# An Exploratory Technique for Finding the Q-matrix for the DINA Model in Cognitive Diagnostic Assessment: Combining Theory with Data

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Catherine Nyambura Close

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Mark L. Davison and Ernest C. Davenport, Jr., Advisers

January, 2012

# Acknowledgements

First, my deepest gratitude goes to my advisers professors Mark L. Davison and Ernest C. Davenport for advising me during graduate school and providing guidance on this thesis. The topic of this thesis arose from numerous discussions and exploration led by Mark Davison. Thank you for making time to answer questions as they came up. I am also thankful to my examining committee, professors Snigdhansu Chatterjee and Michael Rodriguez. Your feedback on the thesis was very helpful and I am honored that you were a part of this journey. I would also like to thank Professor Lawrence Decarlo for providing technical assistance on the Latent Gold software.

Last, I am very thankful to my family. The greatest support and inspiration came from my husband, Jay. Thank you for wholeheartedly embracing all that graduate school brought into our lives especially the difficult times. I am also very grateful to my parents for believing in me and being the best role models.

# Dedication

This thesis is dedicated to Ruthie.

# Abstract

Cognitive diagnostic assessment (CDA) is used to measure the specific knowledge structures and the processing skills that examinees possess. One of the components of CDA is the Q-matrix, a *J* x *K* matrix indicating whether an item *j* requires skill *k* for correct execution. Although the Q-matrix is usually considered known, emerging evidence indicate otherwise. As such, the purpose of this thesis was to investigate a potential exploratory technique that could be used to supplement theory in finding the Q-matrix of a cognitive diagnostic test in data that satisfy the DINA (Deterministic Input Noisy "And") model. The proposed method is based on principal components analysis. The components model is a reparameterization of the DINA model relating examinee responses to cognitive diagnostic tasks. Understanding the relationship between the components and DINA skills can provide information for Q-matrix development. This relationship was investigated by answering four questions, the first two being analytical and the other two being empirical: 1) When does a skill (dimension) in the Q-matrix correspond to a component in the components analysis? 2) When does a skill (dimension) in the Q-matrix fail to correspond to a component in the components analysis? 3) When does the proposed methodology (components analysis) yield a plausible and useful solution for Q-matrix development? And 4) When does the methodology result in a Q-matrix that improves parameter estimation and examinee classification accuracy? The results indicated that components analysis which is akin to item analysis on a block of items can indeed augment theory in developing Q-matrices especially for items that are designed to be diagnostic and that measure a narrow content domain. Some limitations are discussed and potential areas for further work highlighted.

# Table of Contents

# List of Tables

# List of Figures

Chapter 1

# 1 Introduction

Cognitive diagnostic assessment (CDA) is designed to measure the specific knowledge structures and the processing skills that examinees possess so as to provide information about the cognitive strengths and weaknesses of examinees (Leighton & Gierl, 2007**.** Leighton and Gierl (2007) define the term *skill* as the description of knowledge, procedural or declarative, that an examinee needs to successfully complete a task in a specific domain. The terms *skill* and *attribute* are used interchangeably.

The cognitive diagnostic models (CDMs) that CDA uses to relate the latent skills to observed behavior (tasks) require a Q-matrix (Tatsuoka, 1983). Henson and Douglas (2005) have defined this Q-matrix as a matrix having elements $q_{jk}$ for $J$ items and $K$ attributes indicating whether mastery of attribute $k$ is required by item $j$ such that, $q_{jk} = 1$, if item $j$ requires attribute $k$, and 0 otherwise. The Q-matrix embodies the design of the assessment instrument in use and in essence, determines the quality of the diagnostic information obtained through the assessment instrument (Rupp and Templin, 2008). Therefore, cognitive assessment test developers strive to ensure that the necessary procedures and experts in cognitive theory are used in determining the Q-matrix for a set of items written for diagnostic purposes.

Despite the major role that the Q-matrix plays in CDA, little literature exists about the development of the Q-matrix. This dearth in literature is partly because a) the area of cognitive diagnosis is relatively new in psychometrics and b) the Q-matrix is often considered fixed; that is, it does not need to be changed once it has been developed by the content specialists. Of the studies reviewed, two studies investigated the effects of Q-matrix misspecification (Rupp and Templin, 2008; DeCarlo, 2011); several other studies such as, Templin and Henson (2006); and Henson, Templin, and Douglas (2007) addressed the use of factor analysis (closely related to principal components analysis) in diagnosing skill mastery but not for Q-matrix development. Liu, Douglas and Henson (2009) addressed the use of factor analysis for Q-matrix development as discussed in Henson and Templin (2007). One study has examined a method for validating the Q-matrix (de la Torre, 2008).

The study by Rupp and Templin (2008) examined the effect of the Q-matrix misspecification on parameter estimates and skill classification accuracy using the DINA model. Rupp and Templin assumed that the true Q-matrix was known and therefore did not address Q-matrix development. Liu, Douglas and Henson (2009) state that although factor analysis is not typically used for Q-matrix development, it can be expected that factor analysis will give a reasonable solution when the Q-matrix is not overly complex. The study, however, investigated person fit not the use of factor analysis for Q-matrix development.  Henson, Templin, and Douglas (2007) compared the use of sum scores from unidimensional item response theory (IRT) models for mastery classification to the use of item-level factor analysis where the factor levels can determine the mastery of an individual on a set of skills. A simple way of computing factor levels is summing the items with sufficiently high loadings on a particular factor. Although the Henson et al.

(2007) study illuminates a possibility for determining the Q-matrix, the authors do not explicitly address Q-matrix development.

Only de la Torre (2008) has examined a method for validating the Q-matrix. De la Torre named his method, the *delta method*. The delta $\delta_j$ is the difference in the probabilities of the correct response between examinees who have mastered skill *j* and those who have not mastered skill *j*. Delta is therefore, a discrimination index for item *j* in that items that are highly discriminating have high $\delta_j$ values while those that are not highly discriminating have low $\delta_j$ values. This $\delta_j$ changes as the *q*-vector of an item changes. Using the delta method in conjunction with the DINA model, de la Torre conducted a simulation study and showed that the EM -based delta method can correctly replace the *q*-vectors that were misspecified in the Q-matrix with the correct ones while simultaneously retaining the *q*-vectors that had been correctly specified, most of the time. De la Torre concluded that although the delta method provides statistical information about the Q-matrix, it should not be used in isolation but in conjunction with substantive knowledge and domain expertise. In other words, the delta method for validating the Q-matrix does not replace theory. The delta procedure seems best characterized as a method for modestly adjusting an existing Q matrix, rather than a method for deriving a Q matrix as such. In his method the skills are fixed, and only the set of items posited to require the skill is adjusted.

While de la Torre's delta method is a significant step in examining a Q-matrix with a known number of attributes, methods that can determine the number of attributes (if unknown) and the items measuring those attributes are necessary. Clearly, exploratory techniques that can augment theory in determining the Q-matrix for diagnostic testing can be informative. In fact, Templin and Henson (2006) recognize the need to investigate empirical techniques for determining the entries of the Q-matrix. In their own words, “Techniques that allow the empirical data to mold the entries of the Q-matrix would provide helpful feedback for the construction of reliable instruments developed for use with cognitive diagnosis models.”

In the current use of cognitive diagnostic assessment, the Q-matrix is considered known (fixed) as solely determined by content specialists through theory. Recent studies have shown that considering the Q-matrix as known may be misleading. For example, DeCarlo (2011) highlights the ongoing debate about the true Q-matrix of Tatsuoka’s fraction subtraction data (Tatsuoka, 1983) following the findings in studies such as, de la Torre and Douglas, 2004; de la Torre, 2008; and Henson et al., 2009. As such, the purpose of this thesis is to evaluate  the potential of exploratory components analysis as a supplement to theory in finding the Q-matrix of a cognitive diagnostic test in data that satisfy the DINA (Deterministic Input Noisy “And”) model of Haertel (1989).

The evaluation is partly analytical and partly empirical.  In the analytical portion, the components in a components representation of items satisfying the DINA model are compared to the dimensions in the DINA model itself.  This comparison suggests limitations of the exploratory approach as a tool for deriving a Q-matrix, but it also suggests a method for using the results of a components analysis, in conjunction with theory and expert knowledge, to derive a Q-matrix under some circumstances. In the empirical portion, real data are used to derive a Q-matrix and evaluate the accuracy of the derived Q-matrix in classifying examinees.

The proposed components model is a reparameterization of the DINA model relating examinee responses to cognitive diagnostic tasks.  After describing the component form corresponding to the DINA model, the relationship between the components and the DINA skills is described by answering two sets of questions corresponding to the two approaches employed in this thesis: analytical and empirical. In the analytical part, simulated data will be used to illustrate how the proposed components analysis relates to the true Q-matrix under ideal conditions. An ideal situation is defined as a situation in which there are multiple items for each skill set required to answer an item. Two questions are of interest: 1) When does a skill (dimension) in the Q-matrix correspond to a component in the components analysis? And 2) When does a skill (dimension) in the Q-matrix fail to correspond to a component in the components analysis? By answering these questions, one can determine what are referred to here as sufficient and insufficient skills. A sufficient skill is defined as a skill that is the single skill needed to solve some items whereas an insufficient skill is one that must accompany other skills in solving tasks. The answer to the first question leads to identification of the sufficient skills (those that appear in isolation) whereas answering the second question leads to identification of the insufficient skills that never appear in isolation. In the empirical part, real data from real testing situations are used. The questions of interest are: 1) Does the proposed methodology (components analysis) yield a plausible and useful solution for Q-matrix development, and 2) Does the methodology result in a Q-matrix that improves parameter estimation and examinee classification accuracy after testing with a cross-validation sample. The first question is answered by examining the item loadings on each component to determine whether the components are conceptually sensible. The second question is answered by conducting a confirmatory analysis to test the accuracy of the developed Q-matrix.

Holistically, the components analysis can inform the task analysis used to derive a Q-matrix. The answers to the questions above will lead to a proposed method for deriving a Q-matrix from a combination of theory, item task analysis, and the components analysis. In the creation of a Q-matrix, content specialists must employ some form of task analysis to match items with the corresponding skills measured. In this thesis, I elaborate on task analysis based on the components analysis that will be conducted on a block of items rather than single items as is usually the case. The reason for examining blocks of items rather than single items is because components will usually have multiple items loading on them. The reader should be aware that only the DINA model is investigated in this thesis. Although extensions to other conjunctive models for cognitive diagnosis are straightforward, studies need to be conducted to test the proposed method with data sets satisfying other models.

Chapter 2

# 2 Review of Literature

In the recent past, interest has gradually shifted from using unidimensional models that rank order examinees on a continuous latent ability, to assessing the specific cognitive skills that students possess. DiBello and Stout (2007) attribute this shift to the interplay of scientific, political, and educational developments of the last few decades. As an example, the No Child Left Behind Act (NCLB) of 2001 and its accountability requirements has teachers formatively assessing students during the school year to ensure that AYP (adequate yearly progress) standards are met. As such, diagnosing each student’s strengths and weaknesses can provide information for tailoring instruction to the specific needs of the students as compared to unidimensional item response theory (IRT) that yields a single score that may not be diagnostically informative. Therefore, the NCLB mandate, albeit controversial, fosters the need for skills-based formative tests such as the so-called cognitive diagnostic assessments. Formative tests are administered throughout the school year and in addition to indicating mastery/non-mastery of instructional content, they can help predict progress towards summative tests administered at the end of the school year.

The cognitive diagnostic models (CDMs) that cognitive diagnostic assessment (CDA) uses to relate the latent skills to observed behavior (tasks) are described in this thesis. A subset of these models is the focus of this review, which is organized into five sections. First, cognitive diagnostic models commonly found in the literature are presented. These CDMs include continuous and also nonparametric ones. Because the continuous and the nonparametric CDMs are not the focus of this review, only a brief summary is provided to alert the reader to their existence. More relevant here are CDMs that are restricted LCMs (Latent Class Models) in that examinees are classified into *discrete* latent classes based on skill mastery. Moreover, these discrete models are also classified as being either conjunctive, where mastery of all required skills is required for a high probability of a correct response; or disjunctive where mastery of at least one sufficient skill leads to a high probability of a correct response on an item. In particular, this review focuses on conjunctive CDMs that classify examinees in a discrete latent space based on the cognitive perspective that a conjunctive model appears more appropriate for the cognitive assumptions made (Dibello et al., 2007). Second, the role of the Q-matrix in CDA is explained alongside a review of some studies that have addressed issues relevant to the Q-matrix such as, possible techniques of validating the Q-matrix. Only models that use a Q-matrix will be discussed. Third, studies for investigating the CDMs of interest that use both simulated data and real data are reviewed. To conclude, an evaluation of the reviewed literature is presented.

## 2.1 Cognitive Diagnostic Models

One of the earliest methods of skills diagnosis is the Rule Space Approach (RSA) of Tatsuoka (1983). Using a Q-matrix and examinees’ responses, RSA classifies examinees in a Euclidean ‘rule space’ mainly to identify atypical responses (see a detailed description in Gierl, 2007). By modeling atypical responses, the RSA diagnoses skill mastery through error analysis. A closely related method is the Attribute Hierarchy Method (AHM) of Leighton, Gierl, and Hunka (2004), which is a variant of the RSA for estimating the skills that examinees have mastered (see also Gierl, 2007). The RSM and the AHM are similar in that an artificial neural network approach is used to estimate the probability that an examinee possesses specific attributes. The methods differ in that the AHM specifies an attribute hierarchy, a priori, upon which item development is based while the RSA identifies attributes after the test items have been constructed and administered.

Multidimensional item response theory (MIRT) models also exist for diagnostic testing (Dibello et al., 2007; see also Stout, 2007). These models include the MIRT- C (compensatory) of Reckase and McKinley (1991); the MIRT-NC (non-compensatory) of Sympson (1977); the GLTM (General Component Latent Trait) model of Embretson (1985); and the MLTM (Multi-component latent trait model) of Embretson (1985) and Whitely (1980). Although different from MIRT, Almond et al. (2007) present a multidimensional method of diagnostic testing using Bayesian network modeling (BNM). This method is predicated on the evidence centered design (ECD) paradigm of Mislevy, Steinberg and Almond (2003). In BNM, a proficiency model is developed alongside an evidence model. The proficiency model specifies the proficiencies being tested and the relationships between them; whereas the evidence model specifies the relevant tasks together with the links between the tasks and the proficiencies. This network is then subjected to a Bayesian analysis to classify examinees according to the skills mastered. The BNM method offers great flexibility in modeling complex diagnostic situations.

Currently, cognitive diagnostic testing favors discrete CDMs that link observable data, mainly in the form of examinee responses to test items, to a set of categorical latent abilities that are either dichotomous or polytomous (Templin & Henson, 2006). These CDMs are restricted models that are special cases of the general latent class model (LCM) of Lazarsfeld and Henry (1968)**.** The general latent class model is,

$$p(\underline{X}_i = \underline{x}_i) = \sum_{c=1}^{C} p(c) \prod_{j=1}^{J} p(X_{ij} = x_{ij} \mid c) \tag{1}$$

Where, $P(X_{ij} = x_{ij} \mid c)$ is the probability of a response $x_{ij}$ for individual *i* on item *j* (*j*=1,..., *J*), conditional on the class membership *c* (*c=1,…,C*) for individual *i*. P(c)indicates the probability that any given individual is a member of class *c*, with the constraint that $\sum_{c=1}^{C} p(c) = 1$. By the assumption of local independence inherent in this model, the probability of a response vector, $\underline{X_i}$=$\underline{x_i}$ given class *c*, is the product of the *J* conditional item response probabilities given by the latent class analysis model. The number of latent classes, *C*, is determined by the number of dichotomous skills, *K*, measured such that *C*= $2^k$. This model is unrestricted in that estimating $p(c)$ would require $2^k$-*1*parameters to be estimated and these parameters increase exponentially as the number of skills increase (Templin et al., 2008). In other words, the number of classes in the LCM is determined by the number of unique patterns of skill combinations. In effect, *K* dichotomous skills will result in $2^k$-*1* unique latent classes.

Special constraints to the general LCM give rise to the different types of CDMs currently in use. Constraints involve various reparameterizations of the structural part of the latent class model, $p(c)$ , to reduce the number of parameters estimated. An earlier attempt to classify examinees into mastery versus non-mastery classes using a restricted LCM can be found in Macready and Dayton (1977). The restricted LCM that Macready and Dayton used was the foundation for more recent models like the DINA (Deterministic Input Noisy "And" ) model of Haertel (1989).

More recently, Henson, Templin and Willse (2009) have proposed a general log linear model with latent variables that is a general form of a family of cognitive diagnostic models. This general log-linear cognitive diagnosis model (LCDM) is constrained variously to give rise to both conjunctive and disjunctive models. Furthermore, the usual CDMs such as the DINA, the DINO, the compensatory RUM (Reparameterized Unified Model), and the reduced RUM can be recast into log-linear models (these models are described in the following sections). The extension to other models not discussed by Henson et al. (2009) is straight forward. The advantage of recasting the models into the log-linear form is the ability to compare models directly because the constraints applied usually result in nested models.

## 2.2 Disjunctive Models

In disjunctive CDMs, the interaction of skills in examinee responses is modeled such that mastery of at least one sufficient skill leads to a high probability of a correct response on an item in a test. Henson et al. (2009) define disjunctive models as models that define a high probability of a correct response based on mastery of a subset of skills, sometimes one skill. As a result, examinees who have mastered all the skills can be expected to perform similarly to those who have mastered only a few skills. Examples of disjunctive CDMs include the MCLCM-D (Multiple Classification Latent Class Model-Disjunctive) of Maris (1999); the MCLCM-C (Multiple Classification Latent Class Model-Compensatory) of Maris (1999); the DINO (Deterministic Input Noisy "Or") model of Templin and Henson (2006); and the NIDO (Noisy Input, Deterministic "Or") model of Templin, Henson, and Douglas (2006), which is a newer model that is not commonly used (Roussos, Templin & Henson, 2007). Disjunctive models can be used, for instance, in psychiatric or medical diagnosis where the presence of a symptom is indicative of the presence of at least one of the disorders (de la Torre & Douglas, 2004).

Define $X_{ij}$ as a binary indicator [1/0] of whether examinee *i* performed task *j* correctly. 1 indicates a correct response, 0 an incorrect response. $q_{jk}$ = a binary indicator [1/0] of whether attribute *k* is relevant for task *j*. 1 indicates that attribute *k* is required to correctly answer task *j* whereas 0 indicates that the attribute is not required for that item. $\alpha_{ik}$= a binary indicator [1/0] of whether examinee *i* possesses attribute *k*. 1 indicates possession whereas 0 indicates lack of the attribute. Then the IRF (Item Response Function) for the NIDO model is,

$$P(X_{ij} = 1 \mid \alpha_{ik}) = [1 + \exp(\sum_{k=1}^{K} (\tau_k + \beta_k \alpha_{ik}) Q_{jk})]^{-1} \qquad (2)$$

where $\tau_k$ and $\beta_k$ are the parameters representing the examinees who have mastered the attribute and those who have not mastered the attribute, respectively. The IRF for the DINO is,

$$P(X_{ij}=1\,|\,\omega_{ij})=(1-s_j)^{\omega_{ij}}\,g_j^{1-\omega_{ij}} \tag{3}$$

Where, $\omega_{ij}=1-\prod_{k=1}^{K}(1-\alpha_{ik})^{q_{jk}}$ and (1 - $s_j$) > g$_j$. $s_j$ and g$_j$ are slipping and guessing parameters, respectively. The MCLCM-D uses the parameterization of the Unified Model (UM) of Dibello et al. (1995). The item response function (IRF) is,

$$P(X_{ij}=1\,|\,\underline{\alpha}_i)=1-\prod_{k=1}^{K}(1-\pi_{jk})^{\alpha_{ik}q_{jk}}(1-r_{jk})^{(1-\alpha_{ik})q_{jk}} \tag{4}$$

Where, $\pi_{jk}$ is the probability of a correct response given mastery of the required attributes for an item and $r_{jk}$ is a penalty for lacking mastery of an attribute required by an item. In contrast, the IRF for the MCLCM-C is,

$$P(X_{ij}=1\,|\,\underline{\alpha}_i)=\frac{\exp(\sum_{k=1}^{K}a_{jk}\alpha_{ik}-d_j)}{1+\exp(\sum_{k=1}^{K}a_{jk}\alpha_{ik}-d_j)} \tag{5}$$

Where $a_{jk}$ is the amount of increase in probability on the log odds scale when $\alpha_{ik}$ goes from 0 to 1, and $d_j$ is a threshold parameter.

The NIDO is the disjunctive counterpart of the conjunctive NIDA model discussed later. The NIDO has the severe restriction that the probability of correctly applying a specific skill is constant across all the items requiring that skill. Likewise, the DINO is the disjunctive counterpart of the DINA model also to be discussed later. The probability of a correct response is the product of the $g_j$ and the $(1\text{-}s_j)$ terms. It also has the severe restriction that the probability of a correct response, given mastery of at least one skill, does not depend on the number or the type of skills required for the item. The MCLCM-D is an extension of the NIDO where the two parameters measured per skill are allowed to vary across the items. The MCLCM-D is unidentifiable. The MCLCM-C is similar to the multidimensional item response theory model (MIRT) of Reckase and McKinley (1991) (Roussos, Templin & Henson, 2007). Interested readers are referred to Dibello et al. (2007) for an in-depth analysis of the discrete disjunctive CDMs.

## 2.3 Conjunctive Models

In conjunctive models, the interaction of skills in examinee responses is modeled such that mastery of all the skills required by an item leads to a high probability of a correct response. On the other hand, non-mastery of any one of the skills measured by an item greatly reduces the probability of a correct response. Henson, Templin and Willse (2009) define conjunctive models as models such that the lack of mastery in one attribute cannot be compensated for by mastery of other attributes. The models examined in detail in this review are in this category and they include: the DINA (Deterministic Input Noisy "And") model of Haertel (1989); the NIDA (Noisy Input Deterministic "And") model of Junker and Sijtsma (2001); and the reduced RUM (Reparameterized Unified Model) that is a reduced version of the RUM of Hartz (2002).

### 2.3.1 The DINA model

The item response function (IRF) for the DINA model is,

$$p(X_{ij}=1|\underline{\alpha},s,g)=(1-s_j)^{\xi_{ij}}\,{g_j}^{1-\xi_{ij}} \tag{6}$$

Where, $\underline{\xi_{ij}}=\prod_{k=1}^{K}\underline{\alpha_{ik}}^{Q_{jk}}$ , $s_j = p(X_{ij}=0\,|\,\xi_{ij}=1)$, and $g_j = p(X_{ij}=1|\,\xi_{ij}=0)$. In the DINA model, the latent responses are defined by $\xi_{ij}$. $\xi_{ij}$ is the deterministic part of the model and it is binary. $\xi_{ij}$ is 1, if and only if, all the skills required by an item have been mastered, otherwise it is 0. The stochastic nature of the DINA is modeled by the $s_j$ and the $g_j$ parameters to indicate that an examinee who has mastered all the skills required by an item may slip (hence the $s_j$ parameter) and lower the probability of a correct response. Conversely, an examinee who has not mastered an attribute required by an item may guess (hence the $g_j$ parameter) thus increasing the probability of a correct response. Therefore, the latent response variables $\xi_{ij}$ are related to the observed responses $X_{ij}$ through the probabilities $s_j$ and $g_j$ (Junker and Sijtsma, 2001).

### 2.3.2 The NIDA model

The IRF for the NIDA model is given by

$$p(X_{ij}=1|\underline{\alpha},s,g)=\prod_{k=1}^{K}p(\eta_{ijk}=1|\,\alpha_{ik},Q_{jk})=\prod_{k=1}^{K}[(1-s_k)^{\alpha_{ik}}\,{g_k}^{1-\alpha_{ik}}\,]^{Q_{jk}} \tag{7}$$

where,

$X_{ij}=\prod_{k=1}^{K}\eta_{ijk}$ , $s_k = p(\eta_{ijk}=0\,|\,\alpha_{ik}=1,Q_{jk}=1)$ , $g_k = p(\eta_{ijk}=1|\,\alpha_{ik}=0,Q_{jk}=1)$ ,

and $p(\eta_{ijk}=1|\,\alpha_{ik}=a,Q_{jk}=0)\equiv 1$ regardless of the value of $a$ in $\alpha_{ik}=a$ .

The latent variable $\eta_{ijk}$ constitutes the noisy component of the NIDA and $\eta_{ijk}$ is related to the examinee attribute vector, $\boldsymbol{\alpha}_i$, through the probabilities $s_k$ and $g_k$. Because $X_{ij} = \prod_{k=1}^{K} \eta_{ijk}$, these latent variables, $\eta_{ijk}$, which reflect the examinee attributes $\alpha_{ik}$ are combined in a deterministic manner to produce $X_{ij}$, a correct or incorrect response. The difference between the DINA and the NIDA model is that the latent response variable is defined differently in the two models; that is, $\xi_{ij}$ in the DINA and $\eta_{ijk}$ in the NIDA. Notice that $\eta_{ijk}$ includes an extra subscript for skill, *k*, indicating that the NIDA latent response variable is specified at the attribute level whereas, and the DINA latent response variable is specified at the item level (Junker & Sijtsma, 2001). The similarity of the DINA and the NIDA lies in the stochastic nature of the models in that examinees' performance cannot be precisely predicted from the examinees' attribute vectors and therefore, they allow the possibility of slips and guesses (de la Torre & Douglas 2008).

## 2.3.3 The RUM model

Let $Y_{ijk} = 1$ if examinee *i* successfully executes skill *k* on item *j* (or if examinee *i* is not required to execute skill *k* on item *j*) and 0 otherwise. Then, the RUM has the item response function (IRF),

$$P(X_{ij} = 1 \mid \underline{\alpha}_i, \eta_i) = P(\prod_{k=1}^{K} Y_{ijk} = 1 \mid \underline{\alpha}_i) P_{cj}(\eta_i). \qquad (8)$$

where, $\underline{\alpha}_i$, denotes a vector of the state of mastery for the *i*th examinee on each of the skills ($\alpha_{1i,}\alpha_{2i}$, $\alpha_{3i}$, …, $\alpha_{Ki}$). $\eta_i$, is a unidimensional combination of the levels of skill mastery not specified in the Q-matrix for an examinee *i*. c$_j$, is the degree to which the Q-matrix is complete in its skills coverage

$$P(\prod_{k=1}^{K} Y_{ijk} = 1 \mid \underline{\alpha}_j) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{*\,(1-\alpha_{ik})Q_{jk}} ,$$

where,

$\pi_j^* = \prod_{k=1}^{K} \pi_{jk}^{Q_{jk}}$ and,

$$r_{jk}^* = \frac{r_{jk}}{\pi_{jk}}$$

$\pi_j^*$ is the probability of correctly applying all required attributes for the *j*th item given that the examinee has mastered all required attributes for that item whereas $r_{jk}^*$ is the discrimination of the *j*th item for the *k*th attribute (Henson, Roussos, Douglas, & He, 2008).

$P_{cj}(\eta_i)$, is the Rasch model with difficulty –c$_j$ and the ability parameter $\eta_i$, thus

$$p_{cj}(\eta_i) = \frac{1}{1 + \exp\{-1.7[\eta_i - (-c_j)]\}} \quad (9)$$

Because the RUM assumes an incomplete Q-matrix, skills not included in the Q-incidence matrix are modeled by the Rasch component of the RUM (Dibello et al., 2007). In all the studies reviewed here, however, the Rasch component is dropped to give rise to the reduced RUM. The reduced RUM is an abbreviated form of the RUM that omits the Rasch component. The reason for omitting the Rasch component is the perpetual estimation issue (nonconvergence) when the Rasch component is included in the model ( Ruossos, DiBello, Stout, Hartz, Henson, & Templin, 2007).

## 2.4 The Q-Matrix

Although examinee responses are also a component of the application of CDMs, perhaps more important is the Q-matrix. The Q-matrix contains elements 0 or 1 to indicate whether a particular skill is required to execute a particular item. In effect, the Q-matrix embodies the design of the assessment instrument in use and hence, determines the quality of the diagnostic information obtained through the assessment instrument (Rupp and Templin, 2008). Therefore, cognitive assessment test developers are faced with the task of ensuring that the necessary procedures and experts in cognitive theory are used in determining the Q-matrix for a set of items written for diagnostic purposes.

This section of the review is mainly concerned with investigating whether exploratory techniques, in addition to theory, have been proposed for finding the Q-matrix of a cognitive diagnostic test calibrated using the DINA, NIDA, and the reduced RUM.  An example would be an exploratory factor analytic technique for finding the Q-matrix that combines theory with data. To this end, extant literature is scarce. Most studies that address factor analysis focus on its use for cognitive diagnosis compared to CDMs. Templin and Henson (2006) claim that CDMs differ from factor analytic approaches in that the former are better suited to dichotomous latent traits and the accompanying constraints than factor analysis. Templin and Henson add that the latent trait factors in the factor analytic approach are assumed to be continuous and in effect, additional analyses are necessary to determine the thresholds that define group membership. Furthermore, they argue, factor analysis is fit for use if one is interested in both the factor scores and a possible diagnosis but that CDMs are more appealing to examiners because they eliminate the need for additional analyses to establish group membership. Others might argue that even though CDMs are easier to implement, factor analysis may provide a simple method for determining the elements of the Q-matrix while also providing diagnostic information.

Other studies, such as Henson, Templin, and Douglas (2007) have briefly mentioned the factor analytic approach to skills diagnosis. They compare the use of sum scores from unidimensional item response theory (IRT) models for mastery classification to the use of item-level factor analysis where the factor levels can determine the mastery of an individual on a set of skills. A simple way of computing factor levels is summing the items with sufficiently high loadings on a particular factor. Henson et al.(2007) choose to use the RUM to compute sum scores for estimating skills mastery. Henson et al. claim that the use of model-based sum scores can make work easier for the classroom teacher in diagnosing what students know. DiBello and Stout (2007) further reiterate that using such model-based sub-scoring for assessing skill mastery may provide face validity for some users who are interested in understanding the classification methods underlying the reported skills mastery. Clearly, the Q-matrix has received little attention.

Because the area of cognitive diagnostic assessment is relatively new, a review of literature on the use of the DINA, NIDA, and the reduced RUM follows next to familiarize the reader with the measurement issues that have been investigated thus far.

## 2.5 Review of Simulation studies using the DINA, NIDA, and the RUM

The reader should refer to Table A.1 (Appendix A) for the specifications regarding sampling and prior distributions, sample size, number of skills tested, and the assumed correlational structure of the skills in previous simulation studies. Review of the simulation studies that have been conducted using the DINA, NIDA, and the reduced RUM showed that most studies investigated the joint distribution of the latent attributes. Correct modeling of the discrete latent attributes can greatly simplify the CDM in use and in effect, reduce estimation problems that tend to receive considerable attention.  In addition to the joint distribution of the attributes, issues surrounding test construction for diagnostic use; misspecification of the Q-matrix; the reliability of diagnostic tests; person fit; use of sum-scores; retrofitting; and parameter invariance have been investigated.

In cognitive diagnostic assessment, correct modeling of the joint distribution of the attributes is important because it can greatly reduce the number of parameters to be estimated and hence reduce model complexity for identification purposes. To this end, J. Templin, Henson, S. Templin, and Roussos (2008) investigated whether the hierarchical modeling of the joint distribution of the latent skills is robust. In other words, the joint distribution of the attributes is modeled such that the latent attributes, $\alpha_i$s, are locally independent given a general aptitude for acquiring knowledge in a particular domain, $\theta$ (for an intuitive description of the higher order model ,see also de la Torre and Song, 2009). Templin et al. conducted a simulation study using the RUM with three different specifications of the skills association distribution namely: the multivariate normal distribution of the skills (general unconstrained model), a higher-order model of attribute associations, and an independent attributes model. Robustness was investigated in the simulation study where the data generating model was not necessarily the same as the higher order model assumed in the estimation of parameters. Templin et al. used two inter-attribute correlation matrix estimation procedures to evaluate the performance of the estimated attribute correlation matrices: MCMC EAP (expected a posteriori) whose values they named “chained estimates” and the bootstrap procedure in a Monte Carlo simulation of the MCMC estimates of examinee posterior probabilities of mastery, whose values they named “posterior estimates”.

The results showed that both the higher order model and the general method correlational approaches had absolute differences in the average correct classification rates of less than 1%. The independent attribute approach had consistently lower differences in rates of average correct classification ranging from 1% to 3%. In terms of the parameter estimation accuracy, the mean absolute deviation (MAD) for the $\pi^*$ parameter was slightly lower for the higher order model than for the other two approaches; however, MAD was approximately the same for the $r^*$ parameter for all three approaches (range was between .03 and .04). The population parameters, $P_k$, were fixed at .50 for the independent model whereas for the multivariate normal and the higher-order specifications of skill distribution, the $P_k$ parameters were simulated from a uniform distribution, $U$ (.50, .60). The reader should note that specifying the proportion of examinees that have mastery of a skill to be equal across all skills may be unrealistic in practice. Skills cannot be expected to be equally difficult; some will be easier and other more difficult. In the higher order and the general unconstrained model, MAD for the $P_k$ parameters was less than .05 indicating that both approaches performed similarly despite

MAD differences of .01 and .02 between the two approaches. As with the $P_k$ parameters, the correlation parameters of the independent model were not estimated and so only the results of the general unconstrained model and the higher order model are discussed. Both the chained and the posterior estimates showed that the general model had better accuracy in estimates when factors were not highly correlated while the higher order model had better estimates when the factors were highly correlated. This finding is to be expected because highly correlated factors are close to unidimensionality and hence a unidimensional higher order model should produce more accurate estimates than the general model.

These findings indicated that regardless of the true correlational structure of the attributes, the higher order model is very robust with respect to classification and item parameter accuracy and performed just as well as the general method using the multivariate normal model of the attributes. The independence model's poor performance in correct classification of examinees indicates that in most practical testing situations, at least a moderate positive association between the attributes is to be expected. To sum up, if the higher order model of the attributes performs as well as the multivariate normal model of the attributes, then the higher order model is to be preferred because it is parsimonious.

Another study modeling the joint distribution of the latent attributes using a higher–order model that compared several diagnostic models was conducted by de la Torre and Douglas (2008). In this study, de la Torre and Douglas compared the NIDA, the single strategy DINA, and a multiple strategy DINA model in classifying examinees while modeling the joint distribution of the attributes using a higher order latent model. The multiple-strategy DINA is an extension of the single-strategy DINA where each item has distinct strategies for solving that item; in other words, each strategy has a subset of K attributes used conjunctively to solve an item. Two simulation studies were conducted to investigate whether the MCMC( Markov chain Monte Carlo) algorithm used can effectively recover the model parameters (readers should refer to Patz and Junker (1999a, 1999b) for a detailed description of the MCMC algorithm). In simulation I, the NIDA model was used to generate the data that was fitted to both the NIDA and the single strategy DINA. Simulation II was done in two parts. First, the single strategy DINA was used as the data generating model where both the single and the multiple strategy DINA models were fitted. In the second part of simulation II, the only change was that the multiple strategy DINA was the data generating model. Model fit was assessed by use of various indices: Bayes factor (analogous to the likelihood ratio), BIC (Bayesian information criterion), AIC (Akaike information criterion), and DIC (Deviance information criterion). Parameter recovery was assessed using the mean correlation and the root mean square error (RMSE) between $\theta$ and $\hat{\theta}$ (due to the higher-order structure)·

The results of simulation I showed that when the NIDA was the generating model, correlations were higher and RMSE lower if the NIDA was fitted rather than the single strategy DINA, and vice versa. The mean correlation and RMSE between $\theta$ and $\hat{\theta}$ for the NIDA were .75 and .66, respectively; for the DINA, these values were .74 and .67, respectively. Percent of correct attribute classification for the five skills tested ranged from 94 to 96 % for the NIDA and 92 to 95% for the DINA. The results of simulation II were very similar to those of simulation I in terms of parameter estimation accuracy and correct classification rates indicating little effect of model misspecification. Although the NIDA model fit was good with simulated data (because the NIDA was the generating model in simulation I), that was not the case in the real data analysis probably due to the assumption in NIDA that the probabilities of slipping and guessing are the same across all items for each attribute. The –log-marginal likelihood, BIC, AIC, and DIC were about the same and lower for the single-strategy and the multiple-strategy DINA than for the NIDA. De la Torre and Douglas concluded that the single strategy DINA was to be preferred over the multiple-strategy DINA because both models fit the data well and in that case, the more parsimonious model should be chosen.

De la Torre and Douglas (2004) also investigated the DINA and the LLM (linear logistic model) models with a higher order structure. In a simulation study, de la Torre and Douglas investigated model misspecification and the performance of the MCMC algorithm in the estimation of the parameters. De la Torre and Douglas found that if the data were calibrated using the true generating model, then correct classification of the attributes was high. That is, if the DINA model was used to estimate parameters for data generated using the DINA, then the correct classification rate of the attributes was high (e.g., .88 to .97 for all five attributes tested). Perhaps more interesting is that negligible differences in classification occurred when the LLM was used to calibrate data generated using the DINA model and vice versa. The *kappa* statistics (chance-corrected index of agreement) between the two fitted models, DINA and LLM, regardless of the true generating model, were very high (in the .70s and .80s). The MCMC algorithm resulted in stable parameter estimates for both the DINA and the LLM model as assessed by examining the mean estimates and standard deviations over the 25 replications. For instance, none of the mean estimates of the DINA model deviated from the true values (simulated) by more than .02. De la Torre and Douglas concluded that although specifying the correct model resulted in better estimates ( higher correlations between $\theta$ and $\hat{\theta}$ and lower RMSE) , misspecified models did not deviate substantially in parameter estimates indicating that specifying the Q-matrix correctly is, perhaps, more important than identifying the correct response model. De la Torre and Douglas, however, cautioned that further studies were needed before reaching this conclusion.

Henson and Douglas (2005) were interested in test construction, specifically for cognitive diagnosis. In IRT models that rank order examinees on a continuous latent attribute, Fisher information is usually used to choose items that contain the most information because those are the highly discriminating items. Mathematically, Fisher information is the negative expectation of the second derivative of the log-likelihood for the parameter of interest. In CDMs that classify examinees into discrete latent classes (rather than a continuum), Fisher information can no longer be used. Hence, Henson and Douglas proposed a discrimination index based on the Kullback-Leibler information (which does not require a continuous space) that can be used to select items that are highly discriminating between masters and non-masters of a skill.

Kullback-Leibler information is thought of as a measure of the distance, *d*, between any two probability distributions, *f(x)* and *g(x)* because it ranges from 0 to ∞. Kullback-Leibler information is defined as

$$d[f,g] = E_f[\log[\frac{f(x)}{g(x)}]]. \tag{10}$$

$d[f,g]$ is the expectation with respect to *f(x)* (assuming that *f(x)* is the true distribution) of the log-likelihood ratio of any probability density functions *f(x)* and *g(x)*. In CDMs, the *item* that discriminates most between the attribute mastery pattern *α* and an alternative mastery pattern *α** has a large Kullback-Leibler information

$$d_j[\underline{\alpha},\underline{\alpha}^*] = E_\alpha[\log[\frac{p_\alpha(X_j)}{p_\alpha^*(X_j)}]] \tag{11}$$

where $p_\alpha(X_j)$ and $p_\alpha^*(X_j)$ are the probability distributions of $X_j$ given *α* and *α**, respectively. $X_j$ is a binary variable indicating whether an examinee gets item *j* right or wrong. Because Kullback-Leibler information only compares two attribute patterns whereas there are $2^k$ possible attribute patterns and because Kullback-Leibler is not symmetric ( $d[f,g] \neq d[g,f]$, a total of $2^k(2^{k-1})$ possible comparisons exist. These possible comparisons are better organized in a ($2^k$ x $2^k$) matrix, $\boldsymbol{D}_j$ with elements *u,v*, for the *j*th item,

$$D_{juv} = E_{\underline{\alpha}u}[\log[\frac{p_{\alpha u}(x_j)}{p_{\alpha v}(x_j)}]]$$

Attribute patterns that are similar are harder to distinguish and so they should be weighted more when computing a discrimination index. Using the inverse of the squared Euclidean distance,

$$h(\underline{\alpha},\underline{\alpha}^*) = \sum_{k=1}^{K}(\alpha_k - \alpha_k')^2$$

a weighted mean can be computed, the *CDI*$_j$, for a *single* item.

$$CDI_j = \frac{1}{\sum_{u \neq v} h(\underline{\alpha}_u, \underline{\alpha}_v)^{-1}} \sum_{u \neq v} h(\underline{\alpha}_u, \underline{\alpha}_v)^{-1} D_{.uv} \tag{12}$$

Kullback- Leibler information can also be computed at the test level. In that case, compute *CDI.* where $D_{.uv}$ in Equation 12 is replaced by $D_{.}$, the Kullback-Leibler information matrix for the whole test or  simply sum CDI$_j$ over the *J* items.

Henson and Douglas (2005) conducted a simulation study using the RUM and the DINA. Regardless of the correlation between the attributes (ρ=.0 or ρ=.5), both the DINA and the RUM model showed that the CDI index resulted in tests with better correct classification rates of examinees than the randomly generated tests using the same item bank. More specifically, for the DINA model with 4 attributes and ρ=.0, the correct classification rates (CCR) and the marginal CCR were 32.5% and 13.5% higher than for the randomly constructed tests, respectively. With 8 attributes, these values were 18.3% and 16% higher, respectively, than for the randomly generated tests. Similar results were found when ρ=.5. For the RUM model with 4 attributes and ρ=.0 the CCR was 26% higher than for the random tests. For the 8 attributes, the CCR was 26% higher than for the random tests. Similar findings resulted when ρ=.5.

Henson and Douglas (2005) concluded that the discrimination index (CDI) resulted in tests that yielded higher attribute classifications rates than tests that were randomly constructed. The Q-matrix is not used explicitly with the CDI$_j$ index and in this case, the Q-matrix is ignored which can cause the random tests to outperform the CDI$_j$

index. This problem can be corrected by constraining the Q-matrix such that items are selected depending on the constraints that are stipulated for the Q-matrix.

A study that explicitly examined the effects of the Q-Matrix misspecification on parameter estimates and skill classification accuracy using the DINA model was conducted by Rupp and Templin (2008). Rupp and Templin conducted a simulation study using the higher-order DINA with different misspecifications of the Q-matrix and compared parameter estimation and classification accuracy to the parameter and classification accuracy of the true Q-matrix. The misspecifications of the Q-matrix investigated were of two types: (a) underfitting (replacing 1s with 0s), overfitting (replacing 0s by 1s), or a balanced misfit of the Q-matrix ( exchanging 0s and 1s while controlling for the overall number of changes) for blocks of items that required a fixed number of attributes; and (b) incorrect dependency assumptions about two attributes (e.g., a suppression relationship where one attribute cannot occur in the presence of another or a conjunctive relationship where two attributes must co-occur). To check the accuracy of the parameter estimates, Rupp and Templin examined the estimates and the MAD values. Because the standard errors of the parameter estimates were less than .01 for almost all items, the parameter estimates were precisely estimated making it easy to spot the effects of the Q-matrix misspecification.

The results showed a clear indication of the Q-matrix misspecification in the different conditions studied. The effect of misspecification on the slipping and the guessing parameters of the DINA model was a local effect affecting only the items for which the Q-matrix was misspecified. Specifically, the slipping parameters were overestimated when attributes were incorrectly omitted in the Q-matrix while guessing parameters were overestimated when the attributes were unduly added to the Q-matrix. In particular, large values of the slipping and the guessing parameters can provide empirical evidence for Q-matrix misspecification. With respect to examinee classification, there was also evidence of misclassification when the Q-matrix was misspecified as measured by the global correlational measures (Kappa, Cramer's Φ, and contingency co-efficient) and individual cross-classification tables. Because the global correlational indices were not very informative, misclassification rates were computed by taking the total number of respondents within each attribute class divided by the total number of respondents that were classified into that attribute class under each condition of the simulation study. The misclassification was in the direction of either lacking the attribute or possessing the attributes depending on the changes that were made to the Q-matrix; that is, adding or deleting attributes required by an item. For instance, deleting certain attribute combinations such that no items measured those combinations resulted in complete misclassification of respondents possessing those combinations. Class-specific misclassification rates were computed such that a value of 100% indicated that no examinee was classified into a specific attribute class while 0% indicated that all examinees were classified into the correct attribute class. Smaller misclassification rates are desirable. The authors cautioned that the findings of this study would be strengthened by further studies such as, investigating Q-matrix misspecifications with other CDMs.

Only de la Torre (2008) has examined a method for validating the Q-matrix. In this study, de la Torre presented an empirically based method that he named the "delta method”. Suppose $K$ is the number of attributes, $\boldsymbol{\alpha}_l$ is the $2^K$ binary vectors defined by the $K$ attributes such that $l = 0,1, \dots, 2^k\text{-}1$, and $\boldsymbol{\alpha}_o$ corresponds to the null vector (0,0,…,0), and the q-vector is the row for an item $j$, then $\boldsymbol{q}_j$ is the correct Q-vector if

$$\boldsymbol{q}_j = \text{arg max}_{\alpha l}\, [P(X_j = 1|\eta_{ll'}{=}1) - P(X_j = 1|\eta_{ll'}{=}0)] = \text{arg max}_{\alpha l}\, [\delta_{jl}]$$

$$\text{for } l, l' = 1,2,\dots,2^{k-1}, \text{ where } \eta_{ll'} = \prod_{k=1}^{K} \alpha_{l'k}{}^{\alpha_{lk}}\,. \qquad (13)$$

Thus, the delta $\delta_j$ is the difference in the probabilities of the correct responses between examinees who have mastered a skill and those who have not mastered the skill. $\delta_j$ is therefore, a discrimination index of item $j$ in that items that are highly discriminating have high $\delta_j$ values while those that are not highly discriminating have low $\delta_j$ values. This $\delta_j$ changes as the $q$-vector of an item changes. An exhaustive search algorithm for computing $\delta_{jl}$ for each item becomes impractical when $K$ is large. Thus a sequential search algorithm is used as an alternative. The algorithm compares $\delta^*$ based on the single-attribute q-vectors. The attribute, say $\alpha^{(1)}$ that results in the highest $\delta^{(1)}$ must be a required attribute. The search then proceeds to q-vectors requiring two attributes. The attributes in the q-vector with the largest $\delta^{(2)}$ , $\alpha^{(1)}$ and $\alpha^{(2)}$ are the required attributes on condition that $\delta^{(2)} > \delta^{(1)}$. If the $\delta^{(2)} > \delta^{(1)}$ condition is not met, then the algorithm terminates and only $\alpha^{(1)}$ is required otherwise the process proceeds to three attributes and so on. In general, the algorithm terminates after step $s$ if $\delta^{(s)} < \delta^{(s-1)}$ , or when $s = K$ and the q-vector that corresponds to $\max(\delta^{(s-1)}, \delta^{(s)})$ identifies the correct attributes for the item. The sequential search is more efficient than the exhaustive search algorithm because the number of $\delta^*$that need to be computed is $(K_j + 1) – K^2{}_j + K_j\,)/2$, *where* $K_j$ is the correct number of attributes required for item *j*. The exhaustive search algorithm requires $2^k$-1 computations of $\delta^*$ . In real data applications, $\delta_j$ cannot be computed because the true slip and guessing parameters are unknown and also the distribution of the attributes is unknown. In addition, it is harder to have a clean separation between the groups, $\eta_j$ = 0 (non-masters) and $\eta_j$ = 1(masters). Thus observed values are used that can result in a recommended q-vector that has more attributes than necessary. A possible solution is the EM- algorithm with cut-off points. In an EM-based solution, a cut-off point, ε, for the minimum increment in the discrimination index of the item resulting from an additional attribute that is deemed relevant is determined using the criterion $\hat{\delta}_j^{(s)} - \hat{\delta}_j^{(s-1)} > \varepsilon$ .

Using the delta method in conjunction with the DINA model, de la Torre conducted a simulation study to investigate the delta method. When the correct Q-matrix was used, the parameter estimates showed little bias because the mean and the maximum biases were .01 and .04, respectively and δ = .61. Conversely, Q-matrix misspecification resulted in more bias in the parameters and shrunken δ. Five cut-off points ( ε = .00, .01, .05, .10, .20) were used to select the candidate *q*-vectors in the sequential EM- based δ-method. De la Torre found that the EM-based delta method can correctly replace the *q*-vectors that were misspecified in the Q-matrix with the correct ones while simultaneously retaining the *q*-vectors that had been correctly specified with the exception of the rather stringent cut-off ε = .20 . Therefore, the EM-based delta method can be used to check the appropriateness of the Q-matrix. De la Torre points out that although the delta method provides statistical information about the Q-matrix, it should not be used in isolation but in conjunction with substantive knowledge and domain expertise. In other words, the delta method for validating the Q-matrix does not replace theory.

Reliability as realized in classical test theory (CTT) also does not apply to cognitive diagnostic assessment and as such, new measures of reliability must be formulated. A common indicator of reliability in CDA is the correct classification rate (CCR) which is a measure of how accurately a CDM classifies examinees into the correct classes based on skill mastery. An index referred to as the cognitive diagnostic index (CDI, Equation 12) that has a close relationship to the CCR is proposed. Because the CCR is an indicator of reliability in cognitive diagnostic testing, the CDI can also be used to show the reliability of a diagnostic test in that a high correlation between CDI and CCR is an indication of reliability.

Henson, Roussos, Douglas and He (2008) extended the Henson and Douglas (2005) study that did not explicitly use the Q-matrix in designing the CDI. Specifically, they expanded the CDI to a set of indices that measure the discrimination power of an item for each *attribute*, thus making use of the Q-matrix. Henson et. al. (2008) conducted a simulation study to examine the relationship between the CDI and the CCR. From the CDI, they computed two attribute discrimination indices. The first index, $\mathbf{d}_{(\mathrm{A})\mathrm{j}}$, was for comparing attributes that differ by only one component and the second index, $\mathbf{d}_{(\mathrm{B})\mathrm{j}}$, was for comparing attribute patterns with differing likelihoods of occurrence. $\mathbf{d}_{(\mathrm{A})\mathrm{j}}$, was computed from the elements of $\mathbf{D}_j$ with the constraints that attribute comparisons differ only by the *k*th attribute(a situation that is harder to discriminate) while holding attribute mastery constant on the remaining ($K$-1) attributes. $\mathbf{d}_{(\mathrm{B})\mathrm{j}}$ takes the population characteristics into account, that is, the joint probabilities (or their estimates) of the attribute patterns are used to weight the appropriate elements of $\mathbf{D}_j$ giving those values for which α is more likely. The reduced RUM model without the Rasch component was used. Henson, Roussos, Douglas and He found high correlations (ranging from .92 to .97) across all the simulation conditions for both masters and non-masters between the CCR and the log transformed CDI attribute indices for the RUM and concluded that the discrimination indices are reasonable indicators of the correct classification rates and hence, reliability.

Furthermore, new methods of investigating person fit for cognitive diagnostic testing are also needed. Only one study addressed the issue of person fit in CDA. Liu, Douglas and Henson (2009) investigated person fit using two types of likelihood ratio tests adapted for cognitive diagnostic tests.  Define $\rho$ as the probability of responding in an aberrant manner. $\rho$ measures the magnitude of the tendency to act aberrantly beyond a person's true attributes $\boldsymbol{\alpha}$.  'A' is the type of inappropriate behavior exhibited. The two types of misfits examined were: (a) nonmasters who correctly respond to an item (spuriously high scorers) and (b) masters who fail to correctly answer an item (spuriously low scorers). $y_i$ is the dichotomous item response for examinee $i$. The first likelihood ratio test,

$$T_1 = -2\log LR = -2\log\frac{l_0(\hat{\alpha}_i, 0; y_i)}{l_A(\hat{\alpha}_i, \hat{\rho}_i; y_i)} \tag{14}$$

is based on the joint likelihood of the response pattern of person $i$, $l(\alpha_i, \rho_i, \bar{y}_i)$ whereas the second likelihood ratio test ,

$$T_2 = -2\log LR = -2\log \frac{L_0(0; y_i)}{L_A(\hat{\rho}_i; y_i)} \tag{15}$$

is based on the marginal likelihood, $l(\rho_i, \bar{y}_i)$. If a person fits well, $\rho_i = 0$. The hypotheses tested were $H_o$: $\rho_i = 0$ versus $H_A$: $\rho_i \neq 0$. The DINA model was used in the simulation study.

When the tests were long and examinees had strong tendencies for aberrant behavior, these two tests had adequate power to detect misfit. The type of misfit detected largely depended on the items' characteristics as well as the examinee's attribute pattern. For example, a spuriously low scorer is easier to detect if many items are easy or a particular examinee has mastered most of the attributes. Conversely, a spuriously high scorer is easier to detect when many items are difficult or many attributes are missing. These results indicated that the generalized likelihood ratio tests presented in the study were capable of detecting person misfit and moreover, they were robust against inaccurate estimates of the DINA model parameters.

Another study investigated potentially simpler methods of diagnosing skills that are more user-friendly. In an attempt to simplify the cognitive diagnostic testing application by classroom teachers, Henson, Templin, and Douglas (2007) proposed the use of model-based sum scores for diagnosing skill mastery. Henson et.al. state that using the sum-scores to estimate ability such as is done in IRT and factor analysis (item-level) suggests a possibility of generalizing sum-scores to cognitive diagnosis. They proposed three types of sum scores: simple (SSS), complex (CSS), and weighted complex (WCSS) sum scores and sought to determine if these sum-scores could classify examinees in the same manner as the Reduced RUM. To the extent that the sum-scores can approximate the actual CDM, then the sum-scores can be a simpler alternative for cognitive diagnosis when used with a known CDM.

The SSS is computed when skill-level structure is approximately simple structure. That is, an item measures only one skill and is included in the sum score for that skill only. If the Q-matrix is not simple structure, then the Reduced RUM determines the items to be included in each sum –score. Specifically, an item is included in the $k^{\text{th}}$ attribute's sum if the smallest $r^{*}_{jk}$ for the item is for that $k^{th}$ attribute and $q_{jk}$ = 1. The CSS is used when the simple structure does not hold. All the items with $q_{jk}$ = 1 for the *k*th attribute contribute to the *k*th sum-score. In a complex structure, an item can measure more than one skill and hence contribute to more than a single sum-score. The WCSS is a weighted sum score, unlike the first two. A weight is computed using the $r*_{jk}$ and the $\pi_j$ parameter of the RUM. The weight is in turn multiplied with the product of the elements of the Q-matrix and the examinee response matrix and summed over the *J* items. Mathematically, define $\delta jk = \pi^{*}_{j}(1 - r_{jk})$, the weight for a given examinee is $w_{ik} = \sum_{j=1}^{J} \delta_{jk}(q_{jk}x_{ij})$. Cut-off points $\lambda_k$ for each attribute are determined separately for each of the sum-scores. The chosen cut-offs are the ones that maximize the correct classification rate (CCR).

Henson, Templin, and Douglas concluded that when fewer attributes, for instance three attributes,  are tested, the SSS correctly classified examinees into classes but, the CSS and the WCSS should be used when the number of attributes is large ( e.g. eight attributes). Therefore, when model based estimation is not feasible, these sum scores can be used as approximations for skill mastery depending on the number of attributes tested.

Taking into consideration that diagnostic tests based on a cognitive model are still rare, sometimes diagnostic information is desired from an existing instrument that was designed to measure a unidimensional latent trait. In such a case, a method referred to as retrofitting, is used to extract diagnostic information from the unidimensional data. Assuming a hierarchical linear structure of the attributes, de la Torre and Karelitz (2009) examined the diagnosticity of the DINA and the 2PL (2-parameter-logistic) models through retrofitting in a simulation study. In other words, they investigated the ability of measurement models to extract diagnostic information based on the data-generating model. The DINA and the 2PL models were used to both generate and analyze the data. They defined diagnosticity as the degree to which the underlying discrete structures can be adequately represented by data. De la Torre and Karelitz defined three diagnosticity conditions: low discrimination condition (discrimination parameter *a* ~ U(.4, .8)), medium discrimination condition (designed to mirror a typical test), and high discrimination condition ( discrimination parameter *a* ~ U(1.6, 2.0)). To enable comparisons across the models, the item parameters of the 2PL and the DINA models were put on the same scale using the logistic-to-step transformation (LST) (see de la Torre & Karelitz, 2009 for details of the LST method). The overall model fit was assessed using the RMSD ( root mean squared deviation), defined as $\sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij}-E_{ij})^2}$

where $X_{ij}$ is the observed response and $E_{ij}$ is the model-based response, and the -2 log-likelihood.

The low RMSD and -2 log-likelihood values for the items with high diagnosticity indicated that there was slightly better model fit for these items regardless of the model fitted. It did not seem to matter much when items were of low diagnosticity. Item and person estimation accuracy was assessed using the RMSE, signed bias (to show overestimation or underestimation), and misclassification for an attribute and also for the whole attribute vector. The percent of misclassification was consistently lower for the high diagnosticity condition. Also, fitting DINA to CDM data provided more accurate estimates than fitting DINA to the 2PL data. De la Torre and Karelitz concluded that retrofitting a CDM to unidimensional data was not supported, more so when the items are of low diagnosticity. This result reinforces the need for tests to be constructed based on a cognitive model a priori because retrofitting may not be tenable under certain circumstances.

Finally, the absolute invariance of the DINA model parameters has been investigated by de la Torre and Young-Sun (2010). By modeling the joint distribution of the attribute vector using a saturated and a higher order formulation, de la Torre and Young-Sun investigated whether the parameters remained invariant across ability distributions. The saturated formulation means that the probabilities associated with each of the attribute combinations is considered such that $2^k$ probabilities are estimated. The higher order formulation is as previously defined. For the higher-order distributions, three ability distributions were specified with means $\mu_\theta$={-1.0, .0, 1.0} and a common standard deviation of $\sigma_\theta$=1.0. de la Torre and Young-Sun hypothesized that when the data-model fit was attained, there would be invariance across the attribute distributions but that would not be the case if the data did not fit the model.

The accuracy of the item parameter estimates was determined by the mean absolute bias (MAB). MAB was calculated by averaging the absolute difference between the true and the estimated parameters across the items. MAB for the guessing parameter ranged from .000 to .006. MAB for the slip parameters ranged from .000 to .002. With the simulated data, there was perfect model fit. The results of the simulation study confirmed that indeed, the DINA parameters were invariant only when the model fit the data well. Therefore, when the model is correctly specified and when samples with similar characteristics are analyzed, equating of tests is not necessary as the parameters are invariant.

## 2.6 Review of real data application studies using the DINA, NIDA, and the RUM

As in the previous section, the specific details of the real data application studies presented here are summarized in Table A.2 (in Appendix A). Table A.2 tabulates specific information about sample size, method of estimation, number of skills tested, and the correlational structure of the skills. Most of the simulation studies previously reviewed also involved a real data application to test whether the results of the simulation studies held with real data obtained in practice.

In addition to their simulation study, J. Templin, Henson, S. Templin, and Roussos (2008) also used a real data analysis to check whether the robustness exhibited by the higher order structure model would be found in real data. For confidentiality, Templin et al. used a test named test A in which they analyzed only the higher order structure and the general structure (multivariate normal model) of the latent attributes. This test was selected because the number of attributes and the structure of the Q-matrix matched those that were used in the simulation study. As in the simulation study, the chain and posterior estimates were used to evaluate the estimated attribute correlation matrices. The first eigenvalue of the three largest eigenvalues from both the chain and the posterior estimates accounted for over 90% of the variance. This was an indication of a strong single dimension underlying the data. Templin et al. concluded that test A had a strong single dimension underlying the latent attributes and hence the higher order model was a good fit. They added that the hierarchical one factor model provides a parsimonious structural parameterization in cognitive diagnostic testing that is robust to violations of the dimensionality of the proficiency space and so is applicable in many testing applications.

De la Torre and  Douglas (2008) used a subset of Tatsuoka's fraction subtraction data (Tatsuoka, 1983) to verify the findings in the simulation study that the single strategy higher order DINA model resulted in better estimates than the other two models. To determine model fit, de la Torre and Douglas examined the residuals of the observed indices, that is, the residuals of the proportion correct, log-odds ratio (see Equation 16 ), and Pearson product-moment correlation of the fifteen items in the subtraction data. The proportion correct refers to the number of examinees correctly responding to an item. The log-odds-ratio of item-pairs can be used to assess fit as it is a common measure of association between binary random variables. Using the estimated model parameters, the joint distribution for pairs of items can be computed and the log-odds-ratio for items *j* and *j'* is

$$\log[\frac{P(Y_j=1,Y_{j'}=1)P(Y_j=0,Y_{j'}=0)}{P(Y_j=1,Y_{j'}=0)P(Y_j=0,Y_{j'}=1)}] \qquad (16)$$

NIDA had higher mean absolute residuals than the other two models whose residuals were very similar. In addition, the BIC, AIC, and the DIC were used as global fit indices. These values were lower and very similar for the single strategy DINA and the multiple strategy DINA than for the NIDA. In the event that two models fit the data equally well, the more parsimonious model is favored. Findings using the fraction data indicated that the single strategy DINA was a better model than the NIDA and the multiple strategy DINA; thus corroborating the results of the simulation study. De la Torre and Douglas justified modeling the joint distribution of the attributes using a higher order model by stating that it was reasonable to assume the existence of a general ability in addition to the finer grained skills for the subtraction data.

De la Torre and Douglas (2004) also used Tatsuoka's subtraction data as an example of a real data application. Because the attributes were assumed conjunctive in nature, that is, all attributes had to be present for an examinee to obtain a correct response, only the DINA model was fit. Furthermore, NIDA showed poor fit and was subsequently dropped from the analyses. Two versions of the DINA were examined: A higher-order DINA and an independence DINA (no higher-order structure). Model fit was assessed using the mean absolute difference (MAD) between the observed and the expected log-odds ratios of an item averaged over the rest of the items (Equation 16); and also the Bayes factor. MAD was smaller for the higher-order model than for the independence model for all items except one (item 8 in Tatsuoka's data). For the overall fit, MAD across all item pairs was smaller for the higher-order DINA, .43, compared to the independence DINA model, .55. The log Bayes factor was 46.00 providing strong evidence for the higher-order model over the independence model.

Although the item parameters were very similar for the two models, the estimated proportion of examinees possessing specific attributes differed greatly between the two models in addition to the low classification agreement. This low agreement was attributed to noise in the real data. The results reinforce de la Torre and Douglas's warning that the minimal deviations in classification rates when the model was misspecified in the simulation studies may not hold in real settings. Therefore, when using CDMs in real settings, the correct specification of the model, in addition to the correct Q-matrix, is important.

In yet another study, de la Torre (2008) also used Tatsuoka's fraction subtraction data and the 2003 NAEP grade 8 mathematics data to test the delta method of checking the appropriateness of the Q-matrix using the DINA model. Model-data fit was assessed by checking that the sum of the mean guessing and slip parameters was low e.g. below .20 and so the fit was good. For the fraction-subtraction data, the delta-method can recognize and retain the correct *q*-vectors in most cases. For the NAEP data, the resulting parameter estimates had modest improvements; however, the recommended q-vectors from the delta-method could be used to establish or discard q-vector specifications, suggest alternative q-vectors, and confirm non-informative items for most of the q-vectors. Although the real data application results using both Tatsuoka's data and the NAEP data were favorable, de la Torre warns that the Q-matrix should be constructed through the combined efforts of experts in the domain of interest, substantive knowledge, and statistical information such as that yielded by the delta method. For instance, attribute specifications can result that run counter to substantive knowledge when only the statistical method is used to construct the Q-matrix.

Liu, Douglas and Henson (2009) investigated person fit using both the DINA and the RUM models. The data used were from the Examination for the Certificate of Proficiency in English (ECPE). Liu et al. found that the misfitting persons detected by the DINA model were in high agreement with the classifications that resulted from the RUM. Taking into consideration that the DINA and the RUM fit the data well as evidenced by the high agreement in classification rates, the two generalized likelihood ratio statistics performed similarly in both the simulation using the DINA model and the real data analysis using both the DINA and the RUM. In particular, for the DINA model, both the marginal and joint likelihood ratio tests flagged 10.4% and 14.0 % of the examinees, respectively, as having spuriously high scores. 10.6 % and 9.4% were flagged for having spuriously low scores. For the RUM, the two likelihood ratio tests flagged 11.5% and 11.5% respectively, as having spuriously high scores and 10.0% and 8.2% as having spuriously low scores. The Cohen's *Kappa* for testing the agreement between the results of the DINA model and the RUM was .87 and .79 for the marginal and joint likelihood ratio tests respectively (for the spuriously high scores). It was .87 and .82, respectively, for the spuriously low scores.

With respect to retrofitting, Junker and Sijtsma (2001) analyzed data from a set of transitive reasoning tasks, for instance, reasoning that $Y_A < Y_C$ from the premises $Y_A < Y_B$ and $Y_B < Y_C$ without guessing or using any other information. Responses to the tasks were obtained from 417 students in 2nd, 3rd, and 4th grades. Explanations for each answer were recorded. Junker and Sijtsma hypothesized that the total score on these tasks would not be informative if the objective was to know the particular transitive reasoning strategies that the students used in solving the problems. In effect, they fitted the DINA and the NIDA models and extracted structures at the cognitive attribute level from a test that was largely unidimensional (a procedure known as retrofitting). Estimation was carried out using BUGS and the Markov chain Monte Carlo (MCMC) algorithm. The study was largely a demonstration of the fact that *some* diagnostic information can be obtained from traditional tests that were not constructed for diagnosing skill mastery. Furthermore, it is not an issue of good model fit because good unidimensional model fit as indicated by, for instance Rasch fit statistics, does not indicate that there is diagnostic information other than giving confidence in the estimation of the total score. Therefore, if the goal of testing is cognitive diagnosis, discrete attribute models such as the DINA and the NIDA are more appropriate.

Finally, in the DINA parameter invariance study of de la Torre and Young-Sun (2010), Tatsuoka's fraction-subtraction data were also analyzed. Based on the number correct score, examinees were divided into three ability groups: high ability ( above the median score), medium  ability( at the median score), and low ability ( below the median score). The results showed that the guessing parameters from some items in the low ability group were underestimates whereas the guessing estimates from the high ability group for some of the items were overestimates. The Medium ability group was not affected. The trend between the low ability and the high ability groups reversed for the slip parameter: overestimates for the low ability group and underestimates for the high ability group. De la Torre and Young-Sun concluded that when real data were used, the invariance of the DINA parameters may not hold and the attribute distribution used heavily influenced the invariance in terms of which parameters were underestimated or overestimated. Again, noise in real data complicates generalizability of simulation findings to practical settings.

## 2.7 An Evaluation of the Reviewed Literature

With respect to the specifications of the simulation studies (see Table A.1), there are some commonalities.  In most of the studies, the ability parameters are sampled from the normal distribution while the item parameters are sampled from a uniform distribution. The ranges of these parameters are also consistent across the studies with only slight variations. Therefore, although little justification is provided for these particular distributions choices, it is reasonable to assume that the distributions and the parameter ranges are realistic based on the study findings. In addition, all the studies use large samples of examinees perhaps providing for more stable estimates. The number of skills studied range from 3 to 8, a preference that might indicate that CDA functions better with a smaller number of skills. To sum up the simulation specifications, the joint distribution of the latent attributes is modeled with either a higher order model or a multivariate normal model (see Harwell, Stone, Hsu, & Kirisci, 1996, for an introduction to Monte Carlo simulations with item response theory that generalizes easily to CDMs). For the real data analyses (see Table A.2), similar characteristics with the simulation studies are evident. Mostly, large samples of examinees are used; estimation is primarily accomplished in a Bayesian framework using the MCMC algorithm; the number of skills studied range from 3 to 8; and the joint distribution of the latent attributes is either higher order or multivariate normal.

Some general conclusions can be made from the studies reviewed here. First, the higher-order modeling of the joint distribution of the latent attributes seems to outperform other models. Therefore, the more parsimonious higher-order model can be used in real testing applications without serious concerns. J. Templin, Henson, S. Templin and Roussos (2008) found the higher order model to be a better fitting model- for both simulated data and real data – than the independence model of the attributes. Although the multivariate normal also fits equally well, it is not as parsimonious as the higher order model.  In addition, de la Torre and Douglas (2008) compared the single strategy DINA to the multiple strategy DINA and the NIDA model while modeling the attributes using a higher-order model. De la Torre and Douglas concluded that the higher order single strategy DINA was the best fitting model with both the simulation data and the real data. Although de la Torre and Douglas (2004) found that when the higher order structure was used with DINA and the LLM (linear logistic model) there was similar fit in the simulated data; that finding did not hold with real data. This discrepant finding between the simulated data and the real data led de la Torre and Douglas to conclude that when interest is in cognitive diagnostic information from real data, both the Q-matrix and the response model must be specified correctly.

Second, although some studies have been conducted in attempts to validate the Q-matrix, it is still the consensus among researchers that statistical methods should not be used in isolation in developing the Q-matrix. Theory and substantive expertise must be used in conjunction with the statistical methods. For instance, de la Torre (2008) investigated what he referred to as an EM-based delta method for validating the Q-matrix. Although, the delta method succeeded in flagging rows of the Q-matrix that needed revision in both the simulated data and the real data (NAEP and Tatsuoka's fraction data), de la Torre warns against sole reliance on such statistics and reiterates that the delta method does not replace theory when defining the Q-matrix. On the other hand, however, a careful examination of the parameter estimates and examinee misclassification can help point test developers to elements of the Q-matrix worth revising. For instance, Rupp and Templin (2008) state that large $s_j$ and $g_j$ parameters of the DINA model are an indication of Q-matrix misspecification.

Third, the cognitive diagnostic index (CDI) can be used both as an indicator of the discriminating power of an item as well as the reliability of a test. Henson and Douglas (2005) showed that using this index that is based on the Kullback-Leibler information resulted in tests that were better at diagnosing abilities than randomly constructed tests. In addition, Henson, Roussos, Douglas and He (2008) showed that two variant indices of the

CDI correlated highly with the CCR (correct classification rate). Thus, the CDI is also an indirect measure of reliability.

Fourth, other studies reviewed led to some other  important conclusions: (a) although more research is clearly needed to give further empirical evidence, Liu, Douglas, and Henson (2009) showed that two generalized likelihood ratio statistics ( for CDA) can detect misfitting examinees across both the DINA and the RUM using simulated data and also real data; (b) with respect to retrofitting, where a CDM is applied to unidimensional data to extract diagnostic information, de la Torre and Karelitz (2009) and Junker and Sijtsma (2001) concluded that retrofitting is a very limited endeavor. While some diagnostic information might be obtained, retrofitting seems untenable for most real data applications. Therefore, assessment developers must construct tests based on a specific cognitive model if the goal is to test examinees' strengths and weaknesses on a set of skills; (c) parameter invariance, at least with respect to the DINA model, holds with simulated data because the data fit the model perfectly. This invariance, however, does not hold with real data due to noise and so model fit must be evaluated and tests equated when real data are used. Finally, because CDMs tend to be very complex, test experts may endeavor to find simpler ways of providing diagnostic information that gives face validity to users of diagnostic information. To this end, Henson, Templin, and Douglas (2007) suggest the use of simple model based sum scores that can be used in lieu of CDMs and that are easier for the classroom teacher who lacks the technical expertise of the CDMS for easy implementation in the classroom. These sum scores are especially convenient when only a few skills (e.g. three) are to be measured.

In conclusion, it is evident that studies with respect to various aspects of the Q-matrix are needed. Only de la Torre has examined a method for validating the Q-matrix (de la Torre, 2008) and as previously stated, some exploratory techniques for determining the Q-matrix can and should be investigated. Although theory is an integral part of the development of the Q-matrix, an exploratory technique such as a factor analytic one can supplement theory in determining a reasonable Q-matrix for a set of items. In the following chapters, a components analysis that reparameterizes the DINA model into a component form is discussed in detail. Then, simulated and real data are used to investigate the components analysis both analytically and empirically, respectively. Results are then presented and a discussion of the findings concludes this thesis.

Chapter 3

# 3 Method

## 3.1 A Cognitive Diagnostic Model

The DINA model was investigated in this thesis. Define $X_{ij}$ = binary indicator [1,0] of whether examinee *i* performed task *j* correctly. 1 indicates a correct response, 0 an incorrect response. $q_{jk}$ = binary indicator [1,0] of whether attribute *k* is relevant for task *j*. 1 indicates that attribute *k* is required to correctly answer task *j* whereas 0 indicates that the attribute is not required for that item. $\alpha_{ik}$ = binary indicator [1,0] of whether examinee *i* possesses attribute *k*. 1 indicates possession whereas 0 indicates lack of the attribute. Then, the DINA has the item response function (IRF),

$$p(X_{ij}=1|\underline{\alpha},s,g)=(1-s_j)^{\xi_{ij}}\,g_j{}^{1-\xi_{ij}} \qquad (17)$$

where $\underline{\xi_{ij}}=\prod_{k=1}^{K}\underline{\alpha_{ik}}^{Q_{jk}}$ , $s_j = p(X_{ij}=0\,|\,\xi_{ij}=1)$, $g_j = p(X_{ij}=1\,|\,\xi_{ij}=0)$

In the DINA model, the latent responses are defined by $\xi_{ij}$. $\xi_{ij}$ is the deterministic part of the model and it is binary. $\xi_{ij}$ is 1, if and only if, all the skills required by an item have been mastered, otherwise it is 0. The stochastic nature of the DINA is modeled by the $s_j$ and the $g_j$ parameters to indicate that an examinee who has mastered all the skills required by an item may slip (hence the $s_j$ parameter) and lower the probability of a correct response. Conversely, an examinee who has not mastered an attribute required by an item may guess (hence the $g_j$ parameter) thus increasing the probability of a correct response.

### 3.1.1 Component Form of the DINA

The DINA model can be reparameterized into a component form. Due to the deterministic nature of the principal components analysis, the component form is the same as the factor model without the error term. It can be shown that the DINA IRF can be rewritten into a principal components form. Specifically, if $Z_{ij}$ is the z-score of person *i* on variable *j*, the scalar form of the factor model is

$$Z_{ij} = \sum_m \lambda_{jm} f_{im} + e_{ij} \tag{18}$$

where, $\lambda_{jm}$ is the loading of item *j* on factor *m*, $f_{im}$ is the factor score of person *i* on factor *m*, and $e_{ij}$ is the error term. It follows that the expectation of $Z_{ij}$ is the component form,

$$E(Z_{ij}) = \sum_m \lambda_{jm} f_{im} \tag{19}$$

where, $\lambda_{jm}$ is the loading of item *j* on component *m*, $f_{im}$ is the component score of person *i* on component *m*. The DINA IRF can be written in a similar manner,

$$p(X_{ij} = 1 | \underline{\alpha}, s, g) = \sum_m \lambda^*_{jm} f^*_{im} \tag{20}$$

The asterisks indicate that the parameters in Equation 20 are in the raw score model. The parameters in the standard scores model are as shown in Equation 19 and are as previously defined: $\lambda_{jm}$ is the loading of item *j* on component *m*, $f_{im}$ is the component score of person *i* on component *m.*

If the data satisfy the DINA model, then the data have a components solution obtained by setting $\lambda^*{}_{jm}$ equal to $(1-s_j)$ when the skill set corresponding to factor *m* is the minimal skill set sufficient for the solution of item *j*; otherwise it is zero. $f^*{}_{im}$ is either 1 or $(\frac{g_j}{1-s_j})$ depending on whether the examinee possesses the required skills for an item. With this reformulation,

$$p(X_{ij}=1\,|\,\underline{\alpha},s,g)=\sum\nolimits_m \lambda^*_{jm} f^*_{im}=\lambda^*_{jm'} f^*_{im'} = (1-s_j)\ (\frac{g_j}{1-s_j})$$

where factor *m’* is the factor corresponding to the minimal skill set required for item *j*.

The parameters in the raw scores model (Equation 20) can be related to those in the standard score model (Equation 19) as follows:

$$\lambda_{jm}=\frac{\lambda^*_{jm}\hat{\sigma}^*_m}{\hat{\sigma}_j} \quad \text{and} \quad f_{im}=\frac{f^*_{im}-f^*_{.m}}{\hat{\sigma}^*_m}$$

where, $\hat{\sigma}_j$ is the standard deviation for item *j*, $\hat{\sigma}^*_m$ is the standard deviation of the factor scores for a component *m*, and $f^*_{.m}$ is the mean component score for factor *m*. $\hat{\sigma}^*_m$ appears in both the numerator and the denominator respectively of $\lambda_{jm}$ and $f_{im}$ because the variance of the components is 1 for standardized scores. In Equation 19, a component corresponds, not necessarily to a single skill, but a skill set and the items loading on the component are those for which the set is the minimally required combination of skills for the items. Generally, items measuring different skill sets will load on different components whereas items measuring the same skill sets will load on the same component. If the ratio, $(\frac{g_j}{1-s_j})$, is the same for all of the items requiring a given skill set, then all the items will load on the same component . If this condition is not met, a given skill set may define more than one component with items measuring the same skill set loading on different components. This situation is not likely to be present because the slip and the guess parameters are usually small with small variations but, the reader should be aware that items measuring the same skill sets may load on different components under certain circumstances.

Accordingly, the component score $f^*_{im}$ will be 1 if person *i* possesses all the skills required to correctly execute item *j*. If an examinee lacks at least one of the skills, the component score $f^*_{im}$ should be close to zero (unstandardized factor scores, in the raw score model, range from 0 to 1 whereas standardized factor scores range from - ∞ to + ∞). The principal components should follow a simple structure in that every item will load on one and only one rotated component. As an example, if I had a test measuring three skills, the loadings of the items measuring these skills, for all of the combinations of the three skills would be as shown in Table 3.1,

Table 3.1: Components Loadings

| | | **Loadings (dichotomized)** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Skill Combinations (Components)** | Items Measuring Skill Set 1 | Items Measuring Skill Set 2 | Items Measuring Skill Set 3 | Items measuring Skill Set4 | Items Measuring Skill Set 5 | Items Measuring Skill Set 6 | Items measuring Skill Set 7 |
| Item1 | 1 | $1-s_1$ | 0 | 0 | 0 | 0 | 0 | 0 |
| | | | | | | | | |
| Item2 | 2 | 0 | $1-s_2$ | 0 | 0 | 0 | 0 | 0 |
| | | | | | | | | |
| Item3 | 3 | 0 | 0 | $1-s_3$ | 0 | 0 | 0 | 0 |
| | | | | | | | | |
| Item4 | 12 | 0 | 0 | 0 | $1-s_4$ | 0 | 0 | 0 |
| | | | | | | | | |
| Item5 | 13 | 0 | 0 | 0 | 0 | $1-s_5$ | 0 | 0 |
| | | | | | | | | |
| Item6 | 23 | 0 | 0 | 0 | 0 | 0 | $1-s_6$ | 0 |
| | | | | | | | | |
| Item7 | 123 | 0 | 0 | 0 | 0 | 0 | 0 | $1-s_7$ |

Using the component form of the DINA model, the resulting component scores can be represented as shown in Table 3.2.

The analysis involves conducting a principal components analysis with rotation on examinee response data (in binary form). A scree plot can serve as an indicator of the initial number of components to extract. Extraction of components with rotation continues until simple structure is achieved. The rotated component solution with simple structure will correspond to a specific skill or skill set as shown in Table 3.1. In other words, the rows in Table 3.1 correspond to components. The first row for component 1 corresponds to the skill measured by items measuring skill 1 only. Likewise, row 7 is component 7 with all the items that measure skills 1, 2, and 3 loading on it. The same logic applies to the intermediate components. To determine the probability of a correct response, take, as an example, an item that requires skill 1 for correct execution. In Table 3.1 skill 1 corresponds to component 1. If an examinee possesses that skill, then the probability that they get the item correct is ((1-$s_j$)*1) using the information from the first cell in both Tables 3.1 and 3.2. Likewise, if the examinee does not possess the skill, then the probability that they execute the item correctly is ((1-$s_j$)* $(\frac{g_j}{1-s_j})$ ) which amounts to guessing.

## 3.2 The Analysis (Principal Components Analysis)

### 3.2.1 Exploratory Phase

Binary examinee responses from both the simulated data and the real data were used to compute item correlations that were analyzed through principal components analysis (PCA) with promax rotation (kappa=4) to identify the skill sets. Components were extracted until every item loaded on one and only one rotated component. A Q-matrix based on the components analysis was then constructed. The process of constructing the Q-matrix is discussed in more detail under each study in later chapters. Generally, however, the Q-matrix is constructed by conducting a content analysis of the items loading on each component, determining the skills measured by the items, and then constructing an item by skill Q-matrix.

Table 3.2: Component Scores of People Varying in their Skill Sets

| Component Scores | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Skill Combinations Possessed** | Component 1 | Component 1 | Component 3 | Component 4 | Component 5 | Component 6 | Component 7 |
| Person1 | (1) | 1 | $\frac{g_2}{1-s_2}$ | $\frac{g_3}{1-s_3}$ | $\frac{g_4}{1-s_4}$ | $\frac{g_5}{1-s_5}$ | $\frac{g_6}{1-s_6}$ | $\frac{g_7}{1-s_7}$ |
| Person2 | (2) | $\frac{g_1}{1-s_1}$ | 1 | $\frac{g_3}{1-s_3}$ | $\frac{g_4}{1-s_4}$ | $\frac{g_5}{1-s_5}$ | $\frac{g_6}{1-s_6}$ | $\frac{g_7}{1-s_7}$ |
| Person3 | (3) | $\frac{g_1}{1-s_1}$ | $\frac{g_2}{1-s_2}$ | 1 | $\frac{g_4}{1-s_4}$ | $\frac{g_5}{1-s_5}$ | $\frac{g_6}{1-s_6}$ | $\frac{g_7}{1-s_7}$ |
| Person4 | (1,2) | $\frac{g_1}{1-s_1}$ | $\frac{g_2}{1-s_2}$ | $\frac{g_3}{1-s_3}$ | 1 | $\frac{g_5}{1-s_5}$ | $\frac{g_6}{1-s_6}$ | $\frac{g_7}{1-s_7}$ |
| Person5 | (1,3) | $\frac{g_1}{1-s_1}$ | $\frac{g_2}{1-s_2}$ | $\frac{g_3}{1-s_3}$ | $\frac{g_4}{1-s_4}$ | 1 | $\frac{g_6}{1-s_6}$ | $\frac{g_7}{1-s_7}$ |
| Person6 | (2,3) | $\frac{g_1}{1-s_1}$ | $\frac{g_2}{1-s_2}$ | $\frac{g_3}{1-s_3}$ | $\frac{g_4}{1-s_4}$ | $\frac{g_5}{1-s_5}$ | 1 | $\frac{g_7}{1-s_7}$ |
| Person7 | (1,2,3) | $\frac{g_1}{1-s_1}$ | $\frac{g_2}{1-s_2}$ | $\frac{g_3}{1-s_3}$ | $\frac{g_4}{1-s_4}$ | $\frac{g_5}{1-s_5}$ | $\frac{g_6}{1-s_6}$ | 1 |

The probability of a correct response is then simply $\sum_f \lambda f$ .

### 3.2.2 Confirmatory Phase

In addition to the exploratory phase, one study involving real data rather than simulated data included a confirmatory phase. In the confirmatory analysis, the Q-matrix developed in the exploratory phase was evaluated using a cross-validation sample to determine the adequacy of the reconstructed Q-matrix in terms of item parameter estimation accuracy and correct classification of examinees. Liu, Douglas, and Henson (2009) have also argued that, in addition to the accuracy of item parameter estimation, measures of model fit such as, comparison of model predicted versus observed data characteristics can provide evidence of a reasonable Q-matrix. In this thesis, however, I evaluated the accuracy of parameter estimation and the accuracy of examinee classification to determine if the reconstructed Q-matrix outperformed the original one.

## 3.3 Computer Programs

R, freely available software, was used to simulate the data for the simulation studies; SPSS version 18 (SPSS, 2009) was used to run the principal components analysis in the exploratory phase; and Latent Gold 4.5 by Vermunt and Magidson (2007) was used for all the statistical analyses in the confirmatory phase. The computer programs used to simulate and analyze the data are presented in the appendix B.

Chapter 4

# 4 Study I: Illustrative Analysis with Simulated Data

Study I was analytical. An illustrative example using simulated data was used to illustrate how the components analysis method performs under ideal conditions; ideal in the sense that the number of skills and the items measuring those skills are known and the test includes the right combination of items to measure those skills. Generally, examinees' response data generation for cognitive diagnosis requires 1) a Q-matrix, 2) item parameters, and 3) person parameters. Data were generated using the DINA model. For this illustration, the Q-matrix was  fixed (known) so the components results could be compared with the known *Q*-matrix.   The test was simulated as one designed to measure *K*=3 skills with 21 items. In the DINA model, at least $2^K$ *-1* items are required to measure *K* skills. I simulated two sufficient skills (k1 and k2) and one insufficient skill (k3) as shown in the Q-matrix in Table 4.1. As previously stated, a sufficient skill is defined as a skill that appears alone in one or more rows of the *Q*-matrix but may also appear in combination with other skills in some rows whereas an insufficient skill is a skill that never appears alone in a row but rather, in combination with other skills. It can be seen from Table 4.1 that k1 appears alone and is the only skill required by item 1.  Hence, k1 is sufficient for item 1.  However, k1 also appears in combination with other skills as in items 10 and 14.   Skill k2 appears both as a single skill (item 6) and also in combination with other skills (e.g. items 10 and 18, hence is labeled here as a sufficient skill because it is sufficient for at least some of the items, items 6 to 9. Skill K3, however, does not appear alone in any row of the matrix,  and is always in combination with either k1 or k2 hence the label insufficient skill.

Each skill level combination in the Q-matrix was measured by at least four items but no items measured all of the three skills because, in practice, it is uncommon for an item to measure all the skills measured in a diagnostic test. As an example, in Tatsuoka's fraction subtraction data (to be discussed in chapter 5), there are no items that measure all the skills. In addition, some simulation analyses (not shown) suggest that such items are problematic in component analyses in that it is difficult to achieve simple structure when such items are included. Based on the Q-matrix in 4.1 there would be five excepted components (skills or skill sets) from a components analysis corresponding to all the combinations of the skills: (1), (2), (1,2), (1,3, and (2, 3), the skill sets represented in 4.1. )

The item parameters, $s_j$ and $g_j$, were simulated from a random, uniform distribution with the specifications, $s_j \sim U$ (.02, .05) and $g_j \sim U$ (.05, .25). Examinees' latent ability vectors, $\underline{\alpha_i}$, were simulated from a probit model having underlying latent variables that are multivariate normal with mean vector zero and correlations between the skills fixed at .50 such that all the off-diagonal elements in the correlation matrix were .50; that is, $MVN$ ($\underline{\mu}$, $\mathbf{\Sigma}$) where $\underline{\mu}$ is the mean vector and $\mathbf{\Sigma}$ is the correlation matrix. The proportion of the population assumed to have mastered the attributes, $p_k$, was fixed at .50. To dichotomize the continuous latent vector to indicate whether an examinee has mastery of a skill or not, the p$_k$ parameter is converted into a z-score , z$_k$, under the standard normal distribution ( mean=0, standard deviation=1). Then the elements of the latent vector are changed to 1(one) if they are greater than or equal to $z_k$ otherwise they are changed to 0 (zero). These parameter specifications align with those commonly published in journal articles but more importantly, they are only used for illustration purposes in this illustration. The data were simulated to mirror a real cognitive diagnostic testing situation as closely as possible. Principal components analysis was used to analyze the generated examinee responses (n = 4000) generated using the Q-matrix in 4.1 and the DINA model.

Table 4.1: Simulated Q-matrix for 12 Items and 3 Skills

| | K1 | K2 | K3 |
|---:|---:|---:|---:|
| Item1 | *1* | *0* | *0* |
| Item2 | *1* | *0* | *0* |
| Item3 | *1* | *0* | *0* |
| Item4 | *1* | *0* | *0* |
| Item5 | *1* | *0* | *0* |
| Item6 | 0 | 1 | 0 |
| Item7 | 0 | 1 | 0 |
| Item8 | 0 | 1 | 0 |
| Item9 | 0 | 1 | 0 |
| Item10 | *1* | *1* | *0* |
| Item11 | *1* | *1* | *0* |
| Item12 | *1* | *1* | *0* |
| Item13 | *1* | *1* | *0* |
| Item14 | 1 | 0 | 1 |
| Item15 | 1 | 0 | 1 |
| Item16 | 1 | 0 | 1 |
| Item17 | 1 | 0 | 1 |
| Item18 | *0* | *1* | *1* |
| Item19 | *0* | *1* | *1* |
| Item20 | *0* | *1* | *1* |
| Item21 | *0* | *1* | *1* |

## 4.1 Results

The resulting 5 - component solution is shown in Table 4.2 whereas the correlations between the components are shown in Table 4.3. Notice that the correlations between components vary and are not equal to the .50 correlation specified for skills for the data generation. This is to be expected because the components do not necessarily represent single skills; rather, components are skill sets. Because the data were simulated with a known number of skills, the expected number of components that would obtain simple structure was 5. If the number of skills is not known, a scree plot can be used to indicate the initial number of components to be extracted. Extraction then continues with additional components until the rotated solution displays simple structure. In Table 4.2 there are 5 skill combinations hence 5 components as expected. For a better understanding of the recovery of the original simulated Q-matrix, the Q-matrix must be examined simultaneously with the 5-component solution in Table 4.2. The component matrix in Table 4.2 contains one component for each skill set represented in the *Q*-matrix of Table 4.1. More specifically, component 1 corresponds to skill 1 because only the items measuring skill 1 only load on it (see the Q-matrix in Table 4.1). The same logic applies to the other components: component 2 represents skill 2; component 3 represents skills 1and 2; component 4 corresponds to skills 2 and 3; and finally, component 5 corresponds to skill 3. Because skills 1 and 2 are sufficient skills, these two skills have a corresponding component in the components solution in Table 4.2. On the other hand , skill 3 which is insufficient does not have a corresponding component.

Table 4.2: Component Solution of the Simulated Data with Items Loading on a Component in Bold

| | Rotated Components | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| item1 | **.851** | -.032 | .026 | .018 | .061 |
| item2 | **.83** | .01 | -.009 | -.01 | .013 |
| item3 | **.855** | .008 | .016 | -.014 | -.026 |
| item4 | **.871** | .023 | -.049 | 0 | -.035 |
| item5 | **.848** | .001 | .033 | -.006 | .005 |
| item6 | .006 | **.881** | -.017 | .038 | .03 |
| item7 | -.006 | **.927** | .044 | -.081 | -.057 |
| item8 | .002 | **.893** | -.018 | .028 | .037 |
| item9 | .014 | **.881** | -.026 | .03 | .034 |
| item10 | .025 | -.009 | -.027 | .015 | **.846** |
| item11 | .041 | .033 | .021 | .002 | **.794** |
| item12 | .017 | .038 | .016 | .003 | **.811** |
| item13 | -.026 | .009 | .006 | -.016 | **.862** |
| item14 | .114 | -.038 | **.762** | .085 | .022 |
| item15 | .016 | -.032 | **.797** | .051 | .021 |
| item16 | -.029 | .018 | **.872** | -.042 | -.029 |
| item17 | -.024 | .028 | **.842** | -.074 | .005 |
| item18 | .006 | .089 | .029 | **.745** | -.006 |
| item19 | .012 | -.128 | -.101 | **.911** | .032 |
| item20 | -.009 | .02 | .013 | **.788** | -.022 |
| item21 | -.031 | .114 | .086 | **.726** | -.018 |

Table 4.3: The Correlations of the Rotated Components from Simulated Data

| | **Component Correlation Matrix** | | | | |
|---|---|---|---|---|---|
| | ***1*** | ***2*** | ***3*** | ***4*** | ***5*** |
| ***1*** | 1 | 0.293 | 0.606 | 0.314 | 0.607 |
| ***2*** | 0.293 | 1 | 0.316 | 0.597 | 0.61 |
| ***3*** | 0.606 | 0.316 | 1 | 0.54 | 0.531 |
| ***4*** | 0.314 | 0.597 | 0.54 | 1 | 0.525 |
| ***5*** | 0.607 | 0.61 | 0.531 | 0.525 | 1 |

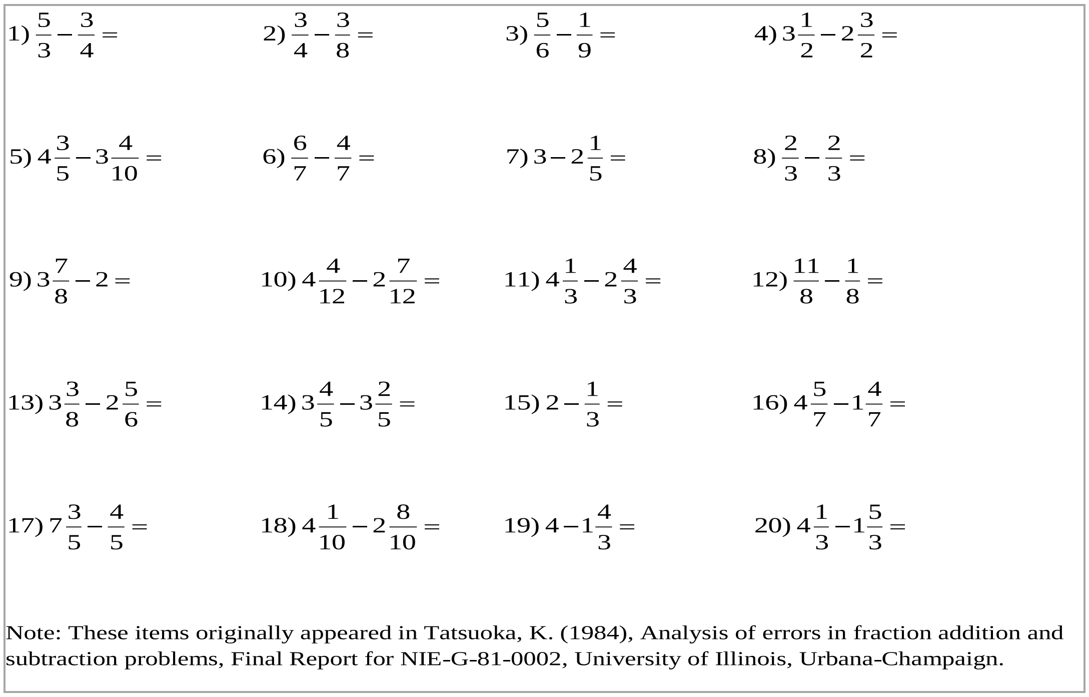
A large sample size can influence the results of a components analysis resulting in a clean simple structure that may not be present with smaller sample sizes. In addition, small samples are common in real testing situations especially with cognitive diagnostic assessments. As such, a random sample of n=572 was drawn from the original sample with n=4000. The components analysis method was applied to the random sample and the resulting rotated component solution was identical to that of Table 4.2 and thus, that solution is not included here. The recovery of the original Q-matrix with a sample of n=572 indicates that the components analysis method will be applicable to smaller sample sizes than that of this simulation study if the right conditions are in place; that is, if there are multiple items measuring each skill set.

The components in Table 4.2 represent the structure of the items, but there is a major conceptual difference between the dimensions of the *Q*-matrix and the components. The items loading on dimension *k* in the *Q*-matrix (i.e. $q_{jk}$=1) is the set of items for which dimension *k* is necessary, but may not be sufficient. The items loading on a component are the items which share a common skill set, a skill set that is both necessary and sufficient for the items. *Q*-matrix dimensions and components differ in two respects. First, *Q*-matrix dimensions correspond to single skills, whereas the components correspond to skill sets containing one or more skills. Second, a *Q*-matrix skill is necessary, but may not be sufficient for solution of the item, whereas the component skill set is both necessary and sufficient for solution of an item loading on the component. In practice, the components solution can be used to identify items sharing a common skill set, by interpreting that component substantively, which involves identifying the skill set associated with the items loading on the component. Identifying that skill set will involve the use of theory, substantive expertise, and a task analysis of items loading on the component as will be illustrated in the real data example that follows.

Chapter 5

# 5 Study II: Tatsuoka's Fraction Subtraction data

Study II was an empirical analysis applying the components analysis method to real data. The real data used were the fraction subtraction data collected by Dr. Kikumi Tatsuoka (Tatsuoka, 1983) that were designed to be diagnostic of students' strengths and weaknesses. The data were obtained from the Royal Statistical Society website. The data set is comprised of dichotomously-scored responses to 20 fraction subtraction test items from 536 middle school students. According to de la Torre and Douglas (2004), these items measure the following eight attributes: 1) Convert a whole number to a fraction, 2) Separate a whole number from a fraction, 3) Simplify before subtracting, 4) Find a common denominator, 5) Borrow from whole number part, 6) Column borrow to subtract the second numerator from the first, 7) Subtract numerators, and 8) Reduce answers to the simplest form. The items are shown in Figure 5.1.



| | | | |
|---|---|---|---|
| 1) $\frac{5}{3}-\frac{3}{4}=$ | 2) $\frac{3}{4}-\frac{3}{8}=$ | 3) $\frac{5}{6}-\frac{1}{9}=$ | 4) $3\frac{1}{2}-2\frac{3}{2}=$ |
| 5) $4\frac{3}{5}-3\frac{4}{10}=$ | 6) $\frac{6}{7}-\frac{4}{7}=$ | 7) $3-2\frac{1}{5}=$ | 8) $\frac{2}{3}-\frac{2}{3}=$ |
| 9) $3\frac{7}{8}-2=$ | 10) $4\frac{4}{12}-2\frac{7}{12}=$ | 11) $4\frac{1}{3}-2\frac{4}{3}=$ | 12) $\frac{11}{8}-\frac{1}{8}=$ |
| 13) $3\frac{3}{8}-2\frac{5}{6}=$ | 14) $3\frac{4}{5}-3\frac{2}{5}=$ | 15) $2-\frac{1}{3}=$ | 16) $4\frac{5}{7}-1\frac{4}{7}=$ |
| 17) $7\frac{3}{5}-\frac{4}{5}=$ | 18) $4\frac{1}{10}-2\frac{8}{10}=$ | 19) $4-1\frac{4}{3}=$ | 20) $4\frac{1}{3}-1\frac{5}{3}=$ |

Note: These items originally appeared in Tatsuoka, K. (1984), Analysis of errors in fraction addition and subtraction problems, Final Report for NIE-G-81-0002, University of Illinois, Urbana-Champaign.

Figure 5.1:Tatsuoka's Fraction Subtraction Items

The examinee response data to the fraction subtraction items were analyzed using components analysis. A random sample of n=136 was used in the exploratory phase (components analysis) and the remaining n = 400 was used in the confirmatory phase to test the reconstructed Q-matrix against the original Q-matrix. The original Q-matrix as presented in de la Torre and Douglas (2004) is shown in Table 5.1.

## 5.1 Results

The analysis resulted in a simple structure solution comprised of the 13 components shown in Table 5.2.To reconstruct the Q-matrix using the 13 - component solution, the substantive content of each of the items loading on a component was examined to determine the skill or skill set corresponding to each component. The content analysis suggested the following skills and skill sets corresponding to the components:

Table 5.1: Original Q-matrix for the Fraction Subtraction Items (de la Torre & Douglas, 2004)

| | skills | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | K1 | K2 | K3 | K4 | K5 | K6 | K7 | K8 |
| Item 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| Item 2 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| Item 3 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| Item 4 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| Item 5 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| Item 6 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Item 7 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| Item 8 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Item 9 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Item 10 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| Item 11 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| Item 12 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| Item 13 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| Item 14 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| Item 15 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Item 16 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| Item 17 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| Item 18 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 |
| Item 19 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| Item 20 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |

Table 5.2: Components Solution for the Fraction Subtraction Data with Items Loading on a Component in Bold

| | The Rotated Matrix | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Component | | | | | | | | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| item1 | .21 | **.841** | .071 | .064 | .087 | -.176 | -.035 | -.037 | -.053 | -.046 | -.115 | .211 | -.073 |
| item2 | -.045 | **.831** | -.045 | -.082 | -.012 | .165 | -.017 | .103 | -.034 | .06 | .065 | -.027 | .077 |
| item3 | -.184 | **.959** | .028 | -.099 | -.077 | .016 | .005 | -.04 | .049 | .109 | .09 | .003 | .083 |
| item4 | .102 | .055 | .026 | .036 | .102 | -.024 | .039 | -.072 | .066 | -.053 | **.899** | -.096 | .014 |
| item5 | .025 | .018 | .012 | .012 | .046 | -.015 | .023 | **.967** | .014 | .007 | -.073 | -.022 | .015 |
| item6 | -.002 | .145 | .003 | .064 | .01 | .07 | .056 | -.02 | .193 | -.144 | -.085 | **.850** | -.009 |
| item7 | .052 | .15 | **.648** | .193 | -.021 | .073 | -.011 | .022 | -.193 | .175 | .071 | -.052 | -.177 |
| item8 | -.053 | -.02 | -.01 | -.019 | **.990** | .009 | -.022 | .045 | .006 | -.011 | .096 | .01 | -.011 |
| item9 | -.081 | -.108 | .009 | **.986** | -.018 | .014 | -.021 | .011 | .035 | .048 | .033 | .066 | .088 |
| item10 | **.846** | .05 | .091 | .034 | -.081 | .049 | -.004 | .055 | .158 | -.08 | .039 | -.26 | .005 |
| item11 | **.940** | -.019 | -.15 | -.03 | -.108 | -.052 | -.028 | .078 | -.091 | -.074 | .219 | .229 | -.059 |
| item12 | -.019 | -.034 | .028 | -.023 | -.023 | -.014 | **.981** | .024 | -.008 | .02 | .041 | .062 | -.012 |
| item13 | .092 | .102 | -.007 | .093 | -.012 | -.049 | -.012 | .016 | -.017 | -.083 | .015 | -.011 | **.910** |
| item14 | .04 | .008 | -.006 | .014 | .008 | **.988** | -.013 | -.013 | .036 | -.098 | -.024 | .069 | -.046 |
| item15 | .034 | .082 | **1.007** | -.008 | .03 | -.044 | .064 | -.03 | .109 | -.231 | -.067 | -.065 | .018 |
| item16 | .053 | -.015 | -.003 | .035 | .007 | .038 | -.008 | .013 | **.788** | .217 | .067 | .215 | -.018 |
| item17 | .111 | .112 | -.066 | .047 | -.01 | -.088 | .016 | .006 | .186 | **1.003** | -.046 | -.142 | -.076 |
| item18 | **.574** | .022 | -.034 | .027 | .108 | .115 | .118 | -.091 | -.168 | .272 | -.117 | -.009 | .148 |
| item19 | -.029 | -.189 | **.693** | -.144 | -.061 | .002 | -.068 | .059 | -.027 | .244 | .11 | .201 | .124 |
| item20 | **.802** | -.121 | .161 | -.119 | .076 | .011 | -.047 | -.062 | .095 | .135 | -.096 | .014 | .057 |

*Note:* Extraction Method: Principal Component Analysis. Rotation Method: Promax with Kaiser Normalization.

Component 1: borrowing, subtracting numerators, and subtracting whole numbers | Component 2: finding a common denominator and subtracting numerators | Component 3: borrowing and subtracting numerators | Component 4: subtracting whole numbers | Component 5: subtracting a number from itself | Component 6: subtracting whole numbers, subtracting numerators, and subtracting a number from itself | Component 7: subtracting numerators and putting fraction into a proper form | Component 8: finding a common denominator, subtracting whole numbers, subtracting numerators, and putting fraction into a proper form | Component 9: subtracting numerators and subtracting whole numbers | Component 10: borrowing, subtracting whole numbers, and subtracting numerators |Component 11: borrowing and subtracting a number from itself | Component 12: subtracting numerators | Component 13: finding a common denominator, borrowing, subtracting numerators, and subtracting a number from itself.

Two examples to illustrate the skill or skill set corresponding to  a component are presented here. In the first example, the items loading on component 1 are items 10, 11, 18, and 20 (see Figure 5.1). These items are similar, with the first numerator being smaller than the second numerator e.g., item 10 is $4\frac{4}{12}-2\frac{7}{12}=?$.To solve such an item, a student must employ three skills: first, borrow from the first whole number (4 in item 10) to add to the first numerator (4 in item 10 which then becomes 16), second ,subtract the numerators (16- 7), and finally, subtract the whole numbers (3-2).  As a second example, items 1, 2 and, 3 load on component 2.  These items, e.g., item 1 ($\frac{5}{3}-\frac{3}{4}=?$)

require that an examinee employ two skills: first, find a common denominator and second, subtract the numerators. For item 1, the common denominator is 12. Take the common denominator divide by each denominator and then multiply by each numerator to get the equation into the form where subtraction of numerators is possible. In this case, the form would be $\frac{20-9}{12}=\frac{11}{12}$. Similarly, it can be shown that the items loading on the other components require the skills or skill sets that correspond to the components under which they load. All items loading on a component require the same skill set. Items loading on different components require different skill sets.

From the results of the content analysis of the each component, six skills were determined. The process involves writing down all the skills corresponding to each component and then making a list of all skills associated with at least one component. The result is a set of skills that is the union of all skill sets associated with a component. The skills in the fraction subtraction data were 1) borrowing, 2) subtract numerators, 3) subtract whole numbers, 4) find a common denominator, 5) subtract a fraction from itself, and 6) put fractions into proper form. Some of the skills are sufficient skills such as those corresponding to components 4 (subtracting a whole number), 5 (subtracting a number from itself), and 12 (subtracting numerators). All of the other skills are insufficient skills because they only appear in combination with at least one sufficient skill. One example of an insufficient skill is *borrowing*. The skills with their corresponding components are in shown in Table 5.3. The reconstructed Q-matrix is shown in Table 5.4 indicating that the components analysis method uncovered six skills compared to the original eight skills in de la Torre and Douglas (2004). More specifically, skill 1(convert a whole number to a fraction), skill 3(simplify before subtracting), and skill 5(borrow from whole number part) in the original Q-matrix of de la Torre are subsumed under skill 1 (borrowing) in the reconstructed Q-matrix. Skill 2 (Separate a whole number from a fraction) does not correspond to a single dimension in the reconstructed Q-matrix but rather, items load across different dimensions. Skill 4 (find a common denominator) is exactly the same as in the reconstructed Q-matrix (skill4). Skill 6 (Column borrow to subtract the second numerator from the first) is subsumed under subtracting numerators in the reconstructed Q-matrix (skill 2). Skills 7 (subtract numerators) and 8 (Reduce answers to the simplest form) are the same as in the reconstructed Q-matrix (skills 2 and 6, respectively) with the exception that some items are excluded in each skill in both Q-matrices.

Table 5.3: Components by Skills Table Marked by x

| | borrowing (skill1) | subtract numerators (skill2) | subtract whole numbers (skill3) | find a common denominator (skill4) | subtract a fraction from itself (skill5) | put fraction into proper form (skill6) |
|---|---|---|---|---|---|---|
| comp. 1 | x | x | x | 0 | 0 | 0 |
| comp. 2 | 0 | x | 0 | x | 0 | 0 |
| comp. 3 | x | x | 0 | 0 | 0 | 0 |
| comp. 4 | 0 | 0 | x | 0 | 0 | 0 |
| comp. 5 | 0 | 0 | 0 | 0 | x | 0 |
| comp. 6 | 0 | x | x | 0 | x | 0 |
| comp. 7 | 0 | x | 0 | 0 | 0 | x |
| comp. 8 | 0 | x | x | x | 0 | x |
| comp. 9 | 0 | x | x | 0 | 0 | 0 |
| comp. 10 | x | x | x | 0 | 0 | 0 |
| comp. 11 | x | 0 | 0 | 0 | x | 0 |
| comp. 12 | 0 | x | 0 | 0 | 0 | 0 |
| comp. 13 | x | x | 0 | x | x | 0 |

*Note:* comp. means component

Table 5.4: Reconstructed Q-matrix for the Fraction Subtraction Items

| | skills | | | | | |
|---|---|---|---|---|---|---|
| | K1 | K2 | K3 | K4 | K5 | K6 |
| Item 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| Item 2 | 0 | 1 | 0 | 1 | 0 | 0 |
| Item 3 | 0 | 1 | 0 | 1 | 0 | 0 |
| Item 4 | 1 | 0 | 0 | 0 | 1 | 0 |
| Item 5 | 0 | 1 | 1 | 1 | 0 | 1 |
| Item 6 | 0 | 1 | 0 | 0 | 0 | 0 |
| Item 7 | 1 | 1 | 0 | 0 | 0 | 0 |
| Item 8 | 0 | 0 | 0 | 0 | 1 | 0 |
| Item 9 | 0 | 1 | 1 | 0 | 0 | 0 |
| Item 10 | 1 | 1 | 1 | 0 | 0 | 0 |
| Item 11 | 1 | 1 | 1 | 0 | 0 | 0 |
| Item 12 | 0 | 1 | 0 | 0 | 0 | 1 |
| Item 13 | 1 | 1 | 0 | 1 | 1 | 0 |
| Item 14 | 0 | 1 | 1 | 0 | 1 | 0 |
| Item 15 | 1 | 1 | 0 | 0 | 0 | 0 |
| Item 16 | 0 | 1 | 1 | 0 | 0 | 0 |
| Item 17 | 1 | 1 | 1 | 0 | 0 | 0 |
| Item 18 | 1 | 1 | 1 | 0 | 0 | 0 |
| Item 19 | 1 | 1 | 0 | 0 | 0 | 0 |
| Item 20 | 1 | 1 | 1 | 0 | 0 | 0 |

In order to compare the two Q-matrices, the original one in Table 5.1 and the reconstructed one in Table 5.4, a confirmatory analysis was conducted in Latent Gold (see code in Appendix B) to compare the models in terms of model fit, examinee classification, and parameter estimation accuracy, and parameter plausibility. The parameter estimates for the two models are shown in Table 5.5. These estimates were computed using the formulas in Decarlo (2011) for converting the regression parameters from Latent Gold to the DINA parameters, $s_j$ and $g_j$. A brief summary of how to transform the slipping and the guessing parameters and compute the corresponding standard errors is shown in Appendix C.

Table 5.5: Parameter Estimates using the DINA Model

| Original Q-Matrix | | | | | Reconstructed Q-Matrix | | | |
|---|---|---|---|---|---|---|---|---|
| item | gj | standard error | 1-sj | standard error | Gj | standard error | 1-sj | standard error |
| 1 | .05 | .01 | .89 | .01 | .03 | .02 | .87 | .02 |
| 2 | .06 | .01 | .97 | .01 | .03 | .02 | .97 | .01 |
| 3 | .01 | .00 | .88 | .01 | .01 | .01 | .86 | .02 |
| 4 | .21 | .01 | .89 | .01 | .11 | .03 | .83 | .03 |
| 5 | .35 | .01 | .84 | .01 | .33 | .04 | .83 | .03 |
| 6 | .44 | .02 | .97 | .00 | .19 | .05 | .97 | .01 |
| 7 | .03 | .01 | .79 | .01 | .10 | .02 | .70 | .03 |
| 8 | .43 | .02 | .88 | .01 | .13 | .18 | .76 | .03 |
| 9 | .35 | .06 | .69 | .01 | .39 | .05 | .77 | .03 |
| 10 | .03 | .01 | .82 | .01 | .04 | .01 | .78 | .03 |
| 11 | .06 | .01 | .93 | .01 | .07 | .02 | .92 | .02 |
| 12 | .36 | .02 | .94 | .01 | .11 | .04 | .93 | .03 |
| 13 | .01 | .00 | .67 | .02 | .01 | .01 | .65 | .04 |
| 14 | .29 | .02 | .97 | .01 | .10 | .04 | .96 | .02 |
| 15 | .02 | .01 | .9 | .01 | .13 | .03 | .80 | .03 |
| 16 | .33 | .02 | .93 | .01 | .16 | .04 | .91 | .02 |
| 17 | .03 | .01 | .87 | .01 | .02 | .01 | .87 | .03 |
| 18 | .13 | .01 | .83 | .01 | .12 | .02 | .82 | .03 |
| 19 | .03 | .00 | .77 | .02 | .02 | .01 | .61 | .04 |
| 20 | .02 | .00 | .84 | .01 | .01 | .01 | .80 | .03 |

A comparison of the guessing parameters obtained from the two Q-matrices indicates that the reconstructed Q-matrix results in somewhat smaller estimates. The estimated guessing parameters from the original Q-matrix range from .01 to .44 whereas the range is from .01 to .39 in the reconstructed Q-matrix. The items were constructed response items so the guessing parameters should be small. The slip parameters, however, seem to be slightly higher for the reconstructed Q-matrix with 1-sj ranging from .61 to .97 compared with .67 to .97 in the original Q-matrix. In general, these parameter estimates are very similar and their standard errors are also small indicating good parameter estimation accuracy. It is safe, therefore, to conclude that both Q-matrices resulted in stable estimates.

The latent class sizes are shown in Table 5.6. A scrutiny of the estimates of the latent classes shows that the proportion of examinees who have mastered the skills in the reconstructed Q-matrix is as can be logically expected. For instance, 95% of the examinees possess skills 5 and 6 which are subtracting a fraction from itself and putting a fraction into proper form, respectively. This is to be expected because those are, presumably, lower level skills. Moreover, skill 1 (borrowing) is the hardest skill with only 59% of the students possessing it, which aligns with the notion that it is a higher level skill than the other 5 skills. The reconstructed Q-matrix, therefore, resulted in latent class sizes that correspond to what would be logically expected given the difficulty level of the skills. On the other hand, the original Q-matrix resulted in some latent class sizes that seem less plausible. As an example, skill 7, subtracting numerators appears to be a higher level skill with latent class size of .67 than skill 6 which has a latent class size of .98 and involves column borrowing to subtract a second numerator from the first. Clearly, there should be more examinees possessing skill 7 than the more complex combination of skills in skill 6 but the original Q-matrix fails to capture that. These discrepancies are discussed in more detail in DeCarlo (2011).These results suggest that the latent class sizes for the reconstructed Q-matrix align more closely with the expected difficulty of the skills than the original Q-matrix.

Table 5.6: Estimates of the Latent Class Sizes for the Skills in the Original Q-matrix and the Reconstructed Q-matrix

| | α1 | α2 | α3 | α4 | α5 | α6 | α7 | α8 |
|---|---|---|---|---|---|---|---|---|
| Original Q-matrix ( de la Torre and Douglas, 2004) | .74<br>(.03) | .93<br>(.02) | .95<br>(.02) | .87<br>(.02) | .75<br>(.03) | .98<br>(.01) | .67<br>(.02) | .94<br>(.02) |
| | α1 | α2 | α3 | α4 | α5 | α6 | | |
| Reconstructed Q-matrix | .59<br>(.03) | .79<br>(.02) | .93<br>(.02) | .76<br>(.03) | .95<br>(.03) | .95<br>(.03) | | |

Although the parameters estimates and the latent class estimates favor the reconstructed Q-matrix, the global fit indices, the AIC and the BIC, indicating how well a Q-matrix fits the data shown in Table 5.7 do not. The differences are rather small but more importantly, the issue in measurement of examinees' mastery of skills is not primarily model fit but rather how well the examinees are classified. In addition, studies such as Jiang et.al (2008) have called into question the appropriateness of global fit indices (e.g., AIC and BIC) in the evaluation of model fit in mixture models.  Therefore, the magnitude of the fit indices is of less concern here. Another interesting observation is that the latent class sizes are not similar to the .50 specification that is usually used in simulation studies.

Table 5.7: Fit Indices

| | Fit Statistics | |
|---|---|---|
| | BIC | AIC |
| Original Q-matrix ( de la Torre and Douglas, 2004) | 7036.15 | 6844.56 |
| Reconstructed Q-matrix | 7174.66 | 6991.05 |

Moreover, the classification accuracy statistics indicate the reconstructed Q-matrix performs better than the original Q-matrix. These statistics are as shown in Tables 5.8 and 5.9 for the original and the reconstructed Q-matrix, respectively. In the original Q-matrix the classification errors range from .01 to .13 whereas in the reconstructed Q-matrix, the range is from .01 to .06.

Table 5.8: Classification Statistics for the Original Q-matrix

| Classification Statistics | α1 | α 2 | α 3 | α 4 | α 5 | α 6 | α 7 | α 8 |
|---|---|---|---|---|---|---|---|---|
| Classification errors | .1295 | .0482 | .0384 | .0676 | .118 | .0168 | .029 | .053 |
| Reduction of errors (Lambda) | .4972 | .3332 | .2086 | .4738 | .5237 | .0669 | .9114 | .1662 |
| Entropy R-squared | .4976 | .4645 | .3046 | .5471 | .5401 | .317 | .8886 | .3217 |
| Standard R-squared | .5077 | .3946 | .2348 | .5163 | .5402 | .1353 | .901 | .2533 |

Table 5.9: Classification Statistics for the Reconstructed Q-matrix

| Classification Statistics | α 1 | α 2 | α 3 | α 4 | α 5 | α 6 |
|---|---|---|---|---|---|---|
| Classification errors | .0665 | .0171 | .0359 | .0683 | .0433 | .0412 |
| Reduction of errors (Lambda) | .8386 | .9201 | .4616 | .7112 | .1257 | .166 |
| Entropy R-squared | .7712 | .9199 | .5741 | .7169 | .4404 | .3877 |
| Standard R-squared | .7963 | .9238 | .5175 | .7202 | .2758 | .2838 |

Chapter 6

# 6 Study III: NAEP Data

Study III was the second empirical study and used the NAEP (National Assessment of Educational Progress) 2003 grade 8 mathematics data. Because of test security only data for items that are publicly available were analyzed. These data are available at http://nces.ed.gov/nationsreportcard/itmrlsx/s508/Result.aspx . The items are from blocks 6, 7, and 10 of the 2003 grade 8 mathematics data. The student responses to these items were obtained from the National Center for Education Statistics (NCES). The NAEP math test contains both constructed response and multiple choice items organized into separate timed blocks of questions but, only the responses to the multiple choice items were analyzed in this study. Analysis was conducted block by block because there were different examinees for each block of items even though the students are usually randomly assigned to blocks and hence no systematic differences between the test takers exist. Block 6 had 21 items and sample size n = 31542; block 7 had 14 items and n = 31420; and block 10 had 10 items with n = 31588. Because the analysis was originally intended to involve both an exploratory and a confirmatory phase, the data were split into two random samples of equal halves within each block. One half was used in the exploratory phase and the half for the confirmatory phase.

The NAEP 2003 grade 8 mathematics test was designed to measure five content strands: Number Sense, Properties, and Operations; Measurement; Geometry and Spatial Sense; Data Analysis, Statistics, and Probability; Algebra and Functions. Because the NAEP math test is not designed for cognitive diagnostic assessment, the first step was to explore the data to determine the number of underlying skill sets (not content strands) for the items. In an exploratory components analysis scree plots for each of the three blocks of items were obtained. These are shown in Figures 6.1, 6.2, and 6.3.

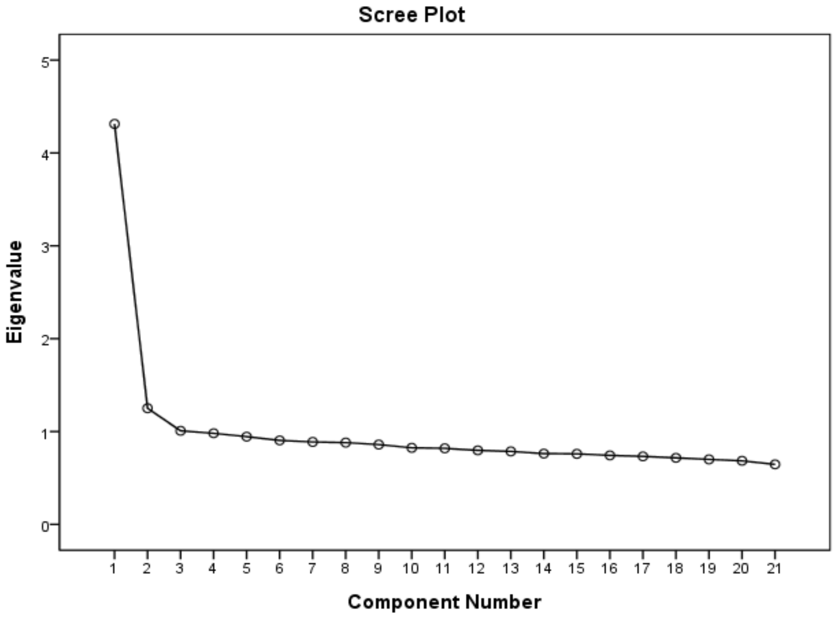


Figure 6.1: Scree Plot for Block 6 NAEP Math Items

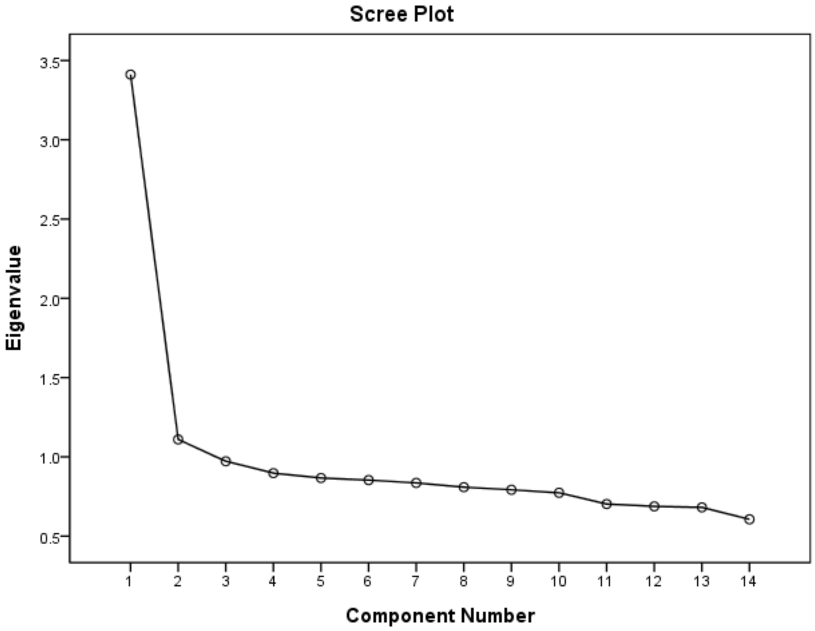


Figure 6.2: Scree Plot for Block 7 NAEP Math Items

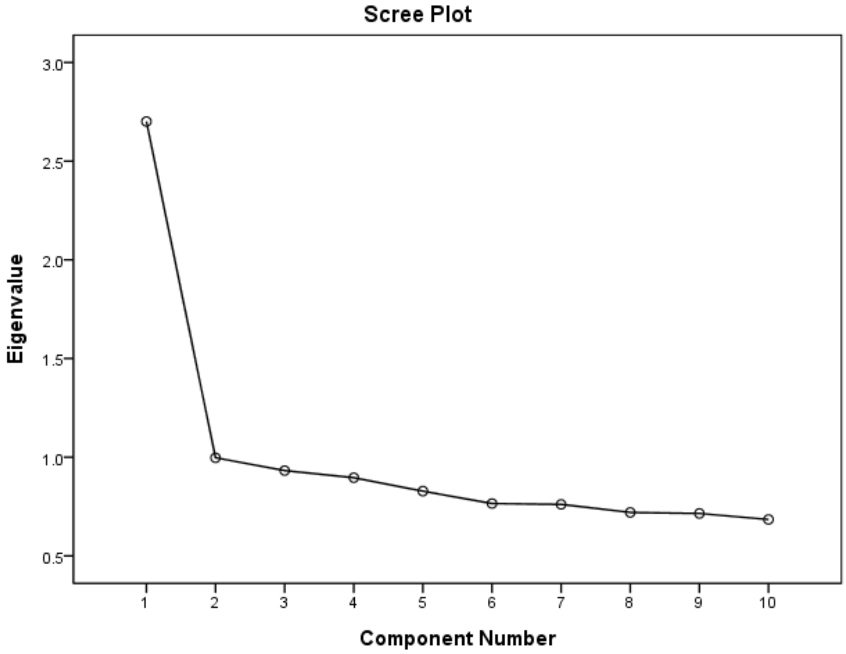


Figure 6.3: Scree Plot for Block 10 NAEP Math Items

As can be seen from the scree plots, there was mainly one discernible component in each block without a clear elbow to determine the number of components to be extracted. Therefore, I examined the first differences (here called the pimple method) where each subsequent eigenvalue is subtracted from the previous one, and these first differences are examined for peaks indicating a relatively large first difference. The reader should keep in mind that the PCA, unrotated solution, is used to compute the first differences. These differences are then plotted to better visualize peaks (pimples) in the differences. The differences are shown in Figures 6.4, 6.5, and 6.6.

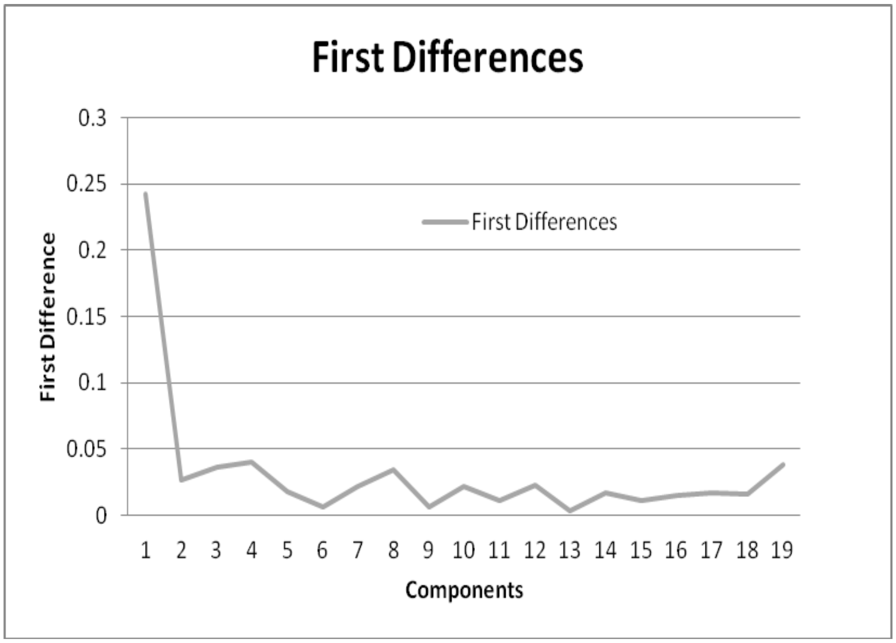


Figure 6.4: First Differences for block 6 NAEP Math Items

Figure 6.4 shows that there are peaks at components 4, 8, 10, 12, and 14. These correspond to 5, 9, 11, 13, and 15 components, respectively, after adding 1 because I omitted the first difference in the plot (it is usually too large and distorts the visual quality of the plot). After extracting the five sets of components, only the 13-component solution resulted in simple structure. Further attempts to extract more components than 13 resulted in a distorted solution. The 13- component solution for block 6 items is shown in Table 6.1. Although the simple structure is evident, the solution was uninterpretable based on the content analysis of the items (The items are attached in Appendix D.) Of the 13 components, 11 components had single items loading on them which suggests that those 11 components correspond to skill sets or skills possessed by only one item. Principal components become unreliable when a test is comprised of many single items that have a unique skill combination. In addition, items that loaded on the first two components, which were the only components with multiple items loading on them, did not appear to be measuring the same skills. As an example, component 1 had items 5, 6, 7, 11, 12, and 20 (see Appendix D). Item 5 involves recognizing the final image after flipping the original image; item 6 involves calculating change (money); item 7 involves computing a total amount; item 11 requires balancing a scale; item 12 involves recognizing which pattern would fold into a cube; and item 12 involves applying equations to a set of conditions to determine the correct answer. Clearly, the items are not measuring the same skill set and therefore, do not make conceptual sense in determining the skill set corresponding to component 1. The same applies to component 2. This may indicate that the components analysis method may not be suited to all situations and there are circumstances when it does not work. This may be a limitation of the method.

Because the PCA method did not uncover a sensible rotated solution, the Q-matrix for the analyzed items could not be reconstructed and no confirmatory analysis could be conducted. As previously described, only two components had multiple items loading them and all of the other components had single items loading on them resulting in a situation with many skill combinations with only a single item measuring them. Principal components become unreliable in such circumstances. An examination of the items in blocks 7 and 10 also showed the same inconclusive results. The components solutions are not shown here but the first differences are shown in Figures 6.5 and 6.6.

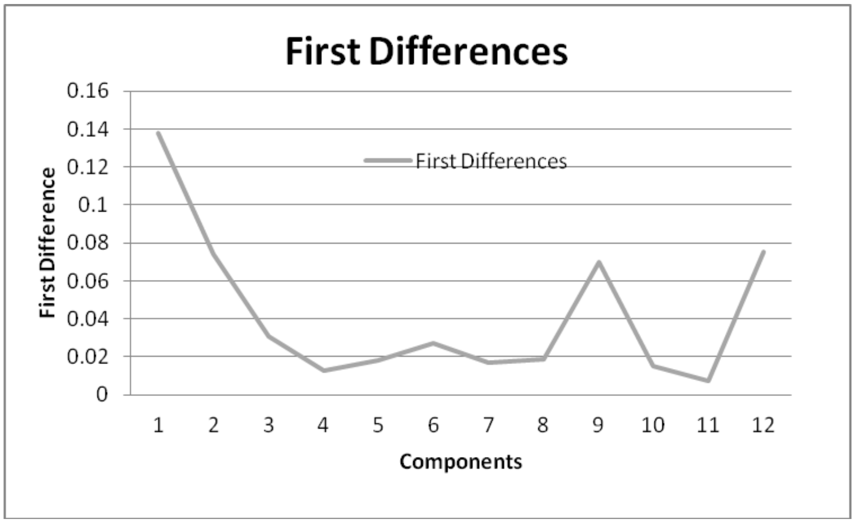


Figure 6.5: First Differences for Block 7 NAEP Math Items

From Figure 6.5, it seems that a 7- component and a 10-component solution might be appropriate. The rotated solutions indicated that both the 7-component and the 10-component solutions achieved simple structure, however, a content analysis of the items that loaded on each component indicated that the components were uninterpretable. Likewise, the first differences from block 10 (Figure 6.6) suggest extraction of 4 and 7 components. The 7-component solution yielded a better simple structure solution after rotation but again, it was uninterpretable. As a result, no confirmatory analysis could be conducted.

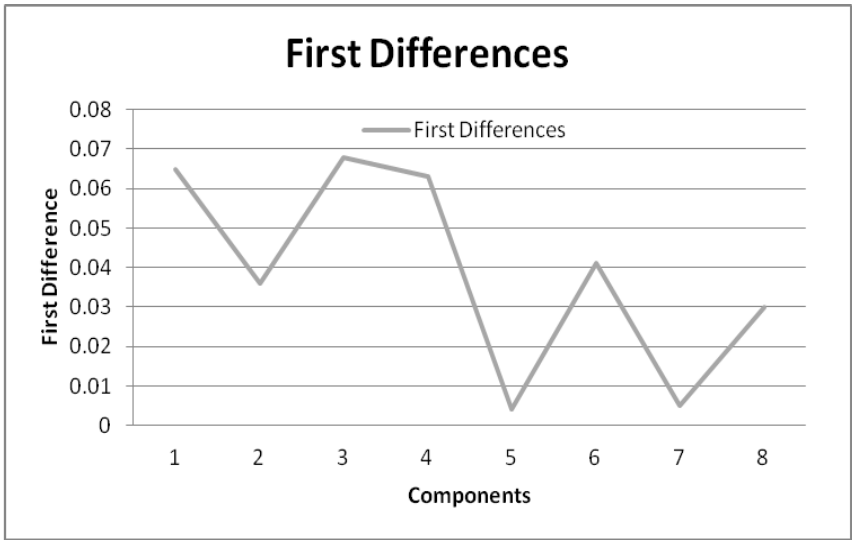


Figure 6.6: First Differences for Block 10 NAEP Math Items

Table 6.1: 13-Component solution for block 6 NAEP Math Items

| | Component | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| item1 | -.023 | -.015 | .035 | 0 | **.997** | .003 | .006 | .005 | -.008 | .016 | -.015 | .012 | -.003 |
| item2 | .008 | .005 | -.013 | .017 | .003 | **.995** | -.002 | -.003 | -.002 | -.003 | .006 | -.017 | .008 |
| item3 | .012 | .005 | -.038 | .013 | .005 | -.003 | .004 | **.995** | .002 | .009 | -.001 | -.011 | .008 |
| item4 | -.004 | -.002 | .015 | .01 | -.008 | -.002 | .002 | .002 | **.997** | .004 | -.007 | -.003 | .005 |
| item5 | **.663** | -.037 | .111 | .029 | -.016 | -.033 | -.022 | -.002 | -.026 | -.05 | .049 | -.081 | .002 |
| item6 | **.646** | -.055 | .04 | -.014 | -.067 | .043 | .069 | -.011 | -.003 | .115 | -.085 | .047 | -.022 |
| item7 | **.628** | -.017 | .001 | .118 | -.037 | -.03 | -.008 | .005 | -.026 | .066 | .012 | -.02 | .036 |
| item8 | .041 | .014 | -.053 | .058 | .012 | -.017 | .013 | -.011 | -.003 | -.019 | .029 | **.955** | .021 |
| item9 | 0 | -.023 | .027 | -.037 | -.003 | .008 | .007 | .008 | .005 | -.01 | -.014 | .022 | **.998** |
| item10 | -.024 | -.027 | .04 | **.987** | 0 | .016 | .003 | .013 | .009 | -.008 | -.015 | .058 | -.037 |
| item11 | **.643** | .009 | -.004 | -.095 | .004 | .017 | .012 | .017 | -.014 | .027 | -.085 | .148 | .029 |
| item12 | **.395** | .179 | -.077 | .148 | .051 | -.018 | -.052 | -.015 | .002 | -.061 | .11 | -.145 | .123 |
| item13 | -.039 | **.725** | -.249 | .018 | .025 | -.01 | .056 | -.057 | .026 | .096 | .017 | -.011 | .042 |
| item14 | .024 | .017 | .027 | -.009 | .016 | -.003 | -.019 | .009 | .004 | **.972** | .025 | -.019 | -.011 |
| item15 | .263 | **.604** | -.073 | .009 | .013 | -.011 | .045 | -.004 | -.008 | -.058 | -.109 | -.046 | -.071 |
| item16 | -.068 | **.576** | .282 | -.01 | -.022 | .01 | -.174 | .011 | -.004 | .006 | .015 | .112 | .01 |
| item17 | .084 | .005 | **.882** | .04 | .035 | -.013 | .058 | -.037 | .015 | .025 | -.009 | -.056 | .025 |
| item18 | -.009 | .038 | .053 | .003 | .006 | -.002 | **.966** | .004 | .002 | -.019 | .025 | .013 | .007 |
| item19 | -.116 | **.679** | .204 | -.057 | -.042 | .019 | .07 | .059 | -.019 | -.029 | .031 | .002 | -.026 |
| item20 | **.659** | -.005 | .005 | -.134 | .064 | .026 | -.047 | .017 | .066 | -.088 | .096 | .05 | -.089 |
| item21 | .015 | -.018 | -.008 | -.014 | -.015 | .006 | .025 | -.001 | -.007 | .025 | **.981** | .029 | -.013 |

Chapter 7

# 7 Study IV: MDE Data

Study IV was the last empirical study involving real data. The 2006 4th grade mathematics data were obtained from the Minnesota Department of Education (MDE). For security reasons, some items were removed from the math test. In addition, constructed response items were removed because only the multiple choice items were analyzed in this study. The 2006 4th grade math test was designed to measure four specific content strands: Number Sense; Patterns, Functions, and Algebra; Data, Statistics, and Probability; Spatial Sense, Geometry, and Measurement. There were 39 multiple –choice items selected with a sample size of n = 5653. For purposes of any confirmatory analysis, the sample was randomly split into an exploratory sample n = 1503 and a cross-validation sample of n = 4150.

As with the NAEP data, the MDE 4th grade math test was not designed for cognitive diagnostic purposes. Therefore, an exploration of the number of skills measured by the 39 items was the initial step. The screeplot is shown in Figure 7.1. The screeplot did not give a clear indication of the number of components that needed to be extracted for the MDE data. As a result, I examined the plot of the first differences shown in Figure 7.2. There were many peaks making it impossible to determine a plausible component solution for these data. As such, the components analysis with rotation could not be conducted. Because the MDE items are not publicly available, they are not included in the appendix.

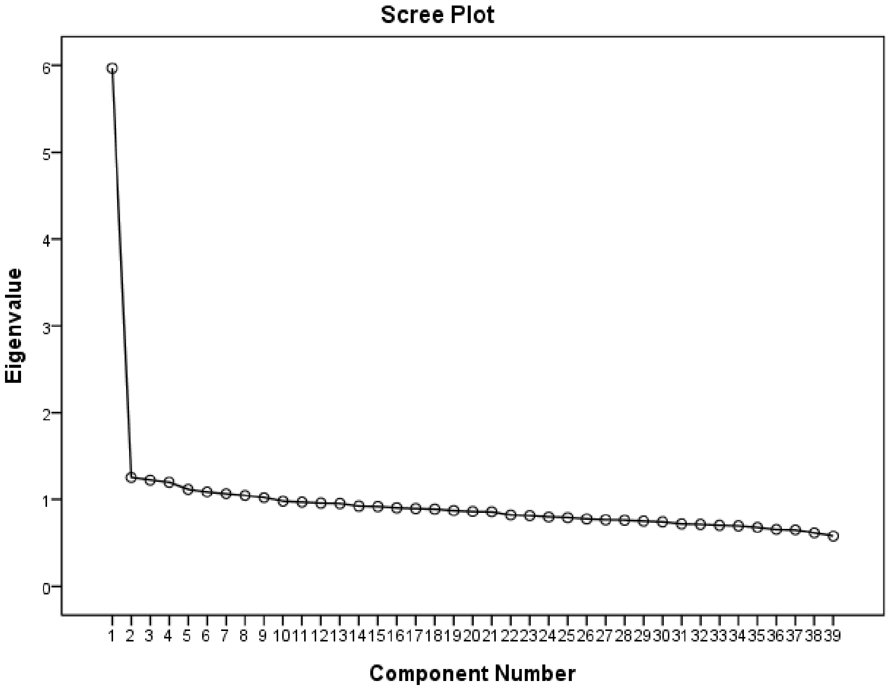


**Figure 7.1: Scree Plot for MDE Data**

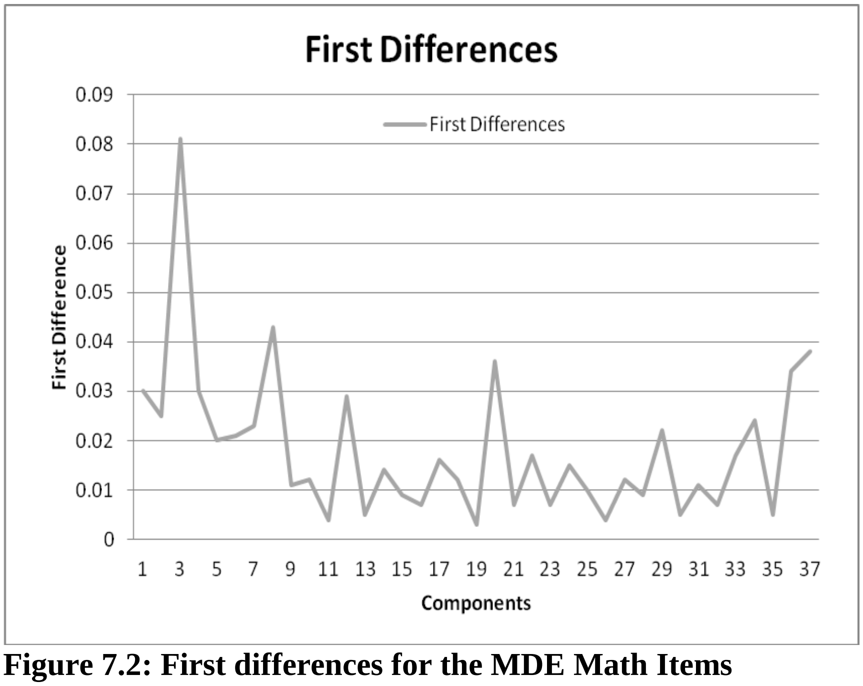


**Figure 7.2: First differences for the MDE Math Items**

Chapter 8

# 8 Conclusion

## 8.1 Discussion

The purpose of this thesis was to investigate principal components analysis as a potential exploratory technique that could be used to supplement theory in finding the Q-matrix of a cognitive diagnostic test in data that fit the DINA model. This investigation was both analytical and empirical: analytical in terms of illustrating  a components model to represent data satisfying the DINA model and how a components analysis based on the model might  be used to find the Q-matrix when there is a satisfactory item-to-skill ratio such that each skill or skill set is measured by multiple items; empirical in terms of examining components analysis with real data sets to help illuminate how the method works to uncover the Q-matrix in less than ideal situations that prevail in real testing situations.

The analytical questions investigated were used to show which skills are sufficient and which are insufficient in a designed Q-matrix. A sufficient skill is defined as a skill that is the single skill needed to solve some items whereas an insufficient skill is one that must accompany other skills in solving tasks. The first question was, when does a skill (dimension) in the Q-matrix correspond to a component in the components analysis?  and the second question was, when does a skill (dimension) in the Q-matrix fail to correspond to a component in the components analysis?  The answer to first question is that a skill in the Q-matrix will correspond to a component in the component analysis if the skill is sufficient for some items.  The answer to the second question is that a skill fails to correspond to a component in the components analysis if the skill is insufficient. From the results of the simulation study shown in Table 4.2, it can be determined, readily, which skills are sufficient and which are insufficient in a test that was designed to measure three skills.  Recall that a sufficient skill appears alone in some rows of the Q matrix  and  in combination with other skills in some rows whereas an insufficient skill never appears in isolation. As such, even without knowledge of the Q-matrix, scrutiny of the components solution in Table 4.2 and the items themselves can uncover the true Q-matrix. In short, one can identify the items loading on a component, identify the skill set common to those items, and then list the set of all skills associated with at least one component. Such an exercise should show the reader that the first two components in Table 4.2, based on the simulated data, correspond to single skills because all the items loading on those components measure a  corresponding skill (to the components) and not any other skills. Furthermore, looking at components three to five should show that the items loading on those components measure combinations of skills. The question then becomes whether skills sets three to five are comprised of insufficient skills. The answer lies in examining whether the skills that appear as skill sets also appear in isolation. In Table 4.2, we already know that skills 1 and 2 are sufficient and the only remaining skill that appears in a skill set and is insufficient is skill 3. The task analysis here is on blocks of items rather than single items as is usually the case. This exercise of examining both item content and loadings on components serves to reinforce that the components analysis method should not be used alone in isolation to determine the Q-matrix. The task of Q-matrix development is a merger of expert knowledge in item development and empirically tested statistical methods that augment theory.

The questions that were explored in the empirical phase of the study were meant to determine whether a) the proposed methodology (components analysis) yielded a plausible and useful solution for Q-matrix development and b) whether the resulting Q-matrix improved parameter estimation and examinee classification accuracy. The process of deriving a Q-matrix using real data begins with  performing a components analysis.  According to my components model, items loading on the same component should require the same skill set, and conversely, items loading on different components should require different skill sets. By examining the item loadings on each component one can identify blocks of items which require the same skill set if the components are conceptually sensible; in other words, do the items that load on a component make logical sense in terms of the skill or skill set that corresponds to that component. After a determination is made regarding the suitability of the components analysis for developing the Q-matrix, the Q-matrix is constructed containing one dimension for each of the skills identified in the components analysis. A confirmatory analysis to test the performance of the developed Q-matrix in terms of how accurately the examinees are classified into classes is then conducted.

Empirical studies use real data and in this thesis, three different types of real data were used. The first dataset was comprised of dichotomously-scored examinee responses to the fraction subtraction items in Figure 5.1. These items were written to be diagnostic.

The other two datasets were the examinee responses to the NAEP 2003 grade 8 mathematics data and the 2006 4th grade mathematics data obtained from the Minnesota Department of Education (MDE). These studies revealed some important findings worth highlighting. First, the fraction subtraction data have been widely analyzed and findings presented in research journals. As such, the original Q-matrix developed for the items is available (de la Torre & Douglas, 2004). The true nature of the Q-matrix remains a subject of debate as summarized in DeCarlo (2011). As a result, it was not surprising that the components analysis of the examinee data resulted in a Q-matrix that was somewhat different from the original Q-matrix in de la Torre and Douglas (2004).

Second, and perhaps more importantly, is that although the reconstructed Q-matrix from the components analysis (Table 5.4) had fewer skills, the results are comparable if not better than those of the original Q-matrix. Specifically, the item parameter estimates in Table 5.5 are more or less similar using both the original and the reconstructed Q-matrix. Moreover, the accuracy of estimation is also similar as indicated by the standard errors. This means that the item parameter estimation was somewhat invariant across the two Q-matrices even though they are different Q-matrices. A plausible conclusion here is that the reconstructed Q-matrix is not grossly misspecified because the resulting item parameter estimates are comparable to those from the original Q-matrix.

Third, although the fit indices presented in Table 5.7 show that the original Q-matrix was a better fit for these data, the sizes of the latent classes (Table 5.6) and the classification statistics (Tables 5.8 and 5.9) favor the reconstructed matrix.. The latent class sizes better align with the expected difficulty of the skills in the reconstructed Q-matrix than in the original Q-matrix (see Decarlo, 2011 for a discussion of the original Q-matrix). The skills that would be considered higher –level skills should have fewer students mastering them whereas the lower level skills should have more students with mastery. A determination of skill difficulty level can be determined by content specialists based on the curriculum for the grade level(s) in question in addition to item statistics such as p-values. In this case, the fraction subtraction data were administered to middle school students and with respect to those grade levels, a skill such as, borrowing can be presumed to be a higher level skill than subtracting numerators. On the same note, items requiring more than one skill to correctly execute should be more difficult than items requiring only one skill for a correct response. In that respect, the reconstructed Q-matrix resulted in latent classes that showed examinee mastery of lower level skills to be greater than mastery of higher level skills as would be expected. That was not the case with the original Q-matrix that has resulted in latent class size discrepancies leading to the uncertainty of what a reasonable Q-matrix is as discussed in DeCarlo (2011). Furthermore, the classification statistics in Tables 5.8 and 5.9 show that the reconstructed Q-matrix is superior in terms of reduction in classification errors. The issue then is whether overall model fit is preferable to classification accuracy.  Arguably, from a measurement perspective, classifying examinees accurately for cognitive diagnostic purposes may be more desirable than a small improvement in how well a Q-matrix fit the data.

The NAEP math data and the MDE math data were not analyzable using the components analysis method. First, the examinations were not designed around a diagnostic model although there are specific content strands that the items measure. The strands and the skills that the items measured seemed different (content-wise) and the strands might not necessarily correspond to the skills measured. Second, because the skills were unknown, a first analysis was used to determine the plausible number of components (skills and skill sets) that the items were measuring. Examination of the scree plots and the first differences in Figures 6.1 to 7.2 indicates that the number of components could not be determined for these data. One explanation for this observation is that the NAEP and the MDE tests are designed to measure broad content because of their summative nature. As a result, some of the skills or skill sets in the test may be measured by single items in which case the exploratory analysis would be expected to yield a large number of components each having only one item loading on that component. As such, it is not possible to derive a useful components solution. The fact that components analysis may not be feasible when tests measure broad content areas with many occurrences of some skills or skill sets measured by single items highlights a situation when the method does not work. In short, the components analysis method seems to work best when items measure narrow content and each skill set is measured by multiple items. This is indeed the case when the tests are designed to be diagnostic of skill mastery. On the contrary, achievement tests designed to measure what students have learned at the end of the school year tend to cover broad content areas and because, for various reasons, there are a limited number of items that can be administered to examinees in a testing session, some skill sets may not be measured by multiple items. The components analysis method, which is based on blocks of items measuring a skill or skill set, would not perform well if at all in such situations. Perhaps, this reiterates the fact that tests intended for diagnostic use should be designed as such and that attempts to extract such information from summative achievement tests may not be worthwhile. The reader should note this is not a discussion about retrofitting where diagnostic information is extracted from a test after the test is administered, but rather an explanation of why the broad content that is the norm in such tests hinders the extraction of the Q-matrix using a components analysis.

## 8.2 Limitations

Some limitations are evident from this thesis. First, an obvious limitation is that the components analysis method cannot be applied to tests that are designed to measure broad content domains because such tests, inadvertently, result in some skill sets being measured by single items so as to ensure adequate content coverage. It then becomes necessary to use components analysis with tests that measure a narrow content area and more importantly, each skill set should be measured by multiple items. This is because components analysis is akin to task analysis on blocks of items not single items as is usually the case. Performing task analysis on a block of items means examining the items loading on each component to determine the skill or skills measured by those items and then using the identified skills to construct the Q-matrix.

Second, while the fraction subtraction data were constructed response items, the NAEP and the MDE math data were multiple choice items. Components analysis failed in the NEAP and MDE data due to what I largely attributed to the broad content covered by the items in the test but, it could be that the item format had an effect that was not investigated in this thesis.

Third, in estimating the parameters during the confirmatory phase, the skills were assumed to be correlated and followed an underlying multivariate normal distribution. There are other correlational structures that could be investigated such as a higher-order structure where the skills are independent after controlling for a general underlying ability. The higher-order structure of the skills has been investigated in literature by J. Templin, Henson, S. Templin, and Roussos (2008), de la Torre and Douglas (2008), de la Torre and Douglas (2004) amongst others. Although both the higher-order and the multivariate normal distribution of skills have been shown to result in similar parameter estimates and classification of examinees, the higher –order specification is more parsimonious. I chose the multivariate normal as a matter of convenience and so that was the distribution that was used in the simulation study.

Finally, although the findings were not expected to differ between principal components analysis (PCA) and factor analysis, PCA was used in all of the analyses. Some might question why factor analysis was not used but a simple reason is factor analysis uses the common variance shared by variables, unlike PCA, and as a result single items that don’t share any variance with other items are lost along with the skills that are measured by those items. In components analysis, it is possible to have a single item loading on a single component that can highlight skill sets that are measured by single items. Despite the advantage of using PCA in this thesis, it would have been beneficial to conduct both factor analysis and PCA for purpose of determining how discrepant or otherwise, the solutions would be but not necessarily relevant to this thesis.

## 8.3 Further work and Conclusion

Further work that might be informative would be: 1) testing the reconstructed Q-matrix using other forms of the DINA model that specify the joint distribution of the skills differently such as, using a higher-order structure; and 2) exploring the use of components analysis with other conjunctive models such as, the reduced RUM because the RUM can also be recast into a components form. By doing so, different models could be compared for fit to the data and the best model chosen for the particular data under analysis.

To conclude, the components analysis method for Q-matrix development appears to be a viable and useful step in generating a Q-matrix when skill sets are measured by more than one item. Once items have been developed by content specialists, these items should be pilot tested and a task analysis using components analysis conducted to finalize the Q-matrix before items are used operationally. A caveat is that items must be designed to be diagnostic of specific skills covering narrow content areas with a sufficient number of items measuring each skill set. In such circumstances, the components analysis method can be a powerful tool for augmenting theory in the development of the Q-matrix.

The component representation of items fitting the DINA model and the DINA model representation itself are based on somewhat different definitions of a “dimension.” A dimension in the components representation corresponds to a set of skills that are necessary and sufficient for a high probability of passing the item. A dimension in the DINA model corresponds to a single skill that is necessary, but not necessarily sufficient by itself, for a high probability of passing the item. This mismatch between the two definitions of a dimension means that there will not be a one-to-one match between components and DINA dimensions. Nevertheless, if there are multiple items corresponding to each (or most) skill sets, a components solution, in combination with theory and content expert knowledge, can be a useful exploratory tool for deriving a Q-matrix.

In summary, the process of deriving a Q-matrix described in this thesis begins by conducting a components analysis on binary examinee responses. The items loading on each of the components are examined to identify blocks of items with the same skill set. If the components solution is sensible in terms of the cohesiveness of the skill sets measured by the items that load on each component, the Q-matrix is then constructed. The constructed Q-matrix has one dimension for each skill identified in the components solution. A confirmatory analysis is then conducted to test the performance of the constructed Q-matrix in examinee classification accuracy.

# 9 References:

Almond, R. G., DiBello, L. V., & Moulder, B, & Zapata-Rivera, J. (2007). Modeling diagnostic assessments with bayesian networks. *Journal of Educational Measurement*, *44*(4), 341.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*, 8-26.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*(4), 343.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*(4), 595.

de la Torre, J., & Karelitz, T. M. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement, 46*(4), 45.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement, 33*(8), 62.

de la Torre, J., & Young-Sun, L. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement, 47*(1), 115.

DiBello, L. V., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement, 44*(4), 285.

DiBello, L. V., Stout, W., Ruossos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In: Nichols, P. D., Chipman, S. F., Brennan, R. L. (Eds.), *Cognitively Diagnostic Assessment.* Erlbaum, Mahwah, NJ, pp.361 – 389.

DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. InC. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics, Volume 26, Psychometrics* (pp. 979–1030). Amsterdam, The Netherlands: Elsevier.

Embretson, S.E. (1985). Multicomponent latent trait models for test design. In: Embretson, S.E. (Ed.), *Test Design: Developments in Psychology and Psychometrics.* Academic Press, New York, pp. 305-321.

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of Educational Measurement, 44*(4), 325.

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

Harwell, M., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101.

Hartz, S. M. (2002). A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending  theory with practicality. Unpublished doctoral dissertation, University of Illinois, Champaign, IL.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*(4), 262.

Henson, R., Roussos, L. A., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement*, *32*(4), 275.

Henson, R., Templin, J. L., & Douglas, J. (2007). Using efficient model based sum-scores for conducting skills diagnoses. *Journal of Educational Measurement, 44*(4), 361.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191.

Henson, R. A., & Templin, J. L. (2007, April). *Large-scale language assessment using cognitive diagnosis models.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Jiang, J., Rao, J. S., Gu, Z. and Nguyen, T. (2008). Fence methods for mixed model selection. *The Annals of Statistics.* **36** 1669–1692.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258.

Lazarsfeld P.F.& Henry, N.W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin..

Leighton, J. P., & Gierl, M. J. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive Diagnostic Assessment for Education: Theory and Applications* (pp. 3–18). New York: Cambridge University Press.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on tatsuoka's rule-space approach. *Journal of Educational Measurement,* , 205.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement, 33*(8), 579.

Macready, G. B. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*(2), 99.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika* 64, 187 – 212.

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

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*(4), 342-366.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to markov chain monte carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146-178.

Reckase, M. D. & McKinley, R. L. (1991). The discriminating Power of items that measure more than one dimension. *Applied Psychological Measurement* 15, 361-373.

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement, 44*(4), 293.

Roussos, L. A., DiBello, L. V., Roussos, Stout, W. F., Hartz, S. M., Henson, R., & Templin, J. L. (2007). In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive Diagnostic Assessment for Education: Theory and Applications* (pp. 275–318). New York: Cambridge University Press.

Rupp, A. A., & Templin, J. L. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78.

SPSS Inc. (2009). *SPSS Base 8.0 for Windows User's Guide*. SPSS Inc., Chicago IL.

Stout, W. (2007). Skills diagnosis using IRT-based continuous latent trait models. *Journal of Educational Measurement, 44*(4), 313.

Sympson, J. B. (1977). A model for testing with multidimensional items. In: Weiss, D. J. (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference.* University of Minnesota, Department of Psychology, Psychometric Methods Program. Minneapolis, PP. 82 - 88.

Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345–354.

Tatsuoka, K. (1990). Toward an integration of item response theory and cognitive error diagnoses. In: Frederiksen, N., Glaser, R. L., Lesgold, A. M., Shafto, M. G. (Eds.), *Diagnostic Monitoring of Skills and Knowledge Acquisition.* Erlbaum, Hillsdale, NJ.

Templin, J. L., Henson, R. A., & Douglas, J. (2006). General theory and estimation of cognitive diagnosis models: Using Mplus to derive model estimates. Manuscript under review.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287.

Templin, J. L., Henson, R. A., Templin, S. E., & Roussos, L. A. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement, 32*(7), 559.

Vermunt, J. K., & Magidson, J. (2007, February). LG-SyntaxTM user's guide: Manual for Latent Gold 4.5 Syntax Module. Belmont, MA: Statistical Innovations.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika 45,* 479 – 494. Name is now Embretson, S.

# Appendix A

## A. Notes on the Design Features of the Simulation and the Real Data Studies

Table A.1: Design Features of Simulation Studies

| Study | Sample Size | Simulation distributions | Priors | Number of items | Number of skills | Correlational structure of the skills |
|---|---|---|---|---|---|---|
| de la Torre & Douglas (2004) | 25 replications, 1000 examinees | $\theta_i$~N(0,1)<br>$\alpha_{ik}$ ~ Ber((1+exp(-1.7$\lambda_{1k}(\theta_i-\lambda_{0k})$))}$^{-1}$<br>λ,s,g,β were fixed across all the replications | λ and $\theta_i$~N(0,1)<br>1-s, g,$\beta_0$,$\beta_{jk}$ ~ 4-Beta(varying values for the parameters) | 30 | 5 | Higher order |
| Henson & Douglas (2005), | 10,000 examinees | DINA - $S_j$, $g_j$ ~ U(.05, .40)<br>RUM – $\pi_j$ ~ U(.75, .95)<br>$r_{jk}$ ~ U(.2, .95) | | 20,20 | 4,8 | Multivariate normal |
| Henson, Templin, & Douglas(2007) | 1500 examinees<br>150 tests | RUM – $\pi_j$ ~ U(.8, .95)<br>$r_{jk}$ ~ U(.2, .95)<br>pk ~ fixed at .50<br>ρ = .30 and .50 | | 20,40 | 3,5,8 | Multivariate normal |
| Rupp & Templin (2008) | 10,000 examinees | DINA – $S_j$ ~ (0, .25)<br>$g_j$ ~ (0, .15)<br>$\theta_i$~N(0,1) | | 15 | 4 | Higher order |

| | | | | | | |
|---|---|---|---|---|---|---|
| Henson, Roussos, Douglas & He (2008) | 10,000 examinees, 1000 tests | RUM - $\pi_j \sim U(.85, .95)$<br>$r_{jk} \sim$ U( $.1+\frac{.6(i-1)}{999}, .3+\frac{.6(i-1)}{999}$ ) where *i* = 1 to the number of tests generated in this example only.<br>$\rho(.50, .75, .95)$ | | 40 | 5 | Multivariate normal |
| Templin, Henson, Templin, & Roussos (2008) | | RUM – $\pi_j \sim U(.8, 1.0)$<br>$r_{jk} \sim U(.65, .95)$<br>$p_k \sim U(.50, .60)$ | | 60 | 8 | Higher order, Multivariate normal, independent attributes |
| de la Torre & Douglas (2008). | 25 replications, 2000 examinees | $\theta_i \sim N(0,1)$<br>$\alpha_{ik} \sim Ber((1+exp(-1.7\lambda_{1k}(\theta_i-\lambda_{0k}))\}^{-1}$<br>λ,s,g,β were fixed to different values across all the replications | λ and $\theta_i \sim N(0,1)$<br>1-s, $g,\beta_0,\beta_{jk}$ ~ 4-Beta(varying values for the parameters) | 20 | 5 | Higher-order |
| de la Torre (2008). | 5000 | DINA – $s_j$ and $g_j$ parameters set to .20 for all items | | 30 | 5 | Equal probability for all attribute patterns |

| | | | | | | |
|---|---|---|---|---|---|---|
| Liu, Douglas & Henson (2009) | 1800 normal examinee, 400 aberrant ones. There were two other simulations with slight differences | DINA – $s_j$ and $g_j$ ~ U(0, .30) | | 30, 45, 60, 90 | 5 | *αis*, randomly selected from 32 possible attribute patterns. |
| de la Torre & Karelitz (2009) | 100 replications, 5000 examinees | 2-P-L – thresholds( -2, -1, 0,1,2)<br>a-parameter low condition~U(.4, .8), a-parameter high condition ~ U(1.6, 2.0),<br>$\theta$~N(0,1) | BILOG-MG<br>OX | 25 | 5 | |
| de la Torre & Young-Sun (2010) | 1000<br>100 iterations | DINA – $s_j$ and $g_j$ parameters set to .10 for all items<br>$\theta_i$~N(0,1)<br>$\lambda$ok fixed to {-1.0, -.5, 0, .5, 1.0} and $\lambda$1 fixed to 1 across all conditions. | OX | 20 | 5 | Higher order, saturated |

*Note:* the 4-beta distribution is like the usual beta distribution but the interval is not (0,1) but rather (a,b)

Table A.2: Study Design Features in Real Data Applications

| **Study** | **Sample Size** | **Estimation** | **Number of items** | **Number of skills** | **Correlational structure of the skills** |
|---|---|---|---|---|---|
| Junker and Sijtsma (2001) | 417 | MCMC in BUGS | 9 | 3 | Ordered series, a < b < c |
| de la Torre & Douglas (2004) | 2144 examinees | MCMC in OX | 20 | 8 | Higher order |
| Templin, Henson, Templin, & Roussos (2008) | 1372 examinees | | 41 | 8 | Higher order and multivariate normal. |
| de la Torre & Douglas (2008). | 2144 examinees | MCMC in OX | 15 | 7 | Higher order |
| de la Torre (2008). | 2144(subtraction data) | | 15 | 5 | |
| | 3823 (NAEP) | | 90 | 9 | |
| Liu, Douglas & Henson | 2922 | | 30 | 3 | |

| | | | | | |
|---|---|---|---|---|---|
| (2009) | | | | | |
| de la Torre & Young-Sun (2010) | 536 | MCMC in OX | 15 | 5 | Higher order, saturated |

Appendix B:

# Computer Programs

## B.1 R Simulation Code for Simulated Data

```
cycles =1   # 1 cycle for illustration purposes #
for (b in 1:cycles)  {

# generating the Q-matrix#

Q <- read.table… # reads the files #
Q <- as.matrix (Q)
colnames(Q) <- NULL
# generating item parameters #
set.seed(99)
items <- 21

s.items <- as.matrix(runif( items, .02, .05))
g.items <- as.matrix(runif( items, .05, .25))

# generating the examinees #

library(MASS)
set.seed(99)
N <- 4000
mu <-rep (0, 3)
rho <- .50
sigma <- matrix(c(1, rho, rho,rho, 1, rho,  rho, rho, 1), 3,3)
alpha.c <- matrix(0, nrow= N, ncol= 3)
  for(i in 1:nrow(alpha.c)){
      for(j in 1:ncol(alpha.c)){
             alpha.c <- as.matrix(mvrnorm(N, mu, sigma))
             }}

## some data checks, means should be zero, sd=1, and corr=.5 ###
meanss <- c(mean(alpha.c[,1]), mean(alpha.c[,2]), mean(alpha.c[,3]))
meanss
stdev <-  c(sd(alpha.c[,1]), sd(alpha.c[,2]), sd(alpha.c[,3]))
stdev
```

```
cor(alpha.c)


## generating the Pk parameters (proportion of examinees mastering each attribute, I
decided to fix it at .5 #
p.k <- matrix(ncol=3, nrow= N)
  for (k in 1: nrow(p.k)){
    for(l in 1: ncol(p.k)){
           p.k[k,l] <- .5
                          }}

z.k <- matrix(1, nrow= N, ncol=3)
  for(m in 1: nrow(z.k)){
           for(n in 1:ncol(z.k)){
z.k <- qnorm(p.k, mean = 0, sd = 1, lower.tail = FALSE)
                                              }}

# scoring the examinees as either masters or non-masters  #


# converting the continuous alpha.c into binary#
alpha.bin <- matrix(0, nrow= 4000,  ncol= 3)
for( r in 1:nrow(alpha.bin)){
    for ( s in 1:ncol(alpha.bin)) {
       if (alpha.c[r,s] >= z.k[r,s])   alpha.bin[r,s] <-  1
       else alpha.bin[r,s] <- 0
       }}

## Generating the item response with the DINA using all the parameters simulated above
##
xsi <- matrix(1,  nrow= 4000, ncol= 21)
resp.prob <- matrix(1, nrow= 4000, ncol= 21)
for (g in 1: nrow(alpha.bin)) {     # persons binary abilities
  for (h in 1: nrow(Q)) {            # the Q-matrix
     for ( a in 1: ncol(Q)) {
     xsi[g,h]= xsi[g,h] * (alpha.bin[g,a] ^ Q[h,a])
     }

     resp.prob[g,h] <- (((1-s.items[h])^xsi[g,h]) *((g.items[h])^(1-xsi[g,h])))
      }
       }

#generating a random u matrix of random numbers between 0 and 1
```

```
u <- matrix(0,nrow=N, ncol=items)
    for(x in 1:nrow(u)) {
      for(y in 1:ncol(u)){
           u <- as.matrix(runif(u,0,1))
                  }}
u <- matrix((u), 4000,21)
print(head(u))
print(tail(u))

# Dichotomizing the responses #

resp.bin <- matrix(0, nrow= 4000,  ncol= 21)
  for( e in 1:nrow(resp.bin)){
    for ( f in 1:ncol(resp.bin)) {
      if (u[e,f] <= resp.prob[e,f])   resp.bin[e,f] <-  1
        else resp.bin[e,f] <- 0
                              }}
```

## B.2 Latent Gold Syntax for Tatsuoka's Data (DeCarlo, 2011)

```
model
options
   algorithm
      tolerance=1e-008 emtolerance=.01 emiterations=350 nriterations=50;
   startvalues
      seed=0 sets=10 tolerance=1e-005 iterations=50;
   bayes
      categorical=1 variances=0 latent=1 poisson=0;
   montecarlo
      seed=0 replicates=500 tolerance=1e-008;
   quadrature nodes=10;
   missing excludeall;
   output
      parameters=first standarderrors probmeans=posterior profile bvr
      identification classification;
   outfile 'Frac_dina_out2.sav' classification;
variables
   dependent i1 cumlogit, i2 cumlogit, i3 cumlogit, i4 cumlogit, i5 cumlogit,
   i6 cumlogit, i7 cumlogit, i8 cumlogit, i9 cumlogit, i10 cumlogit, i11 cumlogit,
   i12 cumlogit, i13 cumlogit, i14 cumlogit, i15 cumlogit, i16 cumlogit,
   i17 cumlogit, i18 cumlogit, i19 cumlogit, i20 cumlogit;
   latent
      a1 ordinal 2 score=(0 1), a2 ordinal 2 score=(0 1),
      a3 ordinal 2 score=(0 1), a4 ordinal 2 score=(0 1),
      a5 ordinal 2 score=(0 1), a6 ordinal 2 score=(0 1),
      a7 ordinal 2 score=(0 1);
equations
   a1-a7 <- 1;
   i1 <- 1 + a2 a4;
   i2 <- 1 + a2 a4;
   i3 <- 1 + a2 a4;
   i4 <- 1 + a1 a5;
   i5 <- 1 + a2 a3 a4 a7;
   i6 <- 1 + a2;
   i7 <- 1 + a1 a2;
   i8 <- 1 + a5;
   i9 <- 1 + a2 a3;
   i10 <- 1 + a1 a2 a3;
   i11 <- 1 + a1 a2 a3;
   i12 <- 1 + a2 a6 a7;
   i13 <- 1 + a1 a2 a4 a5;
   i14 <- 1 + a2 a3 a5;
   i15 <- 1 + a1 a2;
   i16 <- 1 + a2 a3;
   i17 <- 1 + a1 a2 a3;
   i18 <- 1 + a1 a2 a3;
   i19 <- 1 + a1 a2;
   i20 <- 1 + a1 a2 a3;
end model
```

# Appendix C:

## C. Transforming the DINA Parameters and Computing the Standard Errors of the Transformed Variables.

In brief, to estimate the DINA model parameters in Latent Gold, the DINA model is reparameterized in a manner such that Latent Gold runs a series of regression equations. As an example, item 1 in the reconstructed Q-matrix measures skills 2 and 4. Latent Gold estimates the parameters using the equation, item1 <- 1 + skill2 skill4; and outputs parameter estimates as shown in the table below.

**An example of Parameter Estimates for Item 1 in the Reconstructed Q-matrix**

| term | coef | s.e. | z-value | p-value | Wald(0) | df | p-value |
|---|---|---|---|---|---|---|---|
| item1(1) <- 1 | -3.4113 | .5287 | -6.4524 | 1.10E-10 | 41.6332 | 1 | 1.10E-10 |
| item1 <- skill2 skill4 | 5.3082 | .5629 | 9.4295 | 4.10E-21 | 88.915 | 1 | 4.10E-21 |

Note that the skills are denoted by skill2 and skill4. The first coefficient (parameter estimate) = -3.4113 is transformed using the function $\frac{e^{-3.4113}}{1+e^{-3.4113}} = 0.031944172$ to obtain the probability of guessing denoted by the parameter, $g_j$. To obtain the estimated probability minus the slip parameter, 1-$s_j$, add the two co-efficients (estimates), that is, -3.4113+5.3082 = 1.8969 and transform using the function $\frac{e^{1.8969}}{1+e^{1.8969}} = 0.869540264$.

Because the parameters are transformed from the logit scale to the probability scale, the reported standard errors must also be transformed. For the guessing parameter, $\widehat{g_j} = \frac{e^{\hat{\beta}}}{1+e^{\hat{\beta}}}$, the variance of the transformed guessing parameter is then

$\widehat{var(g_j)} \approx \left(\frac{e^{\hat{\beta}}}{(1+e^{\hat{\beta}})^2}\right)^2 * \widehat{var(\hat{\beta})}$. The first term in this product is the square of the derivative of g$_j$ with respect to $\beta$. $\beta$ is estimated by the corresponding coefficient in the table above. As such, the variance of g$_j$ is $\left(\frac{e^{-3.4113}}{(1+e^{-3.4113})^2}\right)^2 * 0.5287^2 = 0.000267302$. and the standard error (S.E) is the square root of the variance, which is, S.E = .016349384. To compute the variance of the estimated probability of 1-s$_j$, $\widehat{1-s_j} = \frac{e^{\hat{\beta}_1+\hat{\beta}_2}}{1+e^{\hat{\beta}_1+\hat{\beta}_2}}$, it can be shown that $var\widehat{(1-sj)} \approx \mathbf{B\Sigma B}^T$where **B** is the partial derivative of 1-s$_j$ with respect to the parameters $\beta_1$ and $\beta_2$ and $\mathbf{\Sigma}$ is variance-covariance matrix of the two parameters. The variance-covariance matrix can be obtained by requesting it in the output section of the Latent Gold syntax. The variance formula is further expanded to show that

$$B = \begin{bmatrix} \frac{e^{\beta_1+\beta_2}}{(1+e^{\beta_1+\beta_2})^2} & \frac{e^{\beta_1+\beta_2}}{(1+e^{\beta_1+\beta_2})^2} \end{bmatrix}$$

Since the partial derivative is the same for both of the estimates that are used to compute 1-s$_j$, $\left(\frac{e^{-3.4113+5.3082}}{(1+e^{-3.4113+5.3082})^2}\right) = \left(\frac{e^{1.8969}}{(1+e^{1.8969})^2}\right) = 0.113439993$.

$$var(1-sj) = \begin{bmatrix} 0.113439993 & 0.113439993 \end{bmatrix} * \begin{bmatrix} 0.2795 & -0.2786 \\ -0.2786 & 0.3169 \end{bmatrix} * \begin{bmatrix} 0.113439993 \\ 0.113439993 \end{bmatrix}$$

$= 0.0005045$

The S.E is the square root of the variance, sqrt(.0005045) = .02246.

## Appendix D:

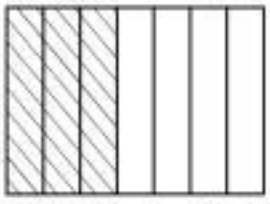
## D. NAEP Items

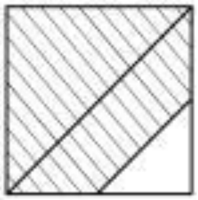
1. Add:

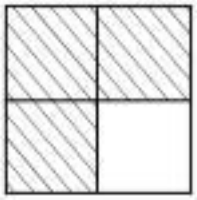
```
   238
+  462
```

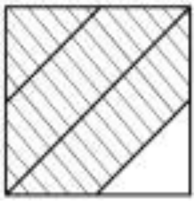
A. 600<br>
B. 690<br>
C. 700<br>
D. 790

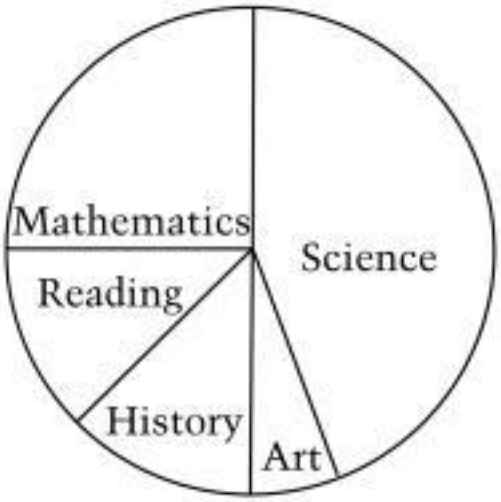
2. Which shows 3/4 of the picture shaded?

A. 

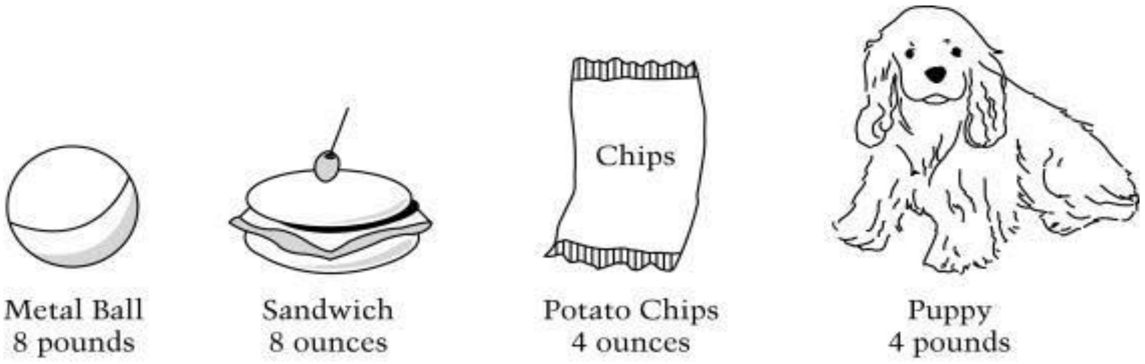
B. 

C. 

D. 

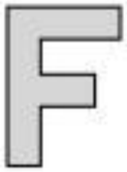


3. The pie chart above shows the portion of time Pat spent on homework in each subject last week. If Pat spent 2 hours on mathematics, about how many hours did Pat spend on homework altogether?
   A. 4
   B. 8
   C. 12
   D. 16

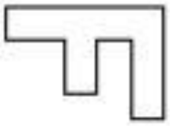


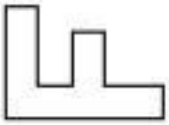
4. In the figure above, which object is heaviest?
   A. The metal ball

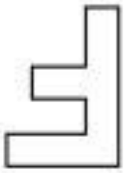
B. The sandwich
C. The bag of potato chips
D. The puppy

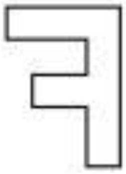


5. The figure above is shaded on the top side and white on the underside. If the figure were flipped over, its white side could look like which of the following figures?

A. 

B. 

C. 

D. 

6. How much change will John get back from $5.00 if he buys 2 notebooks that cost $1.80 each?
A. $1.40
B. $2.40
C. $3.20
D. $3.60

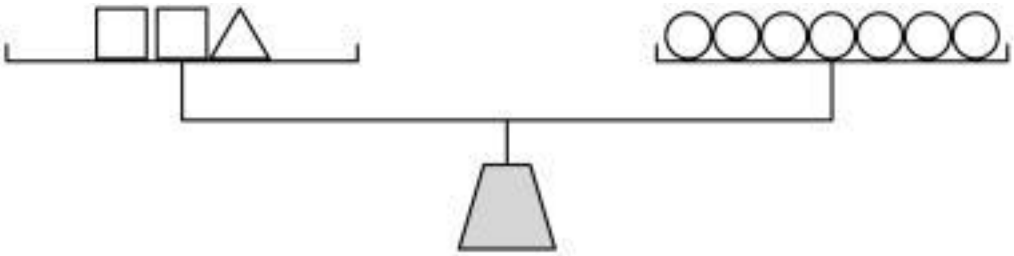
7. Carla has 12 boxes that each weigh the same amount. What would be a quick way for her to find the total weight of the 12 boxes?

A. Add 12 to the weight of one of the boxes.
B. Subtract 12 from the weight of one of the boxes.
C. Divide the weight of one of the boxes by 12.
D. Multiply the weight of one of the boxes by 12.

8. The perimeter of a square is 36 inches. What is the length of one side of the square?
A. 4 inches
B. 6 inches
C. 9 inches
D. 18 inches

9. Six students bought exactly enough pens to share equally among themselves. Which of the following could be the number of pens they bought?
A. 46
B. 48
C. 50
D. 52

10. Carl has 3 empty egg cartons and 34 eggs. If each carton holds 12 eggs, how many more eggs are needed to fill all 3 cartons?
A. 2
B. 3
C. 4
D. 6



11. The objects on the scale above make it balance exactly. According to this scale, if △ balances ○○○, then □ balances which of the following?

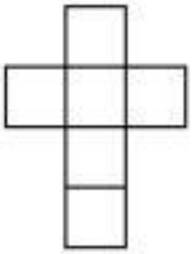
A. ○

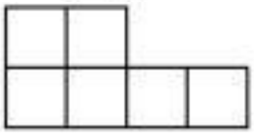
B. ○○

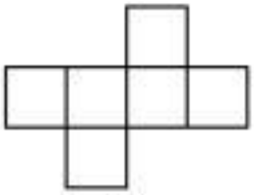
C. ○○○

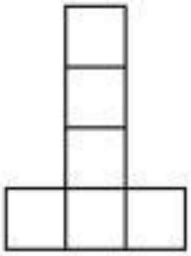
D. ○○○○

12. Which of the following could NOT be folded into a cube?

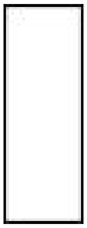
A. 

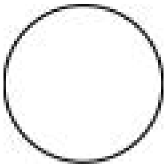
B. 

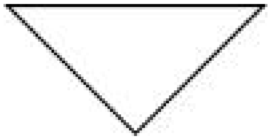
C. 

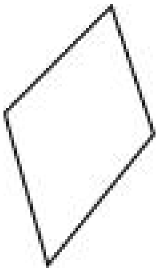
D. 

13. Alan says that if a figure has four sides, it must be a rectangle. Gina does not agree. Which of the following figures shows that Gina is correct?

A. 

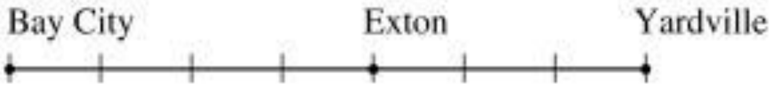
B. 

C. 

D. 

14. Jim has 3/4 of a yard of string which he wishes to divide into pieces, each 1/8 of a yard long. How many pieces will he have?

A. 3
B. 4
C. 6
D. 8

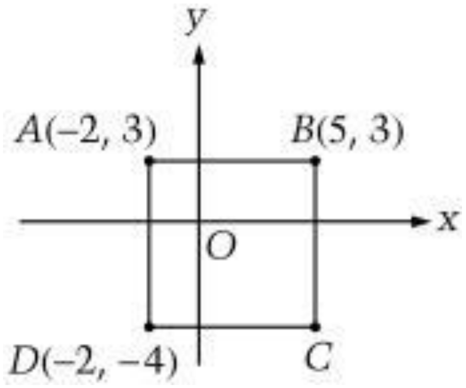


15. On the road shown above, the distance from Bay City to Exton is 60 miles. What is the distance from Bay City to Yardville?
A. 45 miles
B. 75 miles
C. 90 miles
D. 105 miles

16. What are all the whole numbers that make 8 - □ > 3 true?
A. 0, 1, 2, 3, 4, 5
B. 0, 1, 2, 3, 4
C. 0, 1, 2
D. 5

17. Length can be measured to within .05 centimeter accuracy by using a certain type of measuring instrument. A reading of 3.7 centimeters on this instrument means that the actual length is at least
A. 3.20 centimeters
B. 3.65 centimeters
C. 3.69 centimeters
D. 3.70 centimeters
E. 3.75 centimeters

18. A triangle that has sides with lengths 6, 6, and 10 is called
A. acute
B. right
C. scalene
D. isosceles
E. equilateral



19. In the figure above, if *ABCD* is a square, then the coordinates of vertex *C* are

A. (4,5)
B. (3,-4)
C. (3,-2)
D. (5,-4)
E. (5,-2)

20.

| *x* | *y* |
|---|---|
| 0 | -3 |
| 1 | -1 |
| 2 | 1 |

Which of the following equations is true for the three pairs of *x* and *y* values in the table above?

A. $3x + 2 = y$
B. $3x - 2 = y$
C. $2x + 3 = y$
D. $2x - 3 = y$
E. $x - 3 = y$

21. 4, 8, 3, 2, 5, 8, 12

What is the median of the numbers above?

A. 4  
B. 5  
C. 6  
D. 7  
E. 8

1 mile = 5,280 feet

22. How many feet are in 15 miles?

A. 352  
B. 35,200  
C. 79,200  
D. 84,480  
E. 89,760

Did you use the calculator on this question?

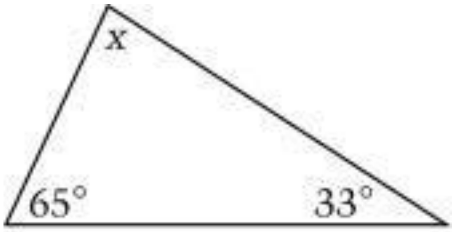
○ Yes ○ No

23. The Breakfast Barn bought 135 dozen eggs at $.89 per dozen. What was the total cost of the eggs?

A. $116.75  
B. $12.15  
C. $135.89  
D. $151.69

Did you use the calculator on this question?

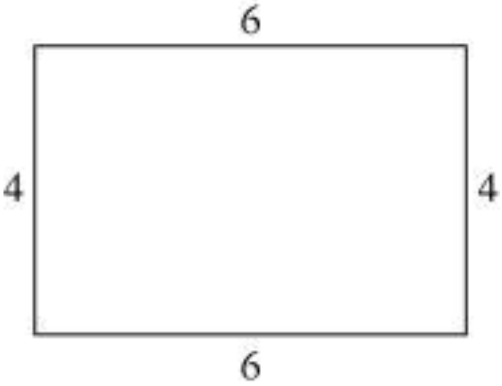
○ Yes ○ No



24. What is the value of x in the triangle above?
    A. 65°
    B. 82°
    C. 90°
    D. 92°
    E. 98°

    Did you use the calculator on this question?

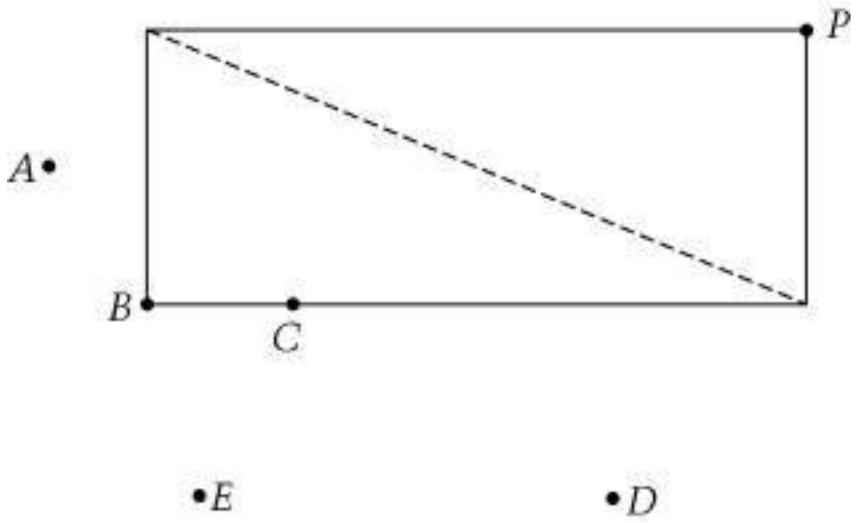
    ○ Yes ○ No



25. Which of the following numerical expressions gives the area of the rectangle above?
    A. 4 × 6
    B. 4 + 6
    C. 2(4 × 6)
    D. 2(4 + 6)
    E. 4 + 6 + 4 + 6

    Did you use the calculator on this question?

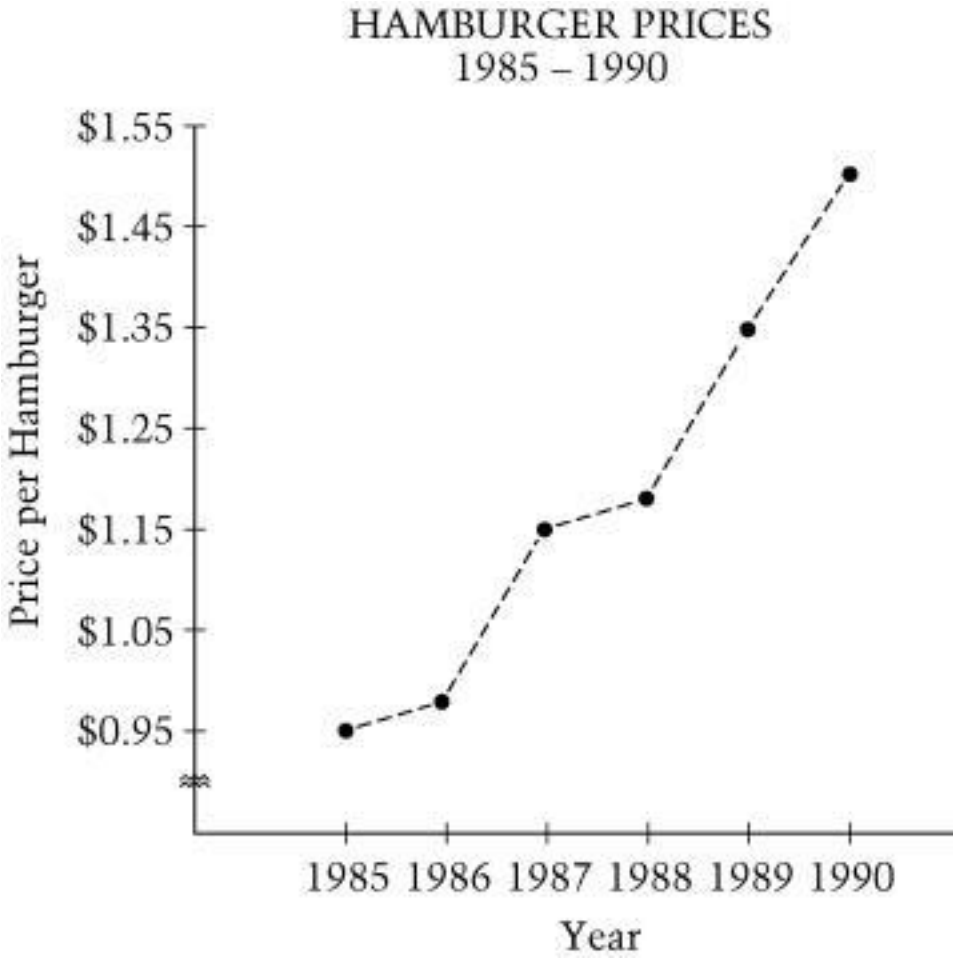
○ Yes ○ No



26. When the rectangle above is folded along the dotted line, point *P* will touch which of the lettered points?
    A. A
    B. B
    C. C
    D. D
    E. E

    Did you use the calculator on this question?

    ○ Yes ○ No



27. According to the graph above, how many times did the yearly increase of the price of a hamburger exceed 10 cents?
   A. None
   B. One
   C. Two
   D. Three
   E. Four

   Did you use the calculator on this question?

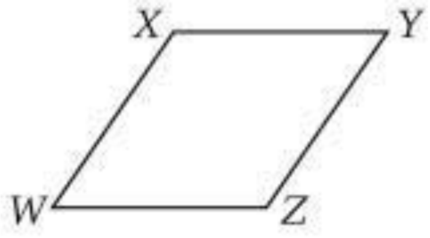
   ○ Yes ○ No

28. Fifteen boxes each containing 8 radios can be repacked in 10 larger boxes each containing how many radios?

A. 8
B. 10
C. 12
D. 80
E. 120

Did you use the calculator on this question?

○ Yes ○ No



29. In the figure above, *WXYZ* is a parallelogram. Which of the following is NOT necessarily true?
A. Side *WX* is parallel to side *ZY*.
B. Side *XY* is parallel to side *WZ*.
C. The measures of angles *W* and *Y* are equal.
D. The lengths of sides *WX* and *ZY* are equal.
E. The lengths of sides *WX* and *XY* are equal.

Did you use the calculator on this question?

○ Yes ○ No

30. Consider the statement "If *n* is an even number, then *n* is two times an odd number." For which of the following values of *n* is the statement FALSE?

A. 2
B. 6
C. 8
D. 10
E. 14

Did you use the calculator on this question?

○ Yes ○ No

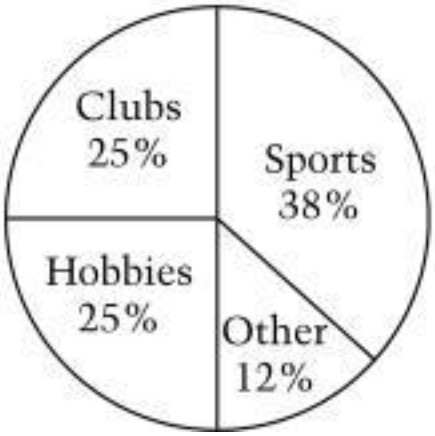
31. If the list of fractions above continues in the same pattern, which term will be equal to .95?

A. The 100th
B. The 95th
C. The 20th
D. The 19th
E. The 15th

Did you use the calculator on this question?

○ Yes ○ No

STUDENT PARTICIPATION IN
ACTIVITIES AT ADAMS MIDDLE SCHOOL



32. There are 1,200 students enrolled in Adams Middle School. According to the graph above, how many of these students participate in sports?
A. 380
B. 456
C. 760
D. 820
E. 1,162

Did you use the calculator on this question?

○ Yes ○ No

33. The diameter of a red blood cell, in inches, is 3 × 10 . This expression is the same as which of the following numbers?
A. .00003
B. .0003
C. .003
D. 3,000
E. 30,000

Did you use the calculator on this question?

○ Yes ○ No

34. Tetsu rides his bicycle $x$ miles the first day, $y$ miles the second day, and $z$ miles the third day. Which of the following expressions represents the average number of miles per day that Tetsu travels?

A. $x + y + z$
B. xyz
C. $3(x + y + z)$
D. $3(xyz)$
E. $(x + y + z)/3$

Did you use the calculator on this question?

○ Yes ○ No

35. If $3 + w = b$, then $w =$

A. $b/3$
B. $b \times 3$
C. $b + 3$
D. $3 - b$
E. $b - 3$

Did you use the calculator on this question?

○ Yes ○ No

36. If the value of the expression $x + 2$ is less than 12, which of the following could be a value of $x$?

A. 16
B. 14
C. 12
D. 10
E. 8

POPULATION

| | |
|---|---|
| Clear Lake | 8,000 |
| Rancho Santa Fe | 4,000 |
| Bull Shoals | 1,500 |
| Beaver City | 750 |
| Jeffersonville | 500 |

37. A pictogram of the data above is to be drawn using 🯅 as the symbol that represents 500 people. How many 🯅 would it take to represent the population of Rancho Santa Fe?

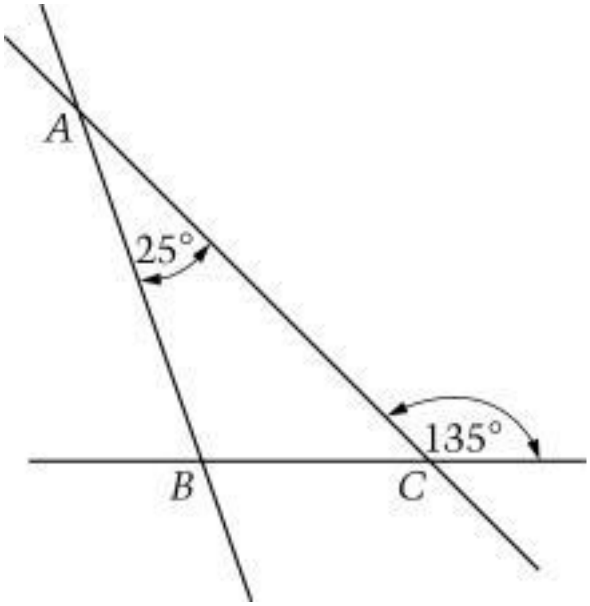
A. 1
B. 4
C. 8
D. 80
E. 4,000

38. Gloria's diving scores from a recent competition are represented in the stem-and-leaf plot shown below. In this plot, 3   4 would be read as 3.4.

```
52  5
61
77
80  2
```

What was her lowest score for this competition?

A. .02
B. 1.0
C. 2.5
D. 5.2
E. 8.0



39. In the triangle, what is the degree measure of $\angle ABC$?
    A. 45
    B. 100
    C. 110
    D. 135
    E. 160

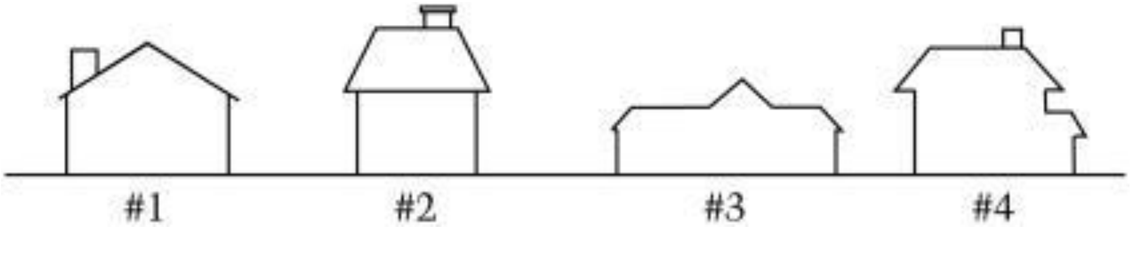
40. Which of the following ratios is equivalent to the ratio of 6 to 4?
    A. 12 to 18
    B. 12 to 8
    C. 8 to 6
    D. 4 to 6
    E. 2 to 3

41. In a coordinate plane, the points (2,4) and (3,-1) are on a line. Which of the following must be true?
    A. The line crosses the *x*-axis.
    B. The line passes through (0,0).
    C. The line stays above the *x*-axis at all times.
    D. The line rises from the lower left to the upper right.
    E. The line is parallel to the *y*-axis.

42. 3 + 15 ÷ 3 - 4 × 2 =
   A. -9
   B. -2
   C. 0
   D. 4
   E. 5

43. The cruise ship Titanic was 882 feet long. Which of the following is closest to that length?
   A. Two moving-van lengths
   B. Fifty car lengths
   C. One hundred skateboard lengths
   D. Five hundred school-bus lengths
   E. One thousand bicycle lengths

44. If *n* represents an even number greater than 2, what is the next larger even number?
   A. $n + 1$
   B. $2n + 1$
   C. $2n$
   D. $n + 2$
   E. $n$



45. Allen, Bridgitte, Chaz, and Diann each live in a different house on the same side of a street. The houses and their numbers are shown above.
   • Only one of the other three people lives next to Bridgitte.

• Chaz lives next to Bridgitte and next to Diann.

Which person could live in house number 2?

A. Allen only
B. Chaz only
C. Diann only
D. Chaz or Diann
E. Any of these four people could live in house number 2.