; therefore a brief explanation of these parameters would be appropriate. First the three item parameters will be described, and then the two person parameters will be explained. Last, the attribute parameter will be discussed.

The parameter π_i^* equals the probability of correctly applying all attributes required by item *i* given that the individual possesses all of the required attributes for the item *i*. As a result, π_i^* reflects the difficulty of item *i* and is referred to as the Q-based item difficulty parameter. It affects a person's capacity to answer the item correctly despite his/her possession of all the involved attributes. Just as with the item difficulty of

classical test theory, the values of π_i^* must remain between zero and unity, and a high value indicates "easiness" rather than "difficulty." Next, the parameter r_{ik}^* represents the proportion of the probability of obtaining a correct response when the examinee does not have the required attribute with the probability of responding correctly to the item when the required attribute is possessed by the examinee. Hence, r_{ik}^* is considered the discrimination parameter of item *i* for attribute *k*. This parameter is described as the penalty for lacking attribute *k*. A high r_{ik}^* value signifies that attribute *k* is not important in producing a correct response to item *i* (Hartz, 2002, Hartz, et al., 2002). This parameter is a ratio of probabilities, and its values also remain between zero and unity. Third, the parameter c_i deals with the item response function's reliance on the residual ability parameter θ_j, which measures the examinee on constructs that are beyond the scope of the Q-matrix. Therefore, the c_i parameter is the completeness index for item *i*. This parameter's value remains between zero and three, with a high value meaning the Q-matrix is complete and therefore a correct response to the item does not rely heavily on θ_j. In sum, a good item would have a high π_i^*, a high c_i and a low r_{ik}^*. Each item has one π_i^* parameter, one c_i parameter, and a r_{ik}^* parameter for each attribute that the given item measures.

The fusion model estimates two types of parameters for the examinees. The first is a vector denoted as $\underline{\alpha}_j$ which identifies the attributes which have been estimated as mastered by examinee *j*. The last parameter specific to the fusion model is the residual

ability parameter θ_j. (Recall that though they are identically denoted, it is not the same as the theta parameter present in typical IRT models. Rather this parameter is comparable to the θ_j parameter in the unified model.) This θ_j parameter measures the "left over" ability construct required by the test but not included in the Q-matrix. This can be thought of as a measure of "higher mental processes" (Samejima, 1995, p. 391), a measure for dealing with multiple solution strategies, or merely a "nuisance parameter," (DiBello, Stout, and Roussos, 1995, p. 370). The inclusion of this parameter is important because it allows us to acknowledge the fact that Q-matrices are not complete representations of what is required by an exam (Hartz, 2002; DiBello, Stout, and Roussos, 1995).

The fusion model also consists of one parameter for each attribute measured in the assessment. This parameter is denoted as p_k and is the cutoff value for attribute k. These values are used to dichotomize the mastery/non-mastery assignments of each examinee, because the values in the attribute vector $\underline{\alpha}_j$ are continuous rather than dichotomous. The continuous values in the attribute vector that are greater than or equal to p_k are assigned the status of mastery for attribute k; the remaining values are assigned the status of non-mastery for that attribute.

Parameter Estimation

Like any model, the fusion model requires an estimation procedure for the parameters involved. The fusion model has several different parameters that must be estimated simultaneously, and therefore requires a powerful procedure for calculating

them. The Markov Chain Monte Carlo (MCMC) method is a versatile estimation procedure that has become an important tool in applied statistics (Tierney, 1997). MCMC is used in parameter estimation in the fusion model because it can elegantly handle the otherwise arduous, perhaps even impossible task of simultaneously estimating the $K+1$ item parameters as well as the K entries in the attribute vector of examinee parameters.

MCMC Estimation. MCMC looks at the probability of the next observation assuming it depends solely on the current observation. Hence, this method is based on a Bayesian framework. The field of Bayesian statistics is based on Bayes' theorem and uses a prior distribution, which involves known information about the parameters of interest, to estimate posterior distribution, which is the "conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data," (Gelman, et al., 1995, p. 3). Bayesian techniques are an advantageous approach to inferential statistical analyses in two important ways. First, the use of prior information utilizes known information about the relationships between the variables involved while the data provides information about unknown characteristics of the variables' relationships. For instance, positive correlations exist between examinee parameters, so it would be appropriate not to waste this information during estimation (Hartz, 2002). A Bayesian framework allows for this sort of flexibility in the prior distribution of parameters (Hartz, 2002). And second, the prior distributions do not untenably influence the posterior distributions (Hartz, 2002). In sum, a Bayesian framework allows the

incorporation of important information already known about the item parameter distributions, but is not adversely affected when such useful information is not available. (For more information regarding Bayesian analytic methods, please see Gelman, et al., 1995.)

To understand how MCMC estimates information about future observations or states, it is easiest to first consider a simple case involving a dichotomous random variable. At any given time, this variable could take on a value of zero or unity. A Markov Chain deals with such a variable over various mathematical states. Consider the probability of achieving a second state given the initial state. This probability can be represented by a matrix with the number of rows and columns matching the number of possible states. When only looking at two states, the transition matrix from state one to state two may be illustrated by the below probability matrix below:

$$p_{s_1 s_2} = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}$$

where

p_{00}= the probability of the variable taking on a value of zero in the first state and zero in the second state.

p_{01}= the probability of the variable taking on a value of zero in the first state and one in the second state.

p_{10}= the probability of the variable taking on a value of one in the first state and zero in the second state.

p_{11} = the probability of the variable taking on a value of one in both states.

Regardless of whether the variable takes on a value of zero or one in either state, this matrix can be used to determine the probability of the transition to state two. A bit of matrix algebra can be used to determine the probability of obtaining a value of one or zero for the second state by multiplying the transition matrix of probabilities and the vector of probabilities of the initial state as illustrated in Equation (9).

$$p_{s_2} = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} \begin{bmatrix} p_{s_1=0} \\ p_{s_1=1} \end{bmatrix} = \begin{bmatrix} p_{00}p_{s_1=0} + p_{01}p_{s_1=1} \\ p_{10}p_{s_1=0} + p_{11}p_{s_1=1} \end{bmatrix} = \begin{bmatrix} p_{s_2=0} \\ p_{s_2=1} \end{bmatrix} \tag{9}$$

Next, consider the probability of observing state 3. The probability of observing a zero or a one for the third state can be similarly determined, as outlined in Equation (10).

$$p_{s_3} = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}^2 \begin{bmatrix} p_{s_1=0} \\ p_{s_1=1} \end{bmatrix} = \begin{bmatrix} (p_{00}{}^2 + p_{01}p_{10})p_{s_1=0} + (p_{00}p_{01} + p_{01}p_{11})p_{s_1=1} \\ (p_{10}p_{00} + p_{11}p_{10})p_{s_1=0} + (p_{10}p_{01} + p_{11}{}^2)p_{s_1=1} \end{bmatrix} = \begin{bmatrix} p_{s_3=0} \\ p_{s_3=1} \end{bmatrix}$$

(10)

Likewise, the probability vector for state 4 or even state 194 and so on can continue to be determined by raising the transition matrix to greater and greater exponents and then multiplying by the vector of probabilities of the initial state. Consequently, as the number of iterations increases, the complexity of the calculation also increases, but more importantly, the reliance on the initial state decreases and the

resulting probability matrix depends more on the transitional matrix (Smith, 2003). As the number of iterations in the chain gets increasingly large, the values in the matrix eventually converges to a single matrix where every row is identical and the values of that row represent the posterior distribution.

This posterior distribution is then used, along with a random number generator, to determine the next state from a current state. A random number between zero and unity is generated from a uniform distribution and this value is compared with the values of the probabilities in the posterior distribution to determine the next state. There is an array of algorithms for this process, and further details are presented in the following section.

The idea here can be extended to variables that can take on polytomous observations, rather than just zeros and ones as illustrated above. MCMC estimation can even be extended to continuous rather than discrete variables. In the continuous case, a "transition kernel" is used instead of a transition matrix (Smith, 2003); a transition kernel is in the form of a function rather than a matrix.

The estimation is referred to as MCMC because the property of a sufficiently long Markov Chain is that the values will converge to a singe matrix. The Monte Carlo element of MCMC estimation describes how the iterations of the chain are randomly selected from the distribution of possible states, which is represented by the rows in the matrix. For more information on MCMC methods and MCMC estimation, please see Smith, 2003.

The Metropolis–Hastings Algorithm. A variety of algorithms for implementing MCMC estimation are available, and the oldest and most widely used is the Metropolis–Hastings (Tierney, 1997). The Metropolis–Hastings (M-H) algorithm was conceived by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller in 1953, and was then generalized by Hastings in 1970 (Chib and Greenberg, 1995). As mentioned earlier, the transitional probabilities are used in combination with a random number generator to determine the mathematical state of the next iteration in the MCMC chain. The M-H algorithm uses the transition kernel (or matrix for the discrete case) to produce a *candidate* for the next state, rather than automatically accepting the state determined by the above mentioned process. The added precaution of examining this state as a candidate is needed because it is possible that there is a propensity to move from one state to another more frequently than vice versa (i.e. the transition from state A to state B occurs more often then from state B to state A), and this lack of "reversibility" must be taken into account (Chib and Greenberg, 1995). So, a comparison is made to reduce the moves from state A to state B to balance out an MCMC process that would otherwise be "lop-sided" and thus not satisfy the required condition of reversibility. This comparative step can be thought of as a filter to always accept moves from state B to state A, but to only accept moves from state A to state B with a certain (non-zero) level of probability (Chib and Greenberg, 1995).

Gibbs Sampling. Gibbs sampling is a special case of the M-H algorithm, which was introduced to the statistical community by Gelfand and Smith in 1990 as an approach

for fitting statistical models (Gelfand, 1997). When estimation is required for multiple parameters, Gibbs sampling helps simplify the process of sampling the candidates by allowing it to focus on one component at a time. In short, the process examines the candidate vector of parameters to be estimated and accepts the candidate with the applicable probability based on the first parameter (say for example the item parameter), then obtains the next candidate and accepts it with the same probability based on the second parameter of interest (say this time it is the examinee parameter). The advantage of Gibbs sampling is that it takes a "divide-and-conquer" approach to estimating the item and examinee parameters simultaneously (Patz and Junker, 1997, p. 7).

This was a brief primer describing the estimation procedure used to calculate the fusion model parameters. For a more comprehensive description of the statistical techniques or backgrounds regarding the use of the M-H algorithm and/or Gibbs Sampling in the context of MCMC estimation, please see Chib and Greenberg, 1995.

Computerized Adaptive Testing

Wide scale assessment must include items able to gauge the abilities of a broad range of examinees, which unfortunately bores the high ability examinees by asking too easy items and frustrates the lower ability examinees with questions that are too hard. Needless to say, this is not an extremely efficient approach to individualized measurement. Computerized adaptive testing (CAT) avoids this issue by honing in on each examinee's ability estimate and asking items that would be informative at his/her level and omitting items that would not be helpful in the estimation process. CAT retains the advantages of group administration while adopting the advantage of individualized assessment (Wainer, 2000). Other advantages of CAT include test security, individual test-takers working proficiently and at their own pace, the absence of physical answer sheets, the possibility of immediate feedback, the ease of administering experimental items, and the variability of item formats available in the computerized interface (Wainer, 2000).

The history of adaptive testing precedes the ubiquity of the personal computer. In the early 1970's new approaches such as flexilevel testing (Lord, 1971a) and multi-stage testing (Lord, 1971b) were being explored as means of directing a test's questions toward the ability level of the individual examinee and avoiding superfluous items. Lord's research in this area is often credited as pioneering the field of adaptive testing (van der Linden and Pashley, 2000). Upon the influx of inexpensive computing power, personal computers became capable of being programmed to administer a test the way an individual examiner would, that is to say, directing the line of questioning towards the

examinees true ability level. CAT has infiltrated the testing arena throughout the country. Large-scale tests that have implemented a computer adaptive version include the Computerized Placement test, the Graduate Record Exam, the Armed Services Vocational Aptitude Battery, as well as licensure exams for individuals in the medical industry (Meijer and Nering, 1999).

The aim of CAT is to construct an optimal test for each examinee (Meijer and Nering, 1999), and as a result, different examinees respond to a different set of items. In order to consistently estimate ability levels across examinees' different administrations of items, CAT employs the field of item response theory (IRT), an area discussed at the beginning of Chapter Two.

CAT algorithms consist of three steps: (1) selecting an initial item, (2) continuing to select all subsequent items, and (3) ending the testing process (Thissen and Mislevy, 2000). Consequently, CAT has a multitude of specialized foci, including item bank development, item selection procedures and ability estimation procedures, which in turn bring up additional issues such as test security and reliability (Meijer and Nering, 1999). The emphasis of this opus, however, is on an approach to optimized item selection.

Item Selection

An important element of CAT administration is determining which items should be presented next given the current estimate of the examinees ability level. The reason item selection is an important process in CAT is because it allows item administration to adaptively correspond to the examinee's ability estimate (Meijer and Nering, 1999).

The two most common selection procedures to determine the best item to administer next are "maximum information" and "maximum expected precision" (Thissen and Mislevy, 2000). The maximum information approach selects an item that maximizes the Fisher information function, presented in Equation (11), for the given ability estimate $\hat{\theta}_j$:

$$I_i\left(\hat{\theta}_j\right) = \frac{[P_i'(\hat{\theta}_j)]^2}{P_i(\hat{\theta}_j)[1 - P_i(\hat{\theta}_j)]} \tag{11}$$

where $P_i\left(\hat{\theta}_j\right)$ is the probability of a correct response by examinee *j* on item *i* given the current ability estimate $\hat{\theta}_j$, and $P_i'\left(\hat{\theta}_j\right)$ is the first derivative thereof. The method of maximum expected precision is a Bayesian approach that seeks to minimize the variance of the posterior distribution. Other possible item selection procedures include Owen's (1975) Bayesian item selection, Veerkamp and Berger's (1997) weighted information criteria, and Davey and Parshall's (1995) posterior weighted information. Other approaches based on variations of the maximum information procedure use Kullback-Leibler (K-L) information rather than Fisher information in item selection (see Chang and Ying, 1996; Eggen, 1999; Xu, et al., 2003). More details regarding K-L information are provided below. Item selection in this research entails the maximum information approach using both Fisher information and K-L information, as well as another information approach, Shannon Entropy.

While utilizing an item selection method that provides the most information about the given examinee is important, item selection does not solely depend on which item is

the best in terms of ability estimation. Theoretically, the next item selected is the one that can best aid in the ability estimation, but in reality, practical considerations must also be taken into account, such as content balancing and items exposure control (Wainer and Mislevy, 2000). In the context of high-stakes CAT, there are three competing goals in a test's construction:

(1) Select items that measure the examinee's ability as quickly, efficiently and accurately as possible.

(2) Ensure each administration of the test measures the same combination of content domains and item formats.

(3) Maintain the test security by controlling the rates of each item's exposure across all administrations (Davey and Parshall, 1995; Stocking and Lewis, 2000).

Generally, satisfying one of these goals imposes a diminution of the other two (Stocking and Lewis, 2000). If items become compromised, they no longer accurately measure all examinees' abilities equitably. As a result, it is appropriate to administer items that are sub-optimal for a given ability estimate when the optimal item happens to be selected so frequently that they may be compromised. Therefore, the precision of measure is sometimes sacrificed for the sake of item security. But in the end, the goal is to obtain an accurate estimate of the examinee's ability with a finite number of items and in a short enough amount of time to minimize the occurrence of fatigue.

Methods for controlling items exposure rates have been developed by Sympson and Hetter (1985), Davey and Parshall (1995), Stocking and Lewis (1998), Chang and

Ying (1999), and Parshall, Harmes and Kromrey (2000), just to name a few. Clearly, the issue of item exposure control is important in the context of CAT. So, accurately approximating individuals' ability levels while maintaining item exposure control as well as content balancing are significant goals in CAT administration. Ergo, an item selection procedure is needed that can simultaneously manage all of such issues. Fortunately, the shadow test approach is capable of handling this challenging endeavor. The shadow test technique can take into account item exposure control along with content balancing and a plethora of other constraints in the item selection process.

The Shadow Test

The idea of shadow testing was proposed by van der Linden and Reese in 1998. Shadow testing is a mode for test assembly that utilizes linear programming (LP) to incorporate constraints into the assembly process. It is an iterative process in which an ideal "shadow" test is formed before the administration of each item in an examination.

A shadow test is a test that is not administered in its entirety to the examinee. Rather, a new shadow test is constructed before the administration of each item, and a whole test had the same number of shadow tests as it has total items. A shadow test must be optimal at the given estimate level while complying with all of the specified constraints, and therefore any item selected from the shadow test will maintain such properties and be a good item for the current estimate level (van der Linden and Chang, 2003). Each shadow test must also contain any items already administered in the overall test. The best item on the shadow test (that has not yet been administered) is then selected as the next item to be administered to the examinee, and unused items are

returned to the pool. The response from this item is then used in the process of formulating the next shadow test. The iterative process of the shadow test for a four-item test is illustrated in Figure 4.

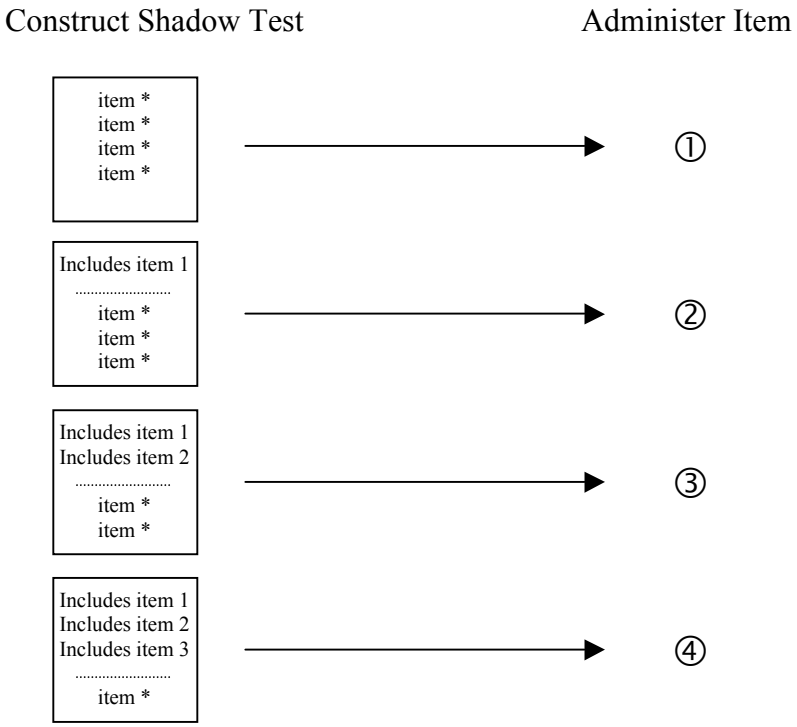

Figure 4: *The iterative process of the shadow test approach.*

The notation "item *" represents an item selected to be in the shadow test that has not already been administered; it is the best item among these that is selected to be the next item given to the examinee. At each step, a full-length shadow test is constructed, each of which preserve three important requirements. First, each shadow test meets all of the constraints; second, it must include all previously administered items; third, it has

maximum information at the current ability estimate (van der Linden, 2000b). "The last shadow test is the actual adaptive test and always meets all constraints," (van der Linden, 2000b, p.33).

The application of this approach results in two major advantages. First, the items actually administered in the adaptive test will certainly follow the constraints because each of the shadow tests meets these specifications. Second, the adaptive test will converge optimally to the true value of the estimator because the shadow tests are assembled to be optimal for the current estimate level, and in turn, each selected item is the optimal one from that shadow test (van der Linden and Chang, 2003).

A shadow test is constructed based on an objective function and a series of constraints. Frequently, the objective function involves maximizing the Fisher information function (van der Linden and Reese, 1998; van der Linden, 2000b; van der Linden and Chang, 2003). With respect to the constraints, test-makers have a wide variety from which to choose (for an extended list, see van der Linden and Reese, 1998, van der Linden, 2000a or van der Linden, 2000b). There are three main types of constraints, including (1) those dealing with categorical item characteristics, (2) those about quantitative features of the items, and (3) those for inter-item dependencies (van der Linden, 2000a; Veldkamp and van der Linden, 2000). The test's administrators may choose any combination of these to include in the list of constraints. Some examples of typical constraints are presented below.

For a fixed length test, the total number of items would be a necessary constraint to include, and for the context of diagnostic assessment, the number of items measuring

each attribute would also be an appropriate constraint. The manual for the Arpeggio software (Stout, et al, 2002) requires each attribute be measured by at least three items, so this would need to be mathematically specified in the list of constraints (Hartz, Roussos and Stout, 2002). Another important constraint deals with content balancing. Each examinee should receive the same number of items in each content area (Green, et al., 1984) and this can be easily taken into account by a mathematical constraint. Item exposure control is another constraint that is beneficial to the test construction process. A constraint dealing with exposure control can be included by merely limiting the frequency of an item's administration (van der Linden and Reese, 1998) or as involved as using the alpha-stratified approach (van der Linden and Chang, 2003). Theoretically, any aspect of a non-adaptive test can be incorporated into the shadow test procedure as long as it can be mathematically represented by a constraint (van der Linden, 2000b). The mathematical constraints involved in this study are listed in Chapter Three. For more information regarding possible constraints, see Stocking and Swanson (1993) or van der Linden and Reese (1998).

In order to formulate a shadow test that is optimal with respect to the objective function and that also obeys the specified constraints, the field of linear programming (specifically, integer programming) must be utilized. A computer software program called CPLEX (ILOG, 2003) is used to solve the linear programming problem of finding the best solution to the objective function under the given constraints. Typically, the solution is found using the branch-and-bound method (van der Linden, 2000b). (Please note that because CPLEX is a proprietary program and hence the source code is not

available to the public, there is no way to guarantee that this is indeed the way that CPLEX finds the solutions.)

The branch-and-bound method is most easily understood in graphical form. Figure 5 illustrates the branch-and-bound method for the simple case of selecting two items from four possible items. The four possible items are represented in parentheses as follows: (item one, item two, item three, item four). A one indicates that the item will be selected and a zero indicates that it will not. In this example, two of the four items are to be selected, so ultimately there shall be two ones and two zeros representing which two of the four items have been selected. An asterisk indicates that a decision has not yet been made about that item.

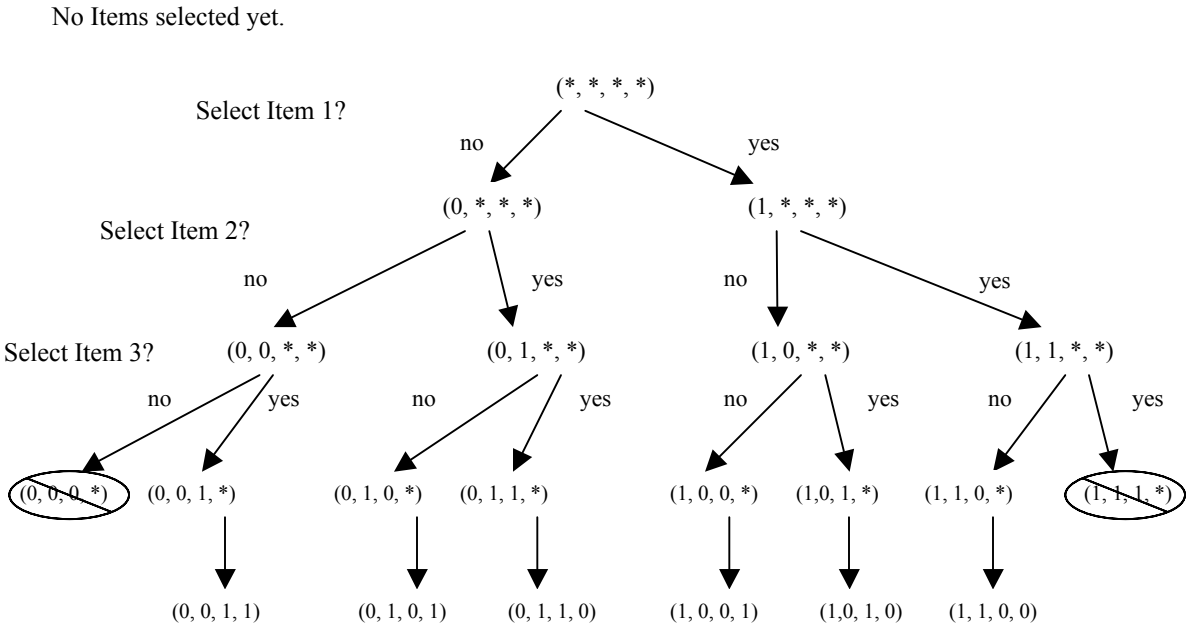

Figure 5: *The branch-and-bound method.*

The first decision is whether or not to include the first item or not, resulting in two possible branches. Each of these branches splits into two more branches to decide whether to include the second item or not. For example, (0, 1, *, *) indicated the first item is not selected, the second item is selected, and a decision has not yet been made about the remaining two items. This branching continues until all of the possible items have been considered. When a particular branch includes a set of items that do not obey the specified constraints, then that branch is no longer explored, that is to say, it is "bound" back and not allowed to "grow" any more. For instance, the branch (0, 0, 0, *) and (1, 1, 1, *) were both "bound" because they would not have yielded a total of two selected items.

These types of problems are often referred to as "knapsack" problems because they can be thought of in terms of trying to select certain objects to be put into a knapsack for a trip. The packer wants to select the best possible combinations of things to put in the knapsack. Most likely, one would select just a few objects from a variety of categories when packing the knapsack. In this analogy, the knapsack is the shadow test and all of the possible objects to choose from are contained in the item bank. The content categories of the items may be thought of as the different categories of the objects to be including. Any time a branch in the algorithm is encountered that does not correspond with all of the requirements in the list of constraints, that part of the branch is abandoned. Once all of the possible combinations of items which obey the list of constraint are determined, the objective function is used to select the best combination. For example, if the objective function is to maximize Fisher information, then the Fisher information is

calculated for every combination of items that obey all of the constraints, and the combination with the greatest Fisher information is selected as the best combination of items. This winning combination then becomes the Shadow Test. For more information about linear programming or the branch-and-bound method, see Bertsimas and Tsitsiklis (1997) or Hawkins (1988).

Once a shadow test is assembled, the best item with respect to the attribute vector estimate is selected from the shadow test to be the next item administered to the examinee. Two strategies, Shannon Entropy and Kullback-Leibler Information, are employed for the selection process as described in Xu, Chang, and Douglas (2003).

Shannon Entropy. Shannon Entropy was introduced in 1948 as a measure of uncertainty from a probability standpoint (Harris, 1988). Shannon Entropy is an indicator of a random variable's uncertainty or disorder. It is a nonnegative concave function of the random variable's probability distribution. In the context of this paper, the goal is to minimize Shannon Entropy; that is to say, it is more desirable to have less uncertainty. Shannon Entropy is described by Equation (12):

$$Sh(\underline{\pi}) = \sum_{i=1}^{K} \pi_i \log\left(\frac{1}{\pi_i}\right) \tag{12}$$

where π_i is the probability that the random variable of interest, call it *Y*, takes on a particular value y_i, and $\underline{\pi}$ is the probability vector containing the π_i's for all possible values of y_i (Xu, et al., 2003). In the context of diagnostic assessment, where we are

interested in estimating attribute vectors, the function for Shannon Entropy becomes Equation 13, as described by Xu, et al. (2003):

$$Sh(\underline{\pi}_n, X_i) = \sum_{x=0}^{1} E_n(\underline{\pi}_n | X_i = x)\, P[X_i = x | \underline{\pi}_{n-1}]$$
$$= \sum_{x=0}^{1} \left\{ E_n(\underline{\pi}_n | X_i = x) \left(\sum_{c=1}^{2^M} P_i^x(\underline{\alpha}_c)[1 - P_i(\underline{\alpha}_c)]^{1-x} \pi_{n-1}(\underline{\alpha}_c) \right) \right\} \quad (13)$$

where X_i is an item in the bank, $\underline{\alpha}_c$ is the possible candidate attribute vector generated by the i^{th} item, and π_{n-1} is the posterior probability distribution of a candidate pattern after n-1 items have been administered.

Kullback-Leibler Information. Kullback-Leibler (K-L) information was introduced in 1951 as a distance measure between probability distributions (Kullback, 1988). It is used as a measure of discrimination between a true distribution and another distribution (Kullback, 1988). More recently, K-L information has be used as a measure of global information for the purpose of item selection in IRT (Chang and Ying, 1996) and as an index in the item selection process in diagnostic assessment (Xu, et al., 2003). The definition of K-L information for continuous probability distributions is given by Equation (14).

$$K(f, g) = \int \log\left(\frac{f(x)}{g(x)} \right) f(x) \mu(dx) \quad (14)$$

For the cognitive diagnosis context, we want to use K-L information as an item selection criterion. The integral becomes a sum when the variables are discrete; then the

sum is taken across all possible attribute patterns. Thus the function becomes Equation (15):

$$K_i(\hat{\underline{\alpha}}) = \sum_{c=1}^{2^M} \left\{ \sum_{x=0}^{1} \log\left(\frac{P(X_i = x|\hat{\underline{\alpha}})}{P(X_i = x|\underline{\alpha}_c)} \right) P\left(X_i = x|\hat{\underline{\alpha}}\right) \right\} \quad (15)$$

where $\hat{\underline{\alpha}}$ is the currant estimate for the attribute vector and $\underline{\alpha}_c$ is the possible candidate attribute vector generated by the i^{th} item (Xu, et al., 2003). This yields an information index relating our current attribute vector estimate with the possible attribute vector estimate resulting from the administration of the next item *i* for every possible remaining item. The item with the largest value of $K_i(\hat{\underline{\alpha}})$ is then selected as the next item.

K-L information is beneficial in the context of cognitively diagnostic assessment because it easily lends itself to the categorical case, unlike alternative forms of information, such as Fisher information. Recall from Equation (11) that Fisher information requires a derivative function, which does not exist for a discrete random variable.

The main concern of a computationally intensive selection procedure like K-L information is an increase in computation time. Because examinees are waiting for the item selection procedure during administration, it is imperative that the procedure occurs in a timely fashion. Cheng and Liou (2000) compared the item selection procedures involving maximizing Fisher information and K-L information and noted that K-L information took longer than Fisher information based algorithms, especially as the size of the item bank increases. Although slower than Fisher information-based algorithms, the use of the K-L information procedure for item selection was still quite fast. In Cheng

and Liou's (2000) study, item selection using K-L information took a quarter of a second on a Pentium II 266 MHz PC. When Eggers (1999) and Chen, Ankenmann, and Chang (2000) compared the use of K-L information with Fisher information in the context of item selection, they found that K-L information performed as well as Fisher information, if not better.

In this study, the minimization of Shannon Entropy and the maximization of K-L information are each used to select the best item from the Shadow Test with respect to the attribute vector. The procedure of this study is explained in the following chapter.

CHAPTER THREE: METHODOLOGY

The research design is comprised of two major sections. The first part deals with the application of the fusion model in analyzing examinees' results on an existing exam in order to provide diagnostic feedback from an exam that would otherwise provide none. The rationale and advantages of this have been previously explained. The second major segment deals with incorporating computerized adaptive testing technologies to adaptively assess examinees from a diagnostic standpoint. This section compares three procedures for administering a computerized adaptive diagnostic assessment, one of which employs the method of shadow testing to select items based on optimal estimates of both the attributes and the conventional IRT ability parameter simultaneously.

PART ONE: Analyze an Existing Test in a Cognitive Diagnosis Framework

This study applies the Fusion model to a pair of assessments required by the Texas Education Agency (TEA). The first dataset of interest involves results from the third grade Texas Assessment of Academic Skills (TAAS) administered in the spring of 2002. The second dataset is from the spring 2003 administration of the eleventh grade Texas Assessment of Knowledge and Skills (TAKS). Both the third grade and eleventh grade tests contain a reading section and a math section. First, a simple random sample of two thousand examinees was obtained from the approximately three hundred thousand students who took each exam. The next requisite for conducting a diagnostic analysis was to obtain a Q-matrix to represent the attributes measured by the items. Three Q-matrices were developed for both portions of the eleventh grade TAKS and third grade TAAS. One of the Q-matrices is based on the test blueprint provided by TEA, the second

is based on the attribute assignments provided by content experts at TEA, and the remaining Q-matrix is based on an intuitive evaluation of the test items by the author. These three approaches to Q-matrix construction reflect different possible methods a psychometrician might adopt if s/he desired to conduct a cognitively diagnostic analysis.

The analyses are then performed using the fusion model software program Arpeggio (Hartz et. al, 2002). This procedure involves a series of three major steps. First, the Arpeggio program is executed and then the estimates for the mastery proportions for each of the measured attributes are examined. It is important these be accurately estimated with a given level of confidence because they are used as the cut-off values for the attribute mastery assignments. When these parameters have been determined to be estimable with in a certain acceptable range (the ninety-five percent confidence interval width must be no greater than three quarters of the overall possible range for the parameter), a stepwise reduction procedure is used to update the Q-matrix to be more useful. This stepwise reduction procedure examines the upper confidence interval bound of the attribute-based item discrimination parameter r_{ik}^{*}. If this value is too high (i.e. above 0.9), then the corresponding Q-matrix entry is removed on the rationale that a high Q-based item discrimination parameter indicates that the designated attribute is not important in obtaining a correct response. (Recall that a low r_{ik}^{*} is desirable.) If any of the fusion model item parameters are determined to not be estimable, additional steps are implemented which include the execution of an application designed for dealing with this issue. Lastly, additional applications are used to evaluate the model's fit in the analysis. More specific details on the process involved in using the

Arpeggio program can be found in Hartz et al., 2002.

The results of the analyses may then be examined to determine the appropriateness of the application of the fusion model to this data. The effectiveness of this model is taken into consideration by examining the proportion of the examinees that are accurately classified.

This portion of the study's design compares the effects of using different Q-matrices in the cognitively diagnostic analysis. This highlights an important issue because of the subjective nature of Q-matrix development, which is a buttressing step in the process of diagnostic assessment. The use of a variety of Q-matrices also addressed the fact that the differing *structures* possible for a Q-matrix, namely simple versus complex structure, may influence the results. The two tests chosen for the analysis exhibit varying levels of complexity; this incorporates a range of possible realistic circumstances.

The portion of the research provides information dealing with the application of a cognitive diagnostic model to an existing mandatory standardized assessment. Some possible future research directions might be to include the rule space methodology in such an analysis or to cross-validate the resulting attribute vectors of the examinees, perhaps by means of interviewing a portion of the sampled students or their teachers.

PART TWO: Adaptively Enhancing the Assessment Process

This part of the research study explores a method for selecting items in a computerized adaptive assessment based on a cognitive diagnosis framework. This portion of the project utilizes a simulation of a computerized adaptive assessment. Computer simulation is an important tool in measurement research because it allows researchers to simulate the testing process with a set of known data. The description of this part of the research first elaborates on the procedure for obtaining the data and item parameters involved, and then expounds on the three different cognitive diagnosis-based approaches to the item selection procedure.

Data and Parameters

This study is based on real responses to items administered by the Texas Education Agency. The item response patterns of a simple random sample of two-thousand examinees from each of three administrations of the third grade TAAS (years 2000 through 2002) are analyzed using BILOG-MG to obtain the three parameter logistic model (3PL) item and ability parameters. The response patterns are also analyzed using the Arpeggio program (Stout, et al., 2002) to calibrate the fusion model parameters, including item, person, and attribute parameters. This analysis leads to a set of values describing mastery levels for every attribute as well as a traditional IRT ability estimate for each examinee.

The mastery/non-mastery assignments of the eleven attributes for each of the two

thousand examinees are determined by dichotomizing the examinee's values in the $\underline{\alpha}_j$ vectors using the master proportion, p_k, as a cutoff value for attribute k. An examinee's mastery level for a specific attribute takes on a value of one (indicating the given examinee has mastered the attribute) if α_{jk} is greater than or equal to the p_k value for the attribute, and a zero (indicating non-mastery) otherwise. The analysis using Arpeggio also revises the Q-matrix for the items by means of the stepwise reduction procedure described in Part One of this chapter.

The values of the ability parameter obtained through the BILOG-MG calibration will be treated as the known or "true" values and will be denoted as θ_0. The final estimates of the ability parameter will be compared with these values. Likewise, the dichotomous ability parameters obtained from the Arpeggio analysis will be considered the known or "true" attribute mastery patterns and will be represent by $\underline{\alpha}_0$.

Once the 3PL item parameters and revised Q-matrix entries are obtained for each item, the process of CAT simulation may begin. Figure 6 outlines how the item parameters are obtained.

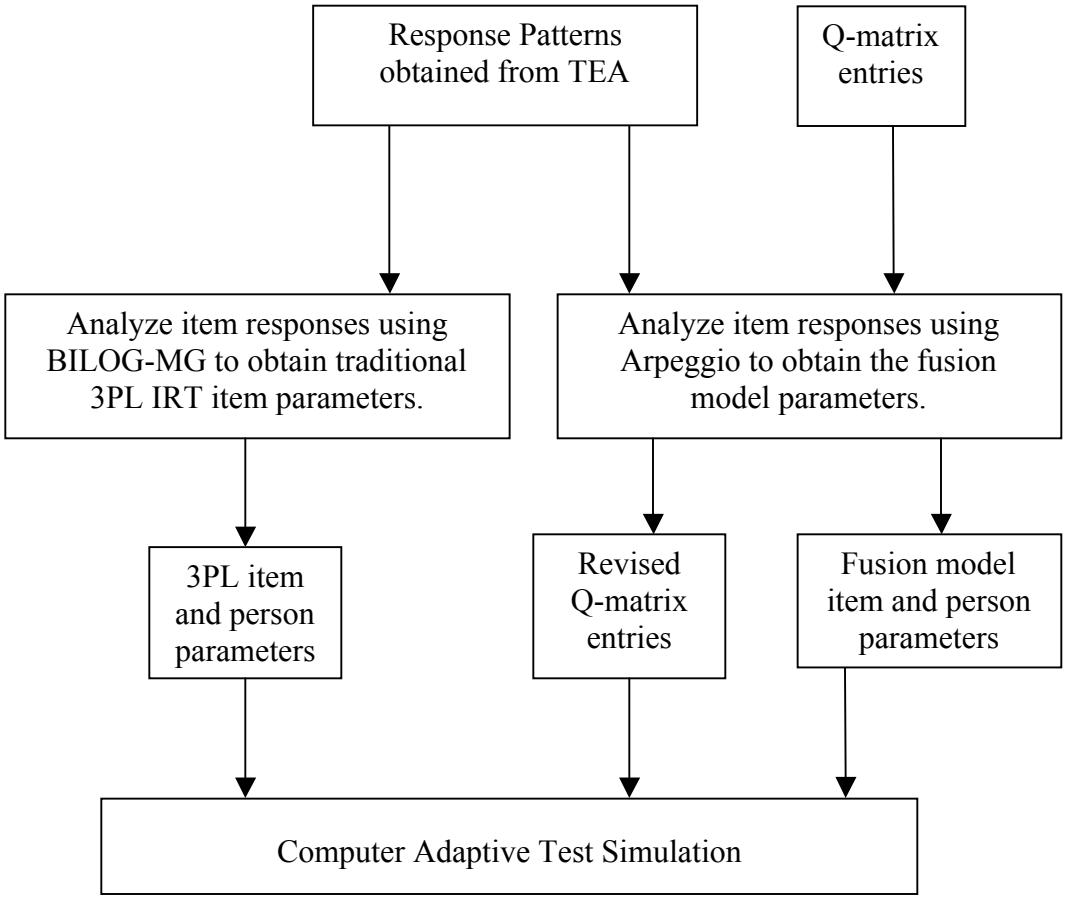

Figure 6: *Using item response patterns to obtain item and person parameters.*

The study's design for the simulation portion includes three conditions for comparison. Each condition reflects a different approach to item selection in adaptive testing. The results are then compared with the known values of the examinees' ability levels and attribute vectors obtained from the observed response data as described above to evaluate the accuracy of the three conditions.

Research Design

The study's design includes three conditions for reflecting three possible items selection methods. One selects items based on the traditional unidimensional IRT-based ability estimate $\hat{\theta}$ only, another selects items based on the cognitive diagnosis-based attribute vector estimate $\underline{\hat{\alpha}}$ only, and the last selects items based on both. The first condition implements the conventional method of focusing solely on $\hat{\theta}$ during item selection. The second condition, focusing solely on $\underline{\hat{\alpha}}$, mimics the approach outlined in Xu, et al. (2003) to select items. The third condition uses the estimates of both $\underline{\hat{\alpha}}$ and $\hat{\theta}$ to select the best item to be administered next in the test. For all three conditions, the goal is to estimate both traditional IRT ability estimates ($\hat{\theta}$) as well as diagnostic attribute mastery levels ($\underline{\hat{\alpha}}$) for each examinee. The difference between the conditions is in the item selection procedure of the adaptive assessment.

The third condition is the heart of this study. It first involves the construction of a shadow test that is optimized according to the ability level $\hat{\theta}$ (as outlined by van der Linden and Reese, 1998) before the administration of each item. Then, the best item for measuring the attribute vector is selected from the shadow test based on the current $\underline{\hat{\alpha}}$ estimate using Shannon Entropy or Kullback-Leibler Information as outlined in Xu, et al. (2003). The first two conditions reflect methods which have already been implemented in previous research. This third method hold the unique contribution of this research study by focusing on both traditional IRT ability estimation as well as the cognitively diagnostic attribute information.

Condition 1: Theta-based Item Selection.

The first condition uses the conventional method for item selection, which focuses solely on the theta estimate $\hat{\theta}$ during item selection. The item to be administered next is determined as the item with maximum Fisher information given the current estimate of $\hat{\theta}$. Once all of the items are administered, and the final ability estimate is obtained, the individual attribute vectors $\underline{\hat{\alpha}}$ are estimated using the maximum likelihood estimation procedure.

Condition 2: Alpha-based Item Selection.

The second condition takes the cognitive attribute vector $\underline{\hat{\alpha}}$ into account in the item selection procedure. In this condition, an item is selected when it is the best item for the current estimate of the attribute mastery vector $\underline{\hat{\alpha}}$ for the given examinee. To do so, this condition mimics the approach outlined in Xu, et al. (2003), which uses Shannon Entropy or Kullback-Liebler Information to determine the best item for a given $\underline{\hat{\alpha}}$ estimate. This study includes both item selection methods of minimizing Shannon Entropy or maximizing K-L information as sub-conditions of condition 2. After all items have been administered and the final estimate of the $\underline{\hat{\alpha}}$ attribute vector has been obtained, the values of the $\hat{\theta}$ ability estimates are calculated from the individual response patterns using the maximum likelihood estimation procedure.

Condition 3: Theta- and Alpha-based Item Selection.

Condition 3 is the part of the project that utilizes the shadow testing approach. The aim of this condition is to simultaneously use $\hat{\theta}$ estimates and $\underline{\hat{\alpha}}$ estimates to select items in a computerized adaptive testing administration. This third condition has two sequential sections. First, a shadow test is constructed that is optimized with respect to the current estimate of $\hat{\theta}$. This ensures that whichever item is selected to be administered to the examinee is optimal for his/her current estimate of $\hat{\theta}$.

Notice the reliance on the "current" estimate of $\hat{\theta}$ is this procedure. Before the first shadow test can be constructed, an initial "current" estimate of $\hat{\theta}$ and $\underline{\hat{\alpha}}$ is required. Generally, the mean value is used as the initial estimate; this is reasonable because it is less likely that a given examinee's true ability level is going to lie in the remote extremes of possible ability values than near the mean (Thissen and Mislevy, 2000). So, the mean of the population of examinees' ability levels is a suitable initial approximation for an examinee's ability estimate (Thissen and Mislevy, 2000), and likewise the mean of the estimates for the attribute vectors is also a sensible initial estimate for a given examinee's attribute mastery vector estimate. However, the mean of a set of dichotomous variables is not very meaningful considering the estimates of the attribute mastery levels must also be dichotomous. Therefore, a more fitting initial estimate of $\underline{\hat{\alpha}}$ would be the most frequently occurring attribute mastery pattern, or in other words, the mode.

Once the initial estimates of $\hat{\theta}$ and $\underline{\hat{\alpha}}$ have been ascertained, a shadow test is assembled that is optimal at the mean value of the theta ability level. Naturally, this will

be the same initial shadow test for all examinees.

As previously mentioned, a shadow test is constructed form an objective function and a list of constraints by means of the branch-and-bound method using the software program CPLEX (ILOG, 2003). The objective function for forming the shadow test follows the conventional maximization of Fisher information, mathematically denoted as $\sum_{i=1}^{I} I(\theta_i)x_i$. The applicable constraints are listed below.

$\sum_{i=1}^{I} x_i = n$	*for i = 0, 1, 2, 3,..., I (I= the # of items in the pool) and a total test length of n.* This regulates the test length.
$\sum_{i=1}^{I} x_i q_{ia} < or > Const_a$	*for a = 0, 1, 2, 3,..., A (A= the # of attributes.)* This ensures there are a specified number of items measuring each attribute.
$\sum_{i \in V_g} x_i = n_g$	*where V_g is a set of items that belong to category g and g=1...G. The value of G is the number of content categories, and g is a content area like geometry or reading comprehension.* This allows that a certain number of items be administered for each content category.
$f_i x_i \leq f_i^{\max}$	*where f_i is the frequency of the exposure of item i.* This is to constrain item exposure control.
$\sum_{i \in c_{k-1}} x_i = k - 1$	*where c_{k-1}=the set of items already given.* This makes sure the shadow test includes all previously administered items.

Other possible constraints could deal with the type of item administration or the item duration involved in answering each item. Recall the original goal of a shadow test is not only to include a set of items that are optimal for a given ability estimate level, but also to ensure these items obey content balancing and any other constraint the test administrator requires.

After inputting this information into CPLEX and completing the analysis, an output file presents the list of the items to be included in the shadow test. The purpose of constructing a shadow test in this manner is to obtain a set of items that are all good items for the current estimate of $\hat{\theta}$ that also obey the assigned constraints. Therefore, no matter which item is selected, it is a good item from the viewpoint of traditional unidimensional IRT measurement.

The second stage of the process involves selecting the next item to be administered to the examinee from this shadow test. This is the stage that takes into account a cognitive diagnosis component during the item selection process. Selecting an item from the shadow test is based on appraising each item's worth in contributing to the estimate of the examinee's $\underline{\alpha}$ vector. The additional information regarding the cognitive attribute vector provided by this approach can be considered supplemental to the conventional method of looking solely at the unidimensional IRT ability level because all items in the shadow test are already optimal with regard to the estimate, $\hat{\theta}$. So, any diagnostic information provided by the estimation of the attribute vector may be considered an extra bonus. Applicable methods for selecting items from the shadow test based on diagnostic information include minimizing Shannon Entropy and maximizing

K-L Information. Just as in condition 2, each of these methods is considered a sub-condition within condition 3.

A computer program will be written to calculate the Shannon Entropy of the items selected by CPLEX for the shadow test and determine the smallest value among them. Likewise another program will be composed to calculate the K-L information for the set of items selected by CPLEX and to single out the greatest value among these. The item selected by this program after the construction of each shadow test will be the item to be administered next. The items in the shadow test that are not selected are retuned to the item pool.

Once an item is selected, the correctness of the response is simulated by comparing the probability of obtaining a correct response with a number randomly drawn from the uniform distribution between zero and unity. The probability is calculated from the 3PL model given the examinee's ability level. If the random number is greater than the probability of obtaining a correct response, then the item is scored as incorrect; otherwise the response is designated as correct.

New estimates of $\hat{\theta}_j$ and $\underline{\hat{\alpha}}_j$ for examinee *j* are then calculated given the responses to all previously administered items using the maximum likelihood estimation procedure for each. This cycle of administering selected items and updating the estimates of $\hat{\theta}$ and $\underline{\hat{\alpha}}$ from the simulated responses is repeated again and again until the desired test length has been administered. The computer adaptive testing simulation procedure for condition 3 is illustrated in Figure 7.

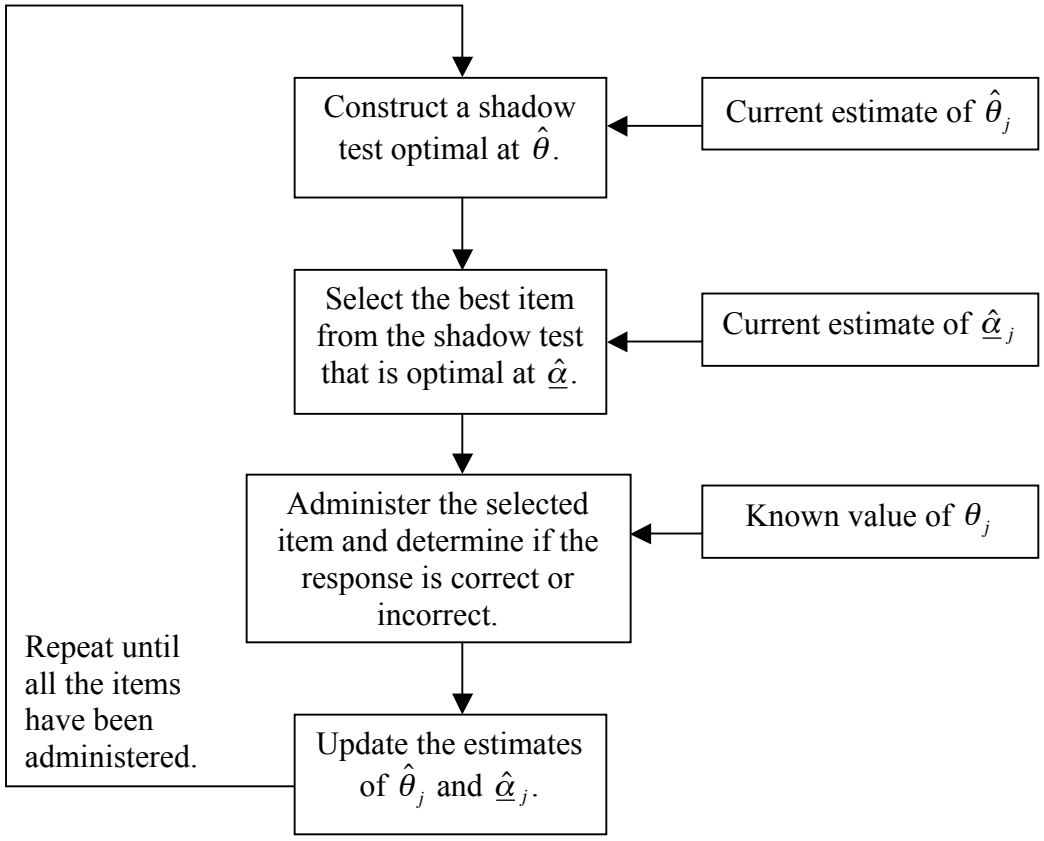

Figure 7: *Computerized adaptive testing simulation process for condition 3*.

Upon completion of the simulation study, the final estimates of $\hat{\theta}$ and $\hat{\underline{\alpha}}$ are compared with the known values, θ_0 and $\underline{\alpha}_0$, obtained from the initial calibration. This comparison is then used to evaluate the various methods and conditions. The various conditions of this study are depicted graphically in Figure 8.

Condition 1:

- Select items based on traditional IRT $\hat{\theta}_j$ estimates only.
- Estimate $\underline{\hat{\alpha}}_j$ values afterwards by MLE.

Items selection is based on the maximum information method.

Condition 2:

- Select items based on cognitive $\underline{\hat{\alpha}}_j$ estimates only.
- Estimate $\hat{\theta}_j$ values afterwards by MLE.

Condition 2a:	Condition 2b:
Item selection is based on minimizing Shannon Entropy.	Item selection is based on maximizing K-L information.

Condition 3:

- Select items based on both traditional IRT $\hat{\theta}_j$ and cognitive $\underline{\hat{\alpha}}_j$ estimates.
  - Construct a shadow test optimal for the current $\hat{\theta}_j$ level.
  - Select best item from the shadow test according to the current estimate of $\underline{\hat{\alpha}}_j$.

Condition 3a:	Condition 3b:
Item selection from the shadow test is based on minimizing Shannon Entropy.	Item selection from the shadow test is based on maximizing K-L information.

Figure 8: *Visual representation of the research design.*

Comparative Evaluation

Results of the three conditions are to be evaluated in terms of the accuracy of both the attribute mastery level estimates and the traditional IRT ability estimates. Evaluation of the attribute vectors is conducted by comparing the final estimated values of $\hat{\underline{\alpha}}_j$ with the attribute mastery levels from the Arpeggio original analysis of the real data, $\underline{\alpha}_{j0}$, for each examinee *j*. Similarly, the final IRT ability estimates $\hat{\theta}_j$ is be compared with the corresponding values of the ability parameter obtained from the BILOG-MG analysis of the original dataset, θ_{j0} for each examinee *j*. If a given condition works well, then the final estimates of both parameters θ and $\underline{\alpha}$ should match the corresponding known parameters θ_0 and $\underline{\alpha}_0$ respectively.

These comparisons are made by means of correlation and scatter plot. The final estimates of $\hat{\underline{\alpha}}_j$ is compared with the attribute vectors provided by the Arpeggio analysis of the original real data by examining the hit rate of each attribute as well as the hit rate of the entire attribute pattern for each examinee. The item selection procedure(s) with the highest correlations and hit rates will be considered superior to the remaining procedures. If the approach(es) with the highest correlation between known and final estimated values of the traditional IRT ability parameter does not happen to also have the best hit rates for the cognitive attribute vectors, then the methods will be evaluated with respect to each criterion individually.

It is expected that condition 3, which selects items based on both the traditional IRT ability and the fusion model cognitive attributes will have the best estimates of both

types of parameters. Condition 1 is expected to have good estimates for θ, but sub-optimal estimates for $\underline{\alpha}$, while vice versa is expected for condition 2. The results will be presented in tabular form in subsequent chapters upon completion of the simulation.

References

Bertsimas, D. and Tsitsiklis, J. N. (1997). *Introduction to Linear Optimization*. Belmont, Massachusetts: Athena Scientific.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord nd M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesely. Chapters 17-20.

Birenbaum, M. and Tatsuoka, K. K. (1993). Applying an IRT-based cognitive diagnostic model to diagnose students knowledge states in multiplication and division with exponents. *Applied measurement in education 6*(4), 225-268.

Campione and Brown. (1990). Guided learning and transfer: Implications for approaches to assessment. In N. Frederiksen, R. L. Glasser, A. M. Lesgold, and M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (p.453-486). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chang, H. & Ying, Z. (1999). *a*-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211-222.

Chen, S., Ankenmann, R. D., and Chang, H. (2000). A Comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement, 24*, 241-255.

Cheng, P. E. and Liou, M. (2000). Estimation of trait level in computerized adaptive testing. *Applied Psychological Measurement, 24*, 257-265.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician, 49*, 327-335.

Chipman, S. F., Nichols, P. D., and Brennan, R. L. (1995). Introduction. In P. D. Nichols, S. F. Chipman, and R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (p. 327-361). Hillsdale, NJ: Lawrence Erlbaum Associates.

Davey, T. and Parshall, C. G. (1995). New algorithms for item selection and exposure control with computerized adaptive testing. Paper presented at the annual meeting of the Americal Educational Research Association, San Francisco, CA.

DiBello, L., Stout, W., and Rousses, L. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, and R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (p. 327-361). Hillsdale, NJ: Lawrence Erlbaum Associates.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with sequential probability ratio test. *Applied Psychological Measurement, 23*, 249-261.

Embretson, S. (1990). Diagnostic testing by measuring learning processes: Psychometric considerations for dynamic testing. In N. Frederiksen, R. L. Glasser, A. M. Lesgold, and M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (p.453-486). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Gelfand, A. E. (1997). Gibbs Sampling. In Kotz, S. Johnson, N. L. and Read, C. B. Eds.), *Encyclopedia of Statistical Sciences, update 1*. (p. 283-291). New York, NY: John Wiley and Sons.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis.* London, England: Chapman & Hall.

Gott, S. (1990). Assisted learning of strategic skills. In N. Frederiksen, R. L. Glasser, A. M. Lesgold, and M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (p.453-486). Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive testing. *Journal of Educational Measurement, 21*, 347-360.

Harris, B. (1988). Entropy. In S. Kotz, N. L. Johnson and C. B. Read, Eds.), *Encyclopedia of Statistical Sciences*, *Vol 2*. (p. 512-516). New York, NY: John Wiley and Sons.

Hartz, S. (2002). A Bayesian framework for the Unified Model for assessing cognitive abilities: blending theory with practice. Doctoral thesis, The University of Illinois at Urbana-Champaign.

Hartz, S.. Roussos, L., and Stout, W. (2002) Skills Diagnosis: Theory and Practice. User Manual for Arpeggio software. ETS.

Hawkins, D. M. (1988). Branch-and-bound method. In S. Kotz, N. L. Johnson and C. B. Read, Eds.), *Encyclopedia of Statistical Sciences*, *Vol 1*. (p. 314-316). New York, NY: John Wiley and Sons.

ILOG, Incorporation. (2003) CPLEX Software Program, version 8.1. Incline Village, NV: CPLEX Division.

Jiang, H. (1996). Applications of Computational Statistics in Cognitive Diagnosis and IRT Modeling. Doctoral thesis, The University of Illinois at Urbana-Champaign.

Kullback, S. (1988). Kullback Information. In S. Kotz, N. L. Johnson and C. B. Read, Eds.), *Encyclopedia of Statistical Sciences vol 4*. (p. 421-425). New York, NY: John Wiley and Sons.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics, 22*(1), 79-86.

Linn, R. (1990). Diagnostic testing. In N. Frederiksen, R. L. Glasser, A. M. Lesgold, and M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (p.453-486). Hillsdale, NJ: Lawrence Erlbaum Associates.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Mongraphs, 61,* No. 285.

Lord, F. M. (1971a). A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and psychological measurement, 31*, 805-813.

Lord, F. M. (1971b). A theoretical study of two-stage testing. *Psychometrika, 36*, 227-242.

Meijer, R. R. and Nering, M. L. (1999). Computerized adaptive testing: Overview and Introduction. *Applied Psychological Measurement, 23*, 187-194.

Parshall, Harmes and Kromrey (2000). Item Exposure control in computerized adaptive testing: The use of freezing to augment stratification. *Florida journal of educational research, 40,* 28-52.

Patz, R. J. and Junker, B. W. (June 1997). A straightforward approach to Markov Chain Monte Carlo methods for item response models. Technical Report 658. Retrieved August 27, 2003, from http://www.stat.cmu.edu/cmu-stats/tr/tr658/tr658.html

Rogers, H. J., Swaminathan, H. and Hambleton, R. K. (1991). *Fundamentals of item response theory: Measurememnt methods for the social sciences volume 2*. Chapter two: Concepts, models and features. Thousand Oaks, CA: Sage Publications.

Samejima, F. (1995). A cognitive diagnosis model using latent trait models: Competency space approach and its relationship with DiBello and Stout's unified cognitive psychometric diagnosis model. In P. D. Nichols, S. F. Chipman, and R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (p. 391-410). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*, 379-423, 623-656.

Smith, A. (April 2003). Markov Chain Monte Carlo simulation made simple. Retrieved August 27, 2003, from http://www.nyu.edu/gsas/dept/politics/grad/classes/quant2/mcmc_note.pdf

Stevens, J. (1996). *Applied multivariate statistics for the social sciences*. (p. 111). Mahwah, NJ: Lawrence Erlbaum Associates.

Stocking, M. L. and Lewis, C. (1998). Controlling item exposure conditional on ability on computerized adaptive testing. *Journal of educational and behavioral statistics, 23*, 57-75.

Stocking, M. L. and Lewis, C. (2000). Methods for controlling the exposure of items in CAT. In W. Van der Linden, and C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (p. 163-182). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Stocking, M. L. and Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied psychological measurement, 17*, 277-292.

Stout, W., et al. (2002). Arpeggio Software Program, version 1.1. Princeton, NJ: Educational Testing Service.

Sympson, J. B. and Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association,* (pp. 973-977). San Diego, CA: Navy Personnel Research and Developemtn Center.

Tatsuoka, K. K. (1995). Architecture of knowledge structure and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, and R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (p. 327-361). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tatsuoka, K. K. (1990). Toward integration of item response theory and cognitive error diagnoses. In N. Frederiksen, R. L. Glasser, A. M. Lesgold, and M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (p.453-486). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of educational measurement 20*(4).

Tatsuoka, K. K. (1984) Caution indices based on item response theory. *Psychometrika 49*(1), 95-110.

Tatsuoka, K. K. and Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of educational statistics 7*(3), 215-231.

Tatsuoka, K. K. and Tatsuoka, M. M. (1984) Bug distribution and pattern classification. *Psychometrika 52*(2), 193-206.

Tatsuoka M. M. and Tatsuoka, K. K. (1989). Rule space. In S. Kots and N. L. Johnson (Eds.) *Encyclopedia of statistical sciences, vol. 8* (p. 217-220). New York: Wiley.

Tatsuoka, K. K. and Tatsuoka, M. M. (1997). Computerized cognitive diagnostic adaptive testing: Effects on remedial instruction as empirical validation. *Journal of educational measurement 34*(1), 3-20.

Thissen, D. and Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (p. 101-134). Mahwah, NJ: Lawrence Earlbaum Associates.

Tierney, L. (1997). Markov Chain Monte Carlo Algorithms. In S. Kotz, N. L. Johnson and C. B. Read, Eds.), *Encyclopedia of Statistical Sciences, update 1* (p. 392-399). New York, NY: John Wiley and Sons.

U.S. House of Representatives (2001), Text of the 'No Child Left Behind Act'. Public Law No. 107-110, 115 Stat. 1425.

van der Linden, W. & Chang, H. (2003). Implementing Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow Test Approach. *Applied Psychological Measurement, 27*, 107-120.

van der Linden, W. (2000a). Optimal assembly of tests with item sets. *Applied Psychological Measurement, 24*, 225-240.

van der Linden, W. (2000b). Constrained adaptive testing with shadow tests. In W. Van der Linden, and C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (p. 27-52). Dordrecht, The Netherlands: Kluwer Academic Publishers.

van der Linden, W. and Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. Van der Linden, and C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (p. 1-25). Dordrecht, The Netherlands: Kluwer Academic Publishers.

van der Linden, W. & Reese, L. (1998, September). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259-270.

Veldkamp, B. P. and van der Linden, W. (2000). Designing item pools for computerized adaptive testing. In W. Van der Linden, and C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (p. 149-162). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Wainer, H. (2000). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (p. 1-22). Mahwah, NJ: Lawrence Earlbaum Associates.

Wainer, H. and Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (p. 61-100). Mahwah, NJ: Lawrence Earlbaum Associates.

Whitesitt, J. E. (1995). *Boolean algebra and its applications*. Mineola, NY: Dover Publications.

Whittaker, T. A., Fitzpatrick, S. J., William, N. J., and Dodd, B. G. (2003). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models. *Applied Psychological Measurement, 27*, 299-300.

Xu, X., Chang, H., & Douglas, J. (2003). A simulation study to compare CAT strategies for cognitive diagnosis. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.