NEW APPROACHES TO Q-MATRIX VALIDATION AND ESTIMATION FOR COGNITIVE DIAGNOSIS MODELS

By

OLASUMBO O. OLUWALANA

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey,

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Education

Written under the direction of

Chia-Yi Chiu

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

MAY 2019

ABSTRACT OF THE DISSERTATION

New Approaches to Q-Matrix Validation and Estimation

for Cognitive Diagnosis Models

by **OLASUMBO O. OLUWALANA**

Dissertation Director:
Chia-Yi Chiu

A primary purpose of cognitive diagnosis models (CDMs) is to classify examinees based on their attribute patterns. The Q-matrix (Tatsuoka, 1985), a common component of all CDMs, specifies the relationship between the set of required dichotomous attributes and the test items. Since a Q-matrix is often developed by content-knowledge experts and can be influenced by their judgment (de la Torre & Chiu, 2016), this can lead to misspecifications in the Q-matrix that can have unintended consequences on examinees' classifications. Incorrect classification of examinees can have tremendous impact since some assessments are high-stake and are used to make important decisions about students, such as selection and placement. Previous research based on the Trends in International Math and Science Study (TIMSS) has predominantly focused on comparing the performances of participating countries using their average scores.

This study focused on fitting data from the TIMSS with a CDM to obtain estimated attribute profiles that will provide information about skill proficiency of students in the participating countries. However, since the test is not specifically designed for use with a CDM, a provisional Q-matrix was developed with input from content experts. As a

preliminary analysis, the TIMSS data was first fitted with the generalized deterministic inputs, noisy, "and" gate (G-DINA) model to obtain examinees' estimated attribute profiles. An evaluation of the estimated attribute profiles however indicated that there are inconsistencies in classification, which may be due to misspecification in the provisional Q-matrix. To ensure that the provisional Q-matrix is appropriately developed, this dissertation proposes one Q-matrix validation method that can be used to correct possible misspecifications in a Q-matrix, and one Q-matrix estimation method for estimating a Q-matrix from scratch.

The proposed methods both integrate the Q-matrix validation procedure (Chiu, 2013) that is based on a nonparametric classification method. The first method, the integrated Q-matrix validation (IQV) technique, uses a joint maximum likelihood estimation (JMLE) procedure for diagnostic classification models (Chiu, Köhn, Zheng, and Henson, 2016) to determine examinees' attribute profiles that are then integrated into the algorithm of Chiu's Q-matrix validation method to validate the Q-matrix. In the second method, the two-step Q-matrix estimation (TSQE) method, factor analysis is first applied to the correlation matrix to obtain a provisional Q-matrix. The provisional Q-matrix is then incorporated into the algorithm of Chiu's Q-matrix validation method, to obtain the true Q-matrix.

The viability of both methods was investigated using simulation studies with various conditions. The TIMSS data was re-analyzed with the G-DINA model using modified Q-matrices obtained from analysis with the proposed methods. An evaluation of the updated estimated attribute profiles indicated that some of the inconsistencies in classification that were previously identified have been resolved.

ACKNOWLEDGMENT

I am extremely grateful to everyone who has been part of this seven-year journey with me.

To my advisor, Dr. Chia-Yi Chiu; for her patience, support, and most especially helping me to develop quality work.

To members of my dissertation committee—Dr. Jiawen Zhou, Dr. Greg Camilli, and Dr. Dake Zhang; for their insight and professional guidance

To my friends and family who supported, encouraged, and prayed for me as I made progress toward this goal

To my greatest cheer leaders Babafemi, Oluwatobi, and Mofeyi; I could not have accomplished this without your love and support – love you loads

And above all, to my Great God, who made ways where there seemed to be no way and with whom NOTHING SHALL BE IMPOSSIBLE

DEDICATION

This dissertation is dedicated to the memory of my mum, an outstanding and selfless educator.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Educational assessment based on traditional psychometric models such as classical test theory (CTT) and item response theory (IRT) use statistical frameworks that identify an examinee's position along a latent ability continuum. The information obtained from these assessments is typically used for ranking or comparing examinees, and for determining how well examinees have performed based on specific standards. In contrast, cognitive diagnosis models (CDMs), sometimes also referred to as diagnostic classification models (DCMs; Rupp, Templin, & Henson, 2010), give information about examinees' mastery or nonmastery of a set of fine-grained attributes required to respond correctly to test items (de la Torre, 2008). Thus, the primary purpose of CDMs is to estimate examinees' attribute profiles based on the attributes required to respond correctly to the items on a test. Although CDMs are predominantly developed and used to analyze educational assessments, they have also been applied to other fields. For example, they have been employed in clinical psychology to identify psychological disorders by using a multidimensional classification of examinees based on their behavioral dispositions (Templin & Henson, 2006).

Many CDMs have been proposed in literature with the main difference between models being the assumptions about the relationship between items and attributes in determining the probability of a correct response. Some frequently used models are introduced in detail in the next chapter. Despite the differences between models, all analyses with CDMs require a Q-matrix (Tatsuoka, 1985), a collection of q-vectors that are individually matched to each item on a test and define the attributes required to answer each item correctly. Given a CDM, information from a Q-matrix and examinees' responses

are used to estimate examinees' attribute profiles. These profiles indicate the attributes each examinee has mastered and the ones they are yet to master. Assessments based on CDMs can thus be used to evaluate examinees' learning and progress, to improve instruction, and to identify appropriate intervention. Therefore, the Q-matrix is an important component of cognitive diagnostic analysis and the quality of the Q-matrix determines the validity of the assessment and inferences made based on the test results (Rupp & Templin, 2008; DeCarlo, 2011).

However, a Q-matrix is often constructed based on the opinions of content experts and this subjective process may lead to misspecifications in the Q-matrix. In addition, many existing assessments are not designed for use with CDMs, and through a process of retrofitting an adhoc Q-matrix is often created based on information given in a test blueprint. However, this process can cause misspecifications in the Q-matrix. These misspecifications and inaccuracies in the Q-matrix often negatively impact model parameter estimations, which may result in the erroneous classification of students, and an inaccuracy of the inferences made based on test results. Since the development of a Q-matrix is one of the most important aspects of using CDMs, Q-matrix validation and estimation methods have been developed to ensure that a Q-matrix is accurately specified. Q-matrix validation methods are used for identifying and correcting the misspecification and inaccuracies that may occur in a provisional Q-matrix due to the subjective process by which it is often developed. These methods are essential because a Q-matrix is often assumed to be correctly constructed and model fit analyses of CDMs often assume that a Q-matrix is correct without substantial evidence of its appropriateness (de la Torre, 2008). On the other hand, Q-matrix estimation methods are used for developing a Q-matrix from

scratch by relying only on information from examinees' responses and the given CDM. Q-matrix estimation methods are particularly useful in situations where content experts are not readily available to guide the process of developing a provisional Q-matrix such as during the process of retrofitting non-cognitive-diagnostic assessment for cognitive diagnostic purposes.

Although several Q-matrix validation and estimation methods currently exist, some limitations have been identified with these methods. For example, some existing Q-matrix estimation methods (Barnes 2010; Liu, Xu, and Ying, 2012; Culpepper, Chen, and Douglas, 2018) have only been used with specific CDMs such as the Deterministic Input, Noisy "And" Gate (DINA; Haertel, 1989; Junker & Sijstma, 2001), Deterministic Input, Noisy "Or" Gate (DINO; Templin & Henson, 2006), and a restricted version of the reparameterized unified model (RRUM: Hartz, 2002; Hartz, Roussos, & Stout, 2002). Furthermore, the robustness and generalizability of the model-based approach to the Q-matrix validation (DeCarlo, 2012; Templin & Henson, 2006a) requires additional exploration especially for situations in which all the misspecified q-entries have not been detected. Results from the simulation study and real data analysis based on the Q-matrix validation method proposed by de la Torre (2008) shows that that the procedure has great potential, however it is unconfirmed if the procedure can be used with models other than the DINA model, and in particular, generalized CDMs. Although the general method of empirical Q-matrix validation (de la Torre & Chiu, 2016) overcomes the shortcoming of validation methods that cannot be used with more general CDMs, the use of the method has limited generalization because of the requirement for an arbitrary cutoff to stop the algorithm from over-correcting. It is therefore vital to develop additional methods of Q-

matrix validation and estimation that are simple and easy to use under varying conditions and with a variety of CDMs.

This study proposes one Q-matrix validation method and one Q-matrix estimation method that are ideal for use under a variety of conditions and with a variety of CDMs. The first method, the integrated Q-matrix validation (IQV) method, uses the Joint Maximum Likelihood Estimation (JMLE) for cognitive diagnostic models (Chiu, Köhn, Zheng, & Henson, 2016) to classify examinees and then validates the q-vectors by applying the Q-matrix validation method (Chiu, 2013). By using the JMLE algorithm to estimate examinees' attribute profiles, the IQV method can be used with more general CDMs beyond the DINA and DINO models, unlike the Q-matrix validation method (Chiu, 2013) which is limited in use with only the DINA model. In the second method, the two-step Q-matrix estimation (TSQE) method, a provisional Q-matrix is estimated by first applying the nonlinear factor analysis to examinees' response data. This provisional Q-matrix is then used as the input for the Q-matrix validation method (Chiu, 2013) to obtain the modified Q-matrix. Apart from its simplicity and minimal computation time, the TSQE method can be used with a variety of CDMs and sample sizes. The performance of both proposed methods is evaluated using two simulation studies each. In addition to the simulation studies, the proposed methods are applied to a subset of the fourth-grade mathematics data from the 2011 Trends in International Mathematics and Science Study (TIMSS). First, the data are fitted with the G-DINA model to estimate examinees' attribute profiles which are then evaluated to determine if there are discrepancies in the classification. The presence of discrepancies is evidence of a likely misspecification in the provisional Q-matrix developed in collaboration with content experts. The IQV method is then applied to

improve the issue of discrepancies in student classification and to correct the misspecifications in the provisional Q-matrix. Similarly, the TSQE method is applied to examinees' response data to obtain the modified Q-matrix. Preliminary results show that both the IQV and TSQE methods have potential in correcting misspecifications in a provisional Q-matrix and in identifying the true Q-matrix respectively.

Chapter 2. Literature Review

2.1 Overview of Cognitive Diagnosis Models

Cognitive diagnosis models (CDMs) are latent class models in which classes are defined by examinees' mastery or nonmastery of a set of skills. CDMs provide fine-grained diagnosis of examinees' strengths and weaknesses in the form of an attribute profile, that can be used to direct instructional improvement plans, provide targeted intervention to meet students' needs, and improve educational outcomes. CDMs can be classified based on some defining characteristics (Rupp & Templin, 2008): (1) the type of measurement scale of the observed response variables (dichotomous or polytomous), (2) the type of measurement scale used by the attributes or skills being measured (dichotomous or polytomous), (3) the way in which the attributes or skills are combined within each item (compensatory or noncompensatory), and how the probability of an examinee correctly responding to an item is determined (Conjunctive or disjunctive). For example, the DINA model, the DINO model and the Noisy Input, Deterministic "And" Gate (NIDA; Junker & Sijstma, 2001) models are based solely on dichotomous response data. Some others CDMs like the General diagnostic model (GDM; von Davier, 2005; Xu & von Davier, 2006) are designed to use both dichotomous and polytomous response data. In compensatory CDMs (e.g., DINO), the absence of an attribute can be compensated for by the presence of other attribute(s), while noncompensatory models (e.g., DINA, NIDA) require mastery of all attributes for an examinee to answer an item correctly. Conjunctive models (DINA, NIDA; Hartz, 2002; Roussos, DiBello, Stout, Hartz, Henson, & Templin, 2007) require mastery of all necessary attributes for an examinee to have a high probability of answering an item

correctly, while in disjunctive models (e.g., DINO) only one attribute needs to be mastered for an examinee to have a high probability of answering an item correctly.

Despite differences in their defining characteristics, all CDMs use a Q-matrix (Tatsuoka, 1983) with dimension $J \times K$. The Q-matrix specifies the attributes required to answer each item on a test correctly. Items are denoted by $j = 1, \ldots, J$, attributes or skills are denoted by $k = 1, \ldots, K$. Each item has a corresponding q-vector, q, that has a length of K. Since attributes are characterized as being discrete and dichotomous, they are either required or not required to answer an item correctly. If the *kth a*ttribute is required to solve the *jth* item, $q_{jk} = 1$, otherwise $q_{jk} = 0$. The combination of skills that examinee i, possesses is represented by an attribute profile, $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iK})$. The attribute profile is a latent vector with length K, in which a 1or 0 in the *kth* entry of the vector indicates mastery or nonmastery of the *kth* attribute respectively. For a test requiring K attributes, examinees will be classified into one of the possible 2^K unique latent classes ($c = 1, \ldots, C$), with each latent class representing a unique combination of attribute mastery and nonmastery patterns. Thus, a primary purpose of CDMs is to appropriately classify examinees into one of the classes based on their mastery or nonmastery of required attributes. The rest of this section provides an overview of the Q-matrix and some commonly used CDMs.

2.1.1 The Q-matrix

The Q-matrix (Tatsuoka, 1983) is a necessary component for analysis involving CDMs as it specifies the attributes that are required to answer each item on a test correctly. The attributes indicate the specific skills, knowledge, or processes that are needed by an examinee to correctly respond to test items. Attributes are discrete and dichotomous which

implies that they are either required for examinees to correctly answer an item (q_{jk} = 1) or not required for examinees to correctly answer the item (q_{jk} = 0). Each item is matched with a corresponding q-vector of length *K*. For example, the entry in the 3rd row of the Q-matrix in Figure 1 indicates that an examinee must master attributes *K*2 and *K*3 to answer item 3 correctly. Thus, by defining the relationship between items and attributes a Q-matrix provides a cognitive specification for each item on a test

Figure 2.1. Example of a Q-matrix to illustrate attribute requirement for items

$$Q = \begin{array}{c} \begin{array}{ccc} K1 & K2 & K3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

For CDMs to be successfully used with educational asssesssment data, it is necessary that the Q-matrix for a test is identified and complete. Completeness is a fundamental requirement that ensures the identifcation of all possible attribute profiles for examinees (Chiu, Douglas, & Li, 2009; Köhn & Chiu, 2016). The Q-matrix for a test based on the DINA or DINO model is said to be complete if and only if it includes all possible single-attribute items. In contrast, the process of establishing compeleteness when other CDMs are used is more complicated. As such, completeness is not assessed based on the structure of a Q-matrix but in reference to the specific CDM in use. Due to this challenge and the difficulties associated with tests that have a large number of items or attributes, assessing the completeness of a Q-matrix is often difficult. As a solution to this issue, Koehn and Chiu (2017) developed a procedure for assessing the completeness of Q-matrices. The procedure is based on the theoritical framework of more generalized CDMs and can therefore be used with CDMs that can be reparameterized as a general CDM. While

an incomplete Q-matrix will affect the identifcation of all possible attribute profiles, problems such as misspecifciations and inaccuracies in the Q-matrix can adverely impact model parameter estimation and lead to misclassification of examinees. To ensure that appropriate inferences about examinees are made based on analysis involving CDMs, it is essential to verify that all required attributes are included in the Q-matrix, and that attributes are appropriately specified for each item on the test. However, the construction of the Q-matrix for most assessments relies heavily on the judgment of content experts, which may lead to errors in the Q-matrix. It is therefore important to develop Q-matrix validation and estimation procedures to ensure that Q-matrices are accurately specified.

2.1.2 The DINA Model

The deterministic inputs, noisy "and" gate (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001) is a conjunctive model since an examinee must possesses all required attributes to answer an item correctly. For each item, the DINA model partitions examinees into two latent groups: examinees in one group have all the required attributes to solve an item correctly, while the examinees in the other group lack at least one of the attributes required to solve the item correctly. This can be attributed to the conjunctive nature of the model, since an examinee must possess all required attributes to correctly answer an item. In the DINA model, the relationship between the latent response variables is represented as follows.

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

where η_{ji}, the ideal item response, shows whether examinee *i* has mastered all the attributes required to answer item *j* correctly. η_{ij} = 1 if examinee *i* has mastered all the required

attributes for item *j*., while $\eta_{ij} = 0$ implies that examinee *i* is missing at least one attribute. The item response function (IRF) for the model is shown below.

$$P(Y_{ij} = 1|\alpha_i) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}$$

Y_{ij} is the observed response for item *j* and examinee *i* and $(1 - s_j)$ is the probability of a correct response by an examinee that has all the required skills. However, since the process is stochastic, examinees can get the item right without having all the required attributes. In addition, examinees can get the item wrong even when all required attributes have been mastered. This is due to the error probabilities, s_j and g_j, the slipping and guessing parameters respectively, which are defined as follows.

$$g_j = P(Y_{ij} = 1|\eta_{ij} = 0)$$

$$s_j = P(Y_{ij} = 0|\eta_{ij} = 1)$$

g_j represents the probability of $Y_{ij} = 1$ when at least one attribute is lacking, and s_j shows the probability of $Y_{ij} = 0$ when all required attributes are present. In the DINA model, the number of parameters for an item is always two, regardless of the number of attributes represented in a Q-matrix. Although the DINA model is one of the most parsimonious CDMs, it is easy to use and interpret (de la Torre, 2008). However, a disadvantage of the model can be attributed to its simplicity since it partitions examinees into two equivalent classes per item and missing one attribute is equivalent to missing all required attributes (Henson & Douglas, 2005).

2.1.3 The NIDA Model

In contrast to the DINA model which does not differentiate between examinees who lack only one of the attributes and those who have not mastered any of the attributes, the noisy input, deterministic, "and" gate (NIDA) model (Maris, 1999; Junker & Sijtsma, 2001) makes a distinction between students who have mastered different combinations of the attributes required to answer an item correctly. Unlike the DINA model, the slipping and guessing parameters in the NIDA model occur at the attribute level, and each attribute uses one slipping parameter and one guessing parameter.

$$S_k = P\left(\eta_{ijk} = 0 \middle| \alpha_{ik} = 1, q_{jk} = 1\right)$$

$$g_k = P\left(\eta_{ijk} = 1 \middle| \alpha_{ik} = 0, q_{jk} = 1\right)$$

η_{ijk} indicates whether examinee *i* has mastered the *kth* attribute required for responding to item *j*.

When an examinee applies attribute *k* correctly for item *i*, $\eta_{ijk} = 1$. $\eta_{ijk} = 0$ when the examinee does not apply attribute *k* correctly. Just like in the DINA model, the parameter α_{ik} is an indicator of attribute mastery for examinee *i*. The guessing parameter (g_k) is the probability of the correct application of attribute *k* in the context of item *j* even though the attribute has not been mastered. Likewise, the slipping parameter (s_k) is the probability of the incorrect application of attribute *k* in the context of item *i* even though the attribute has been mastered. The IRF of the NIDA model is represented as follows.

$$P\left(Y_{ij} = 1 \middle| \alpha_i\right) = \prod_{k=1}^{K} (1 - s_{jk})^{\alpha_{ik}} \, g_{jk}^{(1-\alpha_{ik})^{q_{jk}}}$$

q_{jk} indicates whether attribute *k* is measured by item *i* in the Q-matrix, and $(1 - s_k)$ is the probability of not slipping for attribute *k*. Since the model assumes that the IRF must be

the same for all items that require mastery of the same attributes, the NIDA model is restrictive. An implication of this restriction is that item difficulty for many items will be identical, which is implausible in practice.

2.1.4 The DINO Model

The deterministic input, noisy, "or" gate (DINO) model (Templin & Henson, 2006) is the disjunctive equivalent of the DINA model and just like the DINA model, it has two parameters for each item. The slip parameter (s_j) refers to the probability that examinee *i*, who mastered at least one of the required attributes for item *j*, answered it incorrectly, while the guessing parameter (g_j) refers to the probability of a correct response when an examinee has not mastered any of the required skills. The slip and guessing parameters are represented as follows.

$$s_j = P(x_{ij} = 0|\omega_{ij} = 1)$$

$$g_j = P(x_{ij} = 1|\omega_{ij} = 0)$$

ω_{ij} indicates whether at least one of the attributes required to answer an item correctly has been mastered. Two groups of examinees are represented in the DINO model; examinees that have at least one of the required attributes ($\omega_{ij} = 1$), and examinees who do not have any of the required attributes ($\omega_{ij} = 0$).

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

The IRF for the DINO model is defined as follows.

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

Therefore, in the DINO model the probability of a correct response, given mastery of at least one skill, does not depend on the number and type of skills that are mastered.

2.1.5 The Reduced RUM Model (RRUM)

The reduced reparametrized unified model, RRUM (Hartz, Roussos, & Stout, 2002; Hartz, Roussos, Henson, & Templin 2005; DiBello et al., 2007) is a reduced version of the RUM model in that it omits the Rasch component of the RUM model. RRUM is a generalization of the NIDA model since it allows parameters to differ item-by-item (Chiu, 2013). In the RRUM model, each attribute contributes differently to the probability of a correct response, and the extent to which an attribute contributes to the probability of success can vary from item to item. The model is also based on the assumption that an examinee must master all required attributes to answer an item correctly (Henson et al, 2009). Thus, the RRUM also resolves the issue in the DINA model in which all examinees who have not mastered at least one of the required attributes have the same probability of answering an item correctly. As a result, the RRUM provides for a more flexible impact of attribute mastery on item response probabilities (Rupp, Templin & Henson, 2010). In the item response function, the probability of a correct answer to item j given that an examinee possesses the required attribute pattern α_j is represented as follows.

$$P(Y_{ij} = 1|\alpha_i) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{*qjk(1-\alpha_{ik})}$$

The RRUM model includes two parameters, π^*_j and r^*_{jk}.

$$\pi_j^* = \prod_{k=1}^{K} \pi_{jk}^{\eta_{jk}}$$

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

π^*_j, the baseline parameter, is the probability of a correct response to item *j* provided that an examinee has mastered all the required attributes for the item. π_{jk} is the probability of correctly using the mastered attribute, *k*, to respond to item *j*, and not slipping at the attribute level. A large value of π^*_j for *item j* indicates that an examinee's response correlates with the attributes required to answer item *j* correctly. The penalty parameter r^*_{ik}, *denotes* the amount of reduction to the probability of a correct response to item *j* because of nonmastery of attribute *k*. r_{ik}, is the probability of guessing for attribute *k* when responding to item *k*. When r^*_{jk} for attribute *k* is small, the probability of a correct response is greatly reduced when the attribute is not mastered. Thus, smaller levels of r_{ik} yields higher discrimination levels.

2.1.6 The G-DINA Model

The generalized deterministic inputs, noisy, *"*and*"* gate *(*G-DINA) model, a generalization of the DINA model with more relaxed assumptions, is equivalent to other general models for cognitive diagnosis based on an alternative identity link function (de la Torre, 2011). The G-DINA model incorporates an item-by-item model estimation based on design and weight matrices, and a component for item-by-item model comparison based on the Wald test (de la Torre, 2011). In the G-DINA model, when an examinee has mastered at least one of the required attributes, there is an increase in the examinee's

probability to correctly answer the item. Unlike the DINA model which has 2 parameters for each item *j*, the G-DINA model has $2^{K_j^*}$ parameters for item *j*, where K^*_j is the number of attributes required to answer item j correctly. This accounts for its greater generality whenever $K_j^* > 1$. The DINA model and G-DINA models are however the same when K^*_j =1. The probability of success in the G-DINA model based on the identity link is as follows.

$$P(Y_j = 1|\alpha_{l1}, \alpha_{lK}) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk\alpha_{lk}} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_{j-1}^*} \delta_{jkk'}\alpha_{lk}\alpha_{lk'}\ldots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}$$

where:
δ_{j0} is the intercept for item *j*, i.e., the baseline probability of a correct response when an examinee does not possess any of the required attributes for item *j*
δ_{jk} is the main effect due to α_k i.e., the change in the probability of a correct response when an examinee has mastered a single attribute
δ_{jkk} is the interaction effect due α_k and $\alpha_{k'}$ i.e., the change in probability of a correct response when an examinee masters both α_k and $\alpha_{k'}$.
$\delta_{j12\cdots K^*j}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K^*j}$, i.e., the change in the probability of a correct response when an examinee masters all the required attributes.

The parameter estimation of the G-DINA model is obtained using marginalized maximum likelihood estimation which requires maximizing the log-marginalize likelihood of the response data.

$$l(Y) = \log[LY)] = log \prod_{i-1}^{I} \prod_{l=1}^{L} L(Y_i|\alpha_l)p(\alpha_l)$$

where

$$L(Y_i|\alpha_l) = \prod\prod_{j=1}^{J} P(\alpha_{lj})^{Y_{ij}}[1 - P(\alpha_{lj})]^{(1-Y_{ij})}$$

is the likelihood of the response vector of examinee *i* given the attribute vector α_l.
$P(\alpha_l)$ is the probability of α_l
$P(\alpha_{lj})$ is the probability of a correct response on item *j* and can also be written as $P(\alpha^*_{lj})$

2.2 Q-matrix Validation and Estimation Methods

Although Q-matrix validation and estimation are important aspects of using CDMs and several methods of Q-matrix validation and estimation currently exist, limitations associated with some of these methods makes it necessary to develop additional procedures that can be used with a wide variety of CDMs and with more generalizable application under a variety of conditions. While there are several Q-matrix validation methods that are suitable for use with simple CDMs, the general method of Q-matrix validation (de la Torre & Chiu, 2016) is one of the very few well developed methods of Q-matrix validation that are available for identifying misspecifications in a Q-matrix and that can be used with more complex CDMs. However, the requirement of an arbitrary cutoff to stop the algorithm is a major concern. Q-matrix estimation methods are used for developing a Q-matrix based only on examinees' response data without requiring a provisional Q-matrix. Liu, Xu, and Ying (2012), developed the data-driven approach to identifying a Q-matrix and estimating the associated model parameters. The method is grounded on an estimator that uses the information of the dependence structure of item responses and does not require information about the distribution of the attributes or the slipping or guessing parameters. Some concerns with the use of the method include the need for a sufficiently large sample size which is often unattainable, the excessive computation time which makes the method impractical, and the inappropriateness of the method when both the slipping and guessing parameters are unknown, which is the common practice in reality. In addition, while the approach has been used with the DINA, DINO, and a restricted version of the R-RUM models, the method has not been used with more complex CDMs.

Building on the work of Liu et al (2012), Culpepper, Chen, and Douglas (2018), used the Bayesian framework to develop a Q-matrix estimation method for the DINA model, that is based on the identifiability theory proposed by Chen, Liu, Xu, and Ying (2015). The framework implements an effective Markov chain Monte Carlo (MCMC) algorithm to estimate the Q-matrix by investigating all possible Q-matrices. The method proved to be effective in identifying the Q-matrix both for simulation studies and for empirical applications compared to the method proposed by Chen et al (2015). In addition, unlike the method proposed by Liu et al. (2012), this method eliminates the need for an initial Q-matrix developed by content experts. Since the method estimates the Q-matrix by identifying the one with the most frequent occurrence in the Markov chain, this may be of concern in analysis involving a large number of attributes and examinees.

Q-matrix validation methods (de la Torre, 2008; DeCarlo, 2012; Close, Davison, and Davenport Jr., 2012; Chiu, 2013; de la Torre & Chiu, 2016; Chen, 2017) in contrast, fall in a related but somewhat different category compared to Q-matrix estimation methods. A requirement of these methods is that a provisional Q-matrix associated with the test be known. De la Torre (2008) proposed an empirically based method of Q-matrix validation (the δ method) that is implemented with the DINA model. In the δ method, the difference in the probabilities of a correct response by examinees who have all the attributes required to answer an item correctly ($\eta_j = 1$) and examinees who do not have at least one of the required attributes ($\eta_j = 0$) is based on an item discrimination index, φ_j. φ_j is computed for each item and changes as the q-vector of items changes. A correctly specified q-vector for an item is assumed to be the one for which φ_j is highest since it maximizes the difference between the probabilities of success for the two groups. In contrast, the misspecification of

any q-entry reduces the difference due to either a higher guessing or slipping parameter. By modifying the rule for classifying examinees into groups ($\eta_j = 1$, and $\eta_j = 0$), the method can also be adapted for use with the DINO model. However, the appropriateness of the method with more general CDMs is yet to be determined. DeCarlo (2012) introduced a Bayesian model-based method of Q-matrix validation for data based on the reparameterized DINA (R-DINA; DeCarlo, 2010) model. The method requires a prior identification of q-entries that are likely to be misspecified. These q-entries are identified as random variables and estimated along with other parameters. Although the method showed positive results, further studies are recommended to evaluate its robustness and generalizability and especially in situation in which all misspecified q-entries were not identified (DeCarlo, 2012).

De la Torre & Chiu (2016) developed the general discrimination index (GDI), ς^2, that can be used to validate the Q-matrix with general CDMs. A primary difference between the GDI method and the δ method is that the GDI method can be used with a wide class of CDMs. The GDI first identifies the misspecifications and then substitutes the misspecified entries in a q-vector one q-entry at a time. The method is based on the principle that a correct q-vector should yield homogeneous latent groups with respect to the probability of success (de la Torre & Chiu, 2016). This method was found to be effective in identifying and replacing misspecified q-entries while retaining the correct q-entries. However, despite the benefits of the method such as its generalizability, a major shortcoming of the method is the need to determine in advance an arbitrary cutoff to stop the algorithm. In addition, results of the study based on the method was limited in generalizability because it used specific conditions and did not explore the effect of factors

such as test length, number of attributes, or number of examinees. Chen (2017) developed the residual-based appproach to empirically validate a Q-matrix. The method builds on the absolute fit statistics that is based on the residuals between the observed and expected response patterns (Chen, de la Torre, & Zhang, 2013). Although the approach is suitable for use with several reduced and saturated CDMs that use dichotomous and polythomous attriubtes, additional evidence to support the effecitveness of the fit measures and the process of item adjustment under diverse conditions is required (Chen, 2017).

The Q-matrix validation method proposed by Chiu (2013) is, grounded on the principle that the residual sum of squares (RSS) of a correct q-vector is less than the RSS of the other misspecified vectors for a specific item. The RSS of item *j* for examinee *i* is computed from the observed item response and the ideal item response as shown,

$$RSS_{ij} = (Y_{ij} - \eta_{ij})^2$$

where Y_{ij} and η_{ij} are the observed and ideal item response of examinee *i* to item *j* respectively. Across all examinees, the RSS of item *j* is computed as

$$RSS_j = \sum_{m=1}^{2^k} \sum_{i \in C_m} (Y_{ij} - \eta_{jm})^2$$

where C_m represents examinees' attribute profile *m*. The method uses the nonparametric classification (NPC) procedure (Chiu & Douglas, 2013) to classify examinees. The algorithm for the method initially determines the q-vector that is most likely to be misspecified by identifying the item with the highest RSS. The process begins with the initialization of the search pool and the input Q-matrix which is used to estimate examinees' attribute profiles based on the nonparametric classification method (Chiu & Douglass, 2013). From the estimates of the ideal item responses for obtained, the mean

RSS for each item is determined across examinees for all pairs of observed and ideal responses. The q-vectors for the item with the highest RSS are ranked based on their RSS, and the q-vector with the lowest RSS is added to the input Q-matrix to replace the q-vector for that item. This updated Q-matrix now serves as the input Q-matrix. The input Q-matrix is updated with the q-vector with the lowest RSS for each item and the process is repeated until the RSS of each item remains the same. An advantage of the method is related to its nonparametric nature because the performance of an assessment is not dependent on the quality of the parameter estimations. The method is suitable for use with small and medium sized testing programs because it does not require a large sample size or computation time. Another advantage of the method is that it requires only a few iterations involving $(2^K - 1)\ x\ J$ computations to refine and validate the Q-matrix. As a result, despite situations that may compromise the possibility of obtaining a global optimum, the method is more efficient than other currently proposed algorithms such as Liu et al, 2012 (Chiu, 2013). The method is also viable not only for observed item responses that correspond to the DINA model, but with any CDM that incorporates the ideal item response η, or any ideal item response.

2.3 Joint Maximum Likelihood Estimation (JMLE) for Diagnostic Classification Models

Joint maximum likelihood estimation (JMLE) is a commonly used approach for estimating parameter estimates from response data. In the JMLE approach, both the person parameters θ_i and the item parameters a_j are considered as fixed effects and their estimates are obtained by maximixzing the joint likelihood function, L(α, Θ; Y).

$$L(\alpha, \Theta;\ \mathrm{Y}) = \prod_{i-1}^{N} L_i(\alpha, \Theta;\ Y_i) = \prod_{i=1}^{N} \prod_{j=1}^{J} f(y_{ij}|\theta_j, \alpha_i\ ),$$

where Y represents the NXJ matrix of observed item resposnes, y_i represents the observed item response pattern, and ($\Theta = \theta_1, \theta_2, \ldots, \theta_J$) represents the matrix of item parameters. However in general, JMLE estimators are typicallystatistically incosistent, and in analysis with IRT, estimates for examinees who did not answer any item correctly and examinees who answered all items correctly are excluded (Baker & Kim, 2004; Haberman, 2004). As a result, the JMLE approach is not often used. Chiu, Koehn, Zheng, and Henson (2016) developed a JMLE for CDMs based on Birnbaum' paradigm two-step procedure that resolves the inconsistency issue by using the NPC method (Chiu & Douglas, 2013) to estimate examinees' attribute profiles which are then used as the initial input in the JMLE algorithm. The JMLE for CDM method considers examinees' attribute profiles and item parameters as two distinct entities, with one known and the other unknown. In this case, the known set of parameter is the estimates of examinees atrribute profiles obtained using the NPC method. This reduces the joint likelihood to a function of only the item parameters and the estimates of Θ_j are obtained by maximizing the logarithm of the item likelihood $L_j(\theta_j; y_j, \alpha)$.

$$\log L_j(\theta_j; y_j, \alpha) = \sum_{i=1}^{N} \log(f(y_{ij}|\theta_j, \alpha_i)$$

The item parameters estimates from the above process are then used for re-estimating examinees' attribute profiles by maximizing the (reduced) log-likelihood L($\alpha; Y, \Theta$). The updated examinees attribute profiles are used as input to update the parameter estimates, and through a process of iterations, the examinee attribute profiles and parameter estimates are updated until the estimates converge.

The estimators of item parameters for methods based on the marginal maximum likelihood estimation (MMLE-EM) algorithm or Markov chain Mote Carlo (MCMC) techniques experience computational restrictions when used with complex CDMs, thereby making analysis with large number of replicated datasets unachievable. The JMLE algorithm can overcome this issue for simpler models such as the DINA and DINO models, because the estimators of the item parameters have closed form, which ensure speedy and effective implementations of the EM algorithm (Chiu, Köhn, Zheng, & Henson, 2015).

2.4 The Trends in International Mathematics and Science Study (TIMSS)

2.4.1 Overview of TIMSS

The Trends in International Mathematics and Science Study (TIMSS) is a large-scale international assessment developed and administered by the International Association for the Evaluation of Educational Achievement (IEA), an international organization comprised of research institutions and government agencies. In the United States, the National Center for Education Statistics (NCES) has the primary responsibility of collecting, analyzing, and reporting data on international educational systems to assist local and national education bodies in evaluating the effectiveness of teaching and learning processes, identifying areas in need of improvement, creating educational policies, and making international comparisons (Mullis et al, 2012). Since its first occurrence in 1995, TIMSS has been administered every four years to measure mathematics and science achievement of fourth-grade and eighth-grade students in participating countries. By assessing student achievement in multiple content areas, IEA can gather information that provides insight into the educational processes within individual countries and across

national boundaries. The assessment uses a cross-sectional and quasi-longitudinal nonexperimental design which ensures that the same group of fourth-grade students are assessed in eighth grade. The IEA purposes to use TIMSS to stimulate curricular reforms based on students' performance when they are in fourth grade and to evaluate the effectiveness of the reforms when the students are in eighth grade. Thus, TIMSS uses an international perspective to direct educational policy and practice related to mathematics and science. Since the framework for the assessment is based on broadly defined curriculum, the assessment is generally aligned to curriculum of the participating countries and education systems (Mullis et al, 2012). In addition to the subject-related questions on the assessment, students, teachers, school administrators, and in some instances, parents, provide background information about the instructional contexts of their institutions and factors that affect learning such as characteristics of students, resources available in schools, instructional practices, and family and home support.

To measure student achievement, the assessment is administered to a representative sample of students in each country and educational system. TIMSS 2011, the fifth administration of the assessment, included 57 countries and education systems at the fourth-grade level. The mathematic test is based on a content dimension which focuses on the subject matter and evaluates students in areas such as number, geometric shapes and measures, and data display. The cognitive dimension is based on skills such as knowing, applying, and reasoning. The TIMSS 2011 mathematics assessment consists of 14 blocks of math items with approximately 10-14 items each, from which student booklets are organized. 8 of the 14 blocks are from the 2007 assessment and are used for measuring

trends in the 2011 test. The remaining 6 blocks are made up of released items, from which items used in this study were obtained.

2.4.2 Research on TIMSS Fourth-grade and Eighth-grade Mathematics

Since its inception in 1995, many researchers and organizations, including the NCES, have analyzed the TIMSS data to evaluate fourth and eighth grades students' achievement in mathematics and science. The NCES typically provides results based on performance by average scores and performance on international benchmarks. Reports from the analysis of the 2011 TIMSS by NCES (Mullis et al, 2012) show that average mathematics score for the United States was higher than the international TIMSS scale average. According to the report, the United States was one of the top 15 education system in mathematics, with 8 education systems (Singapore, Korea, Hong Kong-CHN, Chinese Taipei-CHN, Japan, Northern Ireland-GBR, North Carolina-USA, and Belgium (Flemish)-BEL) having higher averages and 6 nations having similar scores. The United States also scored higher on average than 42 education systems. In addition, the report showed an increase in average scores over time for the United States in 2011 compared to 1995 (23 points higher) and 2007 (12 points higher). Differences in performance were also noted within the United States, with North Carolina scoring above the TIMSS scale average and the United States national average in mathematics, while Florida scored above the TIMSS scale average but was not measurably different from the United States national average.

Using a diagnostic-based model, the rule-space method, Tatsuoka et al (2004) analyzed the mathematics data from the revised TIMSS-R, 1999 test. The analysis compared mastery of 23 attributes among 20 countries including the United States. The

result showed significant differences among the countries in their mastery of the attributes. For example, students from the United States showed strong quantitative reading skills but were weaker in areas such as geometry. In their analysis of the 2007 fourth-grade TIMSS mathematics data, Lee et al (2011) focused on the results of two benchmark participants, Massachusetts and Minnesota, with a goal to provide comparison both within and across the United States. The research was based on 25 items encompassing 15 attributes and estimated with the DINA model. The results showed that although both states significantly outperformed the United States in general, there were significant differences in the proportion of items correctly answered and the level of skill mastery. The study also included an evaluation of the model fit between the DINA model and IRT models. The comparison showed that the DINA model had a better model fit and provides more reliability and integrity in terms of the interpretation and meaning of the results (Lee et al, 2011).

Park and Lee (2011) used a cluster analysis to analyze items from the TIMSS 2007 fourth-grade mathematics assessment. To conduct K-means clustering and hierarchical agglomerative cluster analysis (HACA), the study clustered attributes by mapping item responses to an attribute matrix (Park & Lee, 2011). In the study, countries were classified based on their average scale scores as high-performing (Hong Kong SAR and Chinese Taipei), average-performing (Denmark, Sweden, and the United States), and low-performing (Colombia, Kuwait, Qatar, and Yemen). The results indicated that attribute structure for higher-performing countries were explicit and had a more hierarchical structure than the structure of attributes evident in the lower-performing countries. Choi et al (2015) used the DINA model to reanalyze the TIMSS 2003 eighth-grade mathematics

test to compare the performance of students in the United States and Korea. According to the authors, the Q-matrix was specifically constructed for the assessment, by specifying attributes based on the Principles and Standards for School Mathematics published in 2000 by the National Council of Teachers of Mathematics (Choi et al, 2015). The standards were adapted to fit the concepts specified in the items on the assessment. The result compared the discrimination index of the two countries. Based on the findings of the study, the DINA model is recommended for use in empirical research involving large-scale assessments.

Previous research using the TIMSS data focused mostly on comparing the performance of participating countries and educational system using their average scale scores, and retrofitting CDMs to TIMSS data to identify differences in attribute mastery among countries and educational systems. These studies however did not include the estimation or validation of the Q-matrices to ensure that the Q-matrices are correctly specified. The goal of this study is to fit the data to a CDM to identify and correct possible discrepancies in student classification. The proposed Q-matrix validation and estimation methods will then be used to identify the correct Q-matrix after which the data will again be fitted to a CDM to correct possible discrepancies in student classification.

Chapter 3. Analysis of the TIMSS 2011 Fourth-Grade Mathematics Dataset: Part I

The primary purpose of this dissertation study is to analyze a subset of data from TIMSS using a cognitive diagnostic model and an adhoc Q-matrix developed through a process of retrofitting. In the first part of the analysis shown in this chapter, data from eleven countries and educational systems is organized into three groups, with the USA in a fourth group by itself. The data is first analyzed to provide a description of the data set and information about how examinees in each group performed on the test based on their average scores. The data are then fitted with the G-DINA model to obtain examinees' estimated attribute profiles. Part I of the analysis includes a summary of the descriptive statistics, a comparison of group performance by item and by attribute, and examinees' estimated attribute profiles obtained from fitting the data set with the G-DINA model. The estimated attribute profiles will be reviewed with the proportion correct by attribute to ensure that there no discrepancies in classification.

3.1 The Dataset

The study analyzes examinee data for the United States, Singapore, Hongkong, Republic of Korea, Chinese Taipei, Japan, Finland, England, Denmark, Germany, Canada, and Australia. To ensure adequate sample sizes, the countries excluding the United States are classified into three groups based on their average scores provided in the National Center for Education Statistics (NCES) report 2013. The three identified groups are high-performing (Singapore, Korea, Hong Kong, Chinese Taipei), mid-performing (Japan, Finland, England, and Denmark), and low-performing (Canada, Germany, and Australia).

Table 3.1 shows the countries and educational systems in each group and the average score for each country and educational system.

Table 3.1. Average Mathematics Scores of 4th-grade Students on the TIMSS 2011 by country and educational system

Group	Country/Educational System	Average score (Number domain)	Number of Examinees
1 High Performing	Singapore	619	428
	Republic of Korea	606	279
	Hongkong	604	245
	Chinese Taipei- CHN	599	259
2 Mid performing	Japan	584	272
	Finland	545	241
	England	539	192
	Denmark	534	162
3 Low performing	Quebec-CAN	531	215
	Germany	520	199
	Ontario-CAN	504	266
	Australia	508	341
	Alberta-CAN	505	205
4 USA	United States of America	**543**	**798**

The data set consists of responses to 15 released multiple-choice and open-ended items from booklet 6 of TIMSS 2011 fourth-grade mathematics assessment. 8 of the 15 items are multiple choice items and 7 are open-ended items that require students to compute a number or to draw a diagram to illustrate a pattern. Correct responses to the multiple-choice items were coded as 1 and incorrect responses as 0. Correct responses to the open-ended items, were coded as 1 while incorrect responses and responses with partial credits were coded as 0. Information for students with omitted responses is not included in the data set. Items on the TIMSS fourth-grade mathematics test are broadly categorized into two domains: (1) content domains which identify students' knowledge of subject matter and

(2) cognitive domains, which describe thinking processes that are required to answer the items. The content domain includes topics such as number, geometric shapes and measures, and data display, each of which cover 50%, 35%, and 15% of the test respectively. The cognitive domain test skills such as knowing, applying, and reasoning, each covering 40%, 40%, and 20% of the test respectively. Each item on the TIMSS is classified into both domains, however for this analysis, only the content domain is considered, and all 15 items test examinees' knowledge of topics related to the number content domain. The items require students to use addition, subtraction, multiplication, and division skills to solve problems involving whole numbers and fractions.

3.2 Construction of the Q-matrix

The framework for the TIMSS fourth-grade mathematics assessment includes 15 objectives in the number domain that relate to the content matter covered in the mathematics curriculum of participating countries. To determine the attributes that would be included in the Q-matrix, first each item on the test was matched to an objective in the framework. However, since the 15 items were matched to 11 of the objectives in the framework, objectives that require similar skills were consolidated with the help of a mathematics content specialist to avoid overlap and to reduce the number of attributes that will be represented in the Q-matrix. Table 3.2 outlines the objectives in the number domain with the corresponding items and attributes.

Table 3.2. TIMSS fourth-grade mathematics framework for number domain showing items and attributes that correspond to each objective

Domain	Objective	Item	Attribute
Whole Numbers	1. Demonstrate knowledge of place value, including recognizing and writing numbers in expanded form and representing whole numbers using words, diagrams, or symbols	*N/A*	N/A
	2. Compare and order whole numbers.	*J*2,	*K*1
	3. Compute with whole numbers (+, −, ×, ÷) and estimate such computations by approximating the numbers involved	*J*4	
	4. Recognize multiples and factors of numbers.	*J*3, *J*7, *J*10, *J*14	*K*2
	5. Solve problems, including those set in real-life contexts and those involving measurements, money, and simple proportions	*J15*	*K*4
Number: Fractions and Decimals	6. Show understanding of fractions by recognizing fractions as parts of unit wholes, parts of a collection, locations on number lines, and by representing fractions using words, numbers, or models.	*J*8	*K*5
	7. Identify equivalent simple fractions; compare and order simple fractions	*J*5	
	8. Add and subtract simple fractions.	*J*9	
	11. Solve problems involving simple fractions or decimals	*J*9	
	10. Add and subtract decimals.	N/A	N/A
	9. Show understanding of decimal place value including representing decimals using words, numbers, or models.	N/A	
Number Sentences with Whole Numbers	12. Find the missing number or operation in a number sentence (e.g., 17 + ■ = 29).	*J1, J11*	*K*3
	13. Model simple situations involving unknowns with expressions or number sentences.	*J12*	
Number Patterns and Relationships	14. Extend or find missing terms in a well-defined pattern, describe relationships between adjacent terms in a sequence and between the sequence number of the term and the term.	*J*6, *J*10, *J*13	*K*6
	15. Write or select a rule for a relationship given some pairs of whole numbers satisfying the relationship and generate pairs of whole numbers following a given rule (e.g., multiply the first number by 3 and add 2 to get the second number)	*J*7, *J*14	

These 6 attributes define the skills required to correctly solve the 15 items included in this study. While some attributes are a direct match to one objective, other attributes cover more than one objective. For example, attribute *K5* requires skills related to fractions and covers four of the objectives in the framework. As shown in Table 3.3, 12 of the items require examinees to master a single attribute to correctly answer the items, while 3 of the items require multiple attributes.

Table 3.3. Items and Q-matrix

Item (*J*)	Identification number/Item Description	Q-Matrix					
		K1	K2	K3	K4	K5	K6
1	M041107: Identify the correct number sentence	0	0	1	0	0	0
2	M041011: Compare numbers	1	0	0	0	0	0
3	M041122: Circle the factors of 12	0	1	0	0	0	0
4	M041041: Estimate the product of two numbers	1	0	0	0	0	0
5	M041320: Identify the equivalent of a fraction	0	0	0	0	1	0
6	M041115A: Draw the missing pattern	0	0	0	0	0	1
7	M041115B: Determine the number of squares in a given pattern	0	1	0	0	0	1
8	M031210: Compare fractions	0	0	0	0	1	0
9	M031009: Word problem - fractions	0	0	0	0	1	0
10	M031252: Finding missing number in a pattern of 3 6 9 12	0	1	0	0	0	1
11	M031316: Find the missing number	0	0	1	0	0	0
12	M031317: Find the missing number	0	0	1	0	0	0
13	M031079B: Find the missing number pattern	0	0	0	0	0	1
14	M031079C: Determine the number of circles in a given pattern	0	1	0	0	0	1
15	M031043: word problem (time)	0	0	0	1	0	0

K1 = comparing whole numbers, computing with whole numbers, and estimating
K2 = recognizing multiples and factors
K3 = find the missing number or missing operation
K4 = solve real-life problems involving measurements, money, and time
K5 = understanding of fractions, fraction equivalent, and solving problems involving simple fractions
K6 = describing relationship in patterns, their extension, and generating numbers based on a given rule

3.3 Results

3.3.1. Descriptive Statistics

Table 3.4 provides a summary of the descriptive statistics for the dataset. The average score for the high-performing group is 57 points higher than the average score for the mid-performing group, while the mid-performing group has an average score that is about 37 points higher than the low-performing group. USA has an average score that is about 30 points higher than that of the low-performing group, but 7 points lower than that of the mid-performing group. The range, the differences between the lowest and highest average score for each group, shows that the largest difference occurs in the mid-performing group. There are also large differences between the median average scores of the high-performing and low-performing groups.

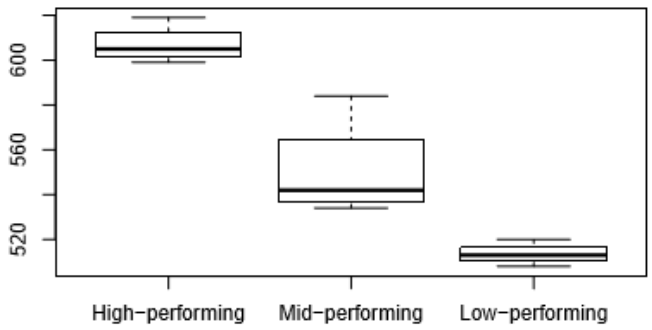
Table 3.4. Summary of Descriptive Statistics

	USA	High-Performing Group	Mid-Performing Group	Low-Performing Group
Number of examinees	798.00	1211.00	867.00	1226.00
Mean	543.00	607.00	550.50	513.60
Median		605.00	542.00	508.00
Standard Deviation		8.52	22.78	11.63
Min		606.00	534.00	504.00
Max		619.00	584.00	531.00
Range		20.00	50.00	27.00
Skewness		0.48	0.66	0.49
Kurtosis		-1.82	-1.75	-1.82

The standard deviation estimates imply that there is more variability in the performance of examinees in the mid-performing group. The variability of the average scores is also illustrated by the boxplots in Figure 3.1. Based on the skewness and kurtosis values, both

distributions are slightly skewed to the right and lightly tailed relative to a normal distribution.

Figure 3.1. Box Plot Comparing Group Average Scores



3.3.2 Comparison of group performance by item and by attribute

Table 3.5 provides information about how well each group performed on the items and the proportion of examinees from each group that answered each item correctly. Items 1 and 11 had the highest number of examinees that responded correctly. Both items require K3 (finding the missing number or missing operation), which may indicate generally that, most examinees in all participating countries are proficient in that specific skill or that the items are easy. On the other hand, the items in which each group performed the poorest was different across groups. For example, examinees in USA performed poorly on items 3 and 14, the high-performing group performed poorly on items 3 and 9, the mid-performing group performed poorly on items 3, 12, and 14, while the low-performing group performed poorly on items 3 and 12. This may possibly be a result of the proficiency

level of each group, differences in content emphasized in their curriculum, or the difficulty level of the items for the examinees.

Table 3.5. Proportion of Items Correct by Group

Item	Required attribute	USA	High-performing group	Mid-performing group	Low-Performing group
*J*1	*K*3	0.94	0.95	0.94	0.63
*J*2	*K*1	0.82	0.93	0.77	0.51
*J*3	*K*2	0.47	0.60	0.28	0.20
*J*4	*K*1	0.72	0.82	0.73	0.51
*J*5	*K*5	0.81	0.82	0.71	0.41
*J*6	*K*6	0.72	0.86	0.80	0.54
*J*7	*K*2, *K*6	0.64	0.80	0.62	0.42
*J*8	*K*5	0.66	0.68	0.65	0.38
*J*9	*K*5	0.49	0.57	0.63	0.37
*J*10	*K*2, *K*6	0.86	0.88	0.84	0.56
*J*11	*K*3	0.89	0.97	0.90	0.57
*J*12	*K*3	0.52	0.80	0.48	0.20
*J*13	*K*6	0.77	0.93	0.88	0.59
*J*14	*K*2, *K*6	0.46	0.69	0.48	0.28
*J*15	*K*4	0.60	0.82	0.66	0.40

Table 3.6 shows the attribute prevalence estimates, denoted as P_k, for $k = 1,\ldots, 6$, indicating the proportion of examinees in each group who have mastered attribute *k*. P_6 is highest for all groups, P_5 is lowest for the mid-performing group, and P_2 is lowest for the other three groups. This could be interpreted that in all countries, most examinees have mastered the concept of patterns and relationships, while most of them are yet to reach proficiency in factors and multiples (*K*2). The result might also suggest that items requiring examinees to use knowledge of patterns and relationships are relatively easier than the items that require examinees to use their knowledge of factors and multiples. Examinees in the high-performing group were more proficient on all 6 attributes compared to the other

three groups. A comparison of the other three groups shows that the mid-performing group had higher attribute prevalence for *K*1, *K*2, *K*3, *K*4, and *K*6, while the United States had a higher attribute prevalence for *K*5. This result suggests that examinees from the mid-performing group may have found items that require *K*5 difficult or that they have not really mastered concepts related to fraction. This result is also of note because although

Table 3.6. Attribute Prevalence Estimates by Group

	USA	High-performing group	Mid-performing group	Low-Performing group
*K*1	0.68	0.80	0.68	0.61
*K*2	0.56	0.69	0.61	0.52
*K*3	0.60	0.75	0.70	0.61
*K*4	0.55	0.73	0.63	0.54
*K*5	0.59	0.73	0.58	0.54
*K*6	0.69	0.82	0.73	0.69

there is a large difference in average scores between the mid-performing group (584) and the United States (543), examinees from United States seem to be more proficient in *K*5, fractions, than the examinees in the mid-performing group. A comparison of the attribute prevalence estimates for the low-performing group and the United States shows that although the United States has higher estimates for the 6 attributes, the number of examinees who have mastered for *K*3, *K*4, and *K*6 are very similar for both groups.

3.3.3 Review of examinees' estimated attribute profiles

To obtain examinees' estimated attribute profile, the data for each group was fitted with the G-DINA model using the provisional Q-matrix. The model fitting procedure was implemented in the R package G-DINA (Ma & de la Torre, 2018). Examinees' estimated

attribute profiles were evaluated by reviewing examinees' estimated attribute profiles and proportion correct by attribute. to gain more insight into whether examinees with the same proportion correct for an attribute are classified the same for that attribute. The proportion correct by attribute is a measure of the number of times an examinee correctly applied an attribute. For example, *K*3 is required for items 1, 11, and 12. An examinee with a proportion correct estimate of 0.67 for *K*3, successfully applied this attribute 2 out of the 3 required times.

Table 3.7. Estimated attribute profiles, item response profiles, and proportion correct estimates for a random sample of examinees

Examinee ID	Estimated Attribute Profiles	Item Response Profile	Proportion Correct by Attribute					
1065	111000	110100111111100	1.00	0.50	1.00	0.00	0.67	0.60
3130	111111	110100111111101	1.00	0.50	1.00	1.00	0.67	0.60
1500	100101	110101011110111	1.00	0.50	0.67	1.00	0.67	0.80
2111	100101	110101011110111	1.00	0.50	0.67	1.00	0.67	0.80
1060	111111	110101011111111	1.00	0.50	1.00	1.00	0.67	0.80
163	000000	110110000111100	1.00	0.25	1.00	0.00	0.33	0.40
226	111101	110110000111111	1.00	0.50	1.00	1.00	0.33	0.60
4028	111100	110111000111001	1.00	0.25	1.00	1.00	0.33	0.40
1235	101101	110111000111101	1.00	0.25	1.00	1.00	0.33	0.60
3439	000001	110110100110110	1.00	0.75	0.67	0.00	0.33	0.80
1092	111111	110110101110101	1.00	0.50	0.67	1.00	0.67	0.60
3155	111110	110110011110101	1.00	0.25	0.67	1.00	1.00	0.40
1931	100010	110110011110110	1.00	0.50	0.67	0.00	1.00	0.60
3136	101010	110110010111100	1.00	0.25	1.00	0.00	0.67	0.40
4075	111110	110110010111101	1.00	0.25	1.00	1.00	0.67	0.40
3252	100011	110111111110100	1.00	0.50	0.67	0.00	1.00	0.80

The results for a random sample of examinees from each group shown in Table 3.7 indicates that there are discrepancies in classification and that there may be some inconsistencies in the estimated attribute profiles. For example, examinees 1065 and 3130 have the same proportion correct by attribute for K5 and K6, indicating that the two

examinees applied the attributes correctly the same number of times. While examinee 3130 is shown to have mastered both attributes, examinee 1065 is classified as nonmastery for both. Similarly, the proportion correct estimates for *K*2 and *K*5 are the same for examinees 1500, 2111, and 1060. However, examinee 1060 is shown to have mastered the two attributes while the other two examinees are classified as nonmastery for the attributes. Additional inconsistencies observed with examinees 163 and 226 show that examinee 226 is classified as mastery for K1 and K3 while examinee 63 is classified as nonmastery for both even though the proportion correct estimates are the same. In another situation, examinees 1235 and 4028 have the same proportion correct estimate for *K*2. Examinee 4028 is classified as mastery while examinee 4028 is classified as nonmastery for the attribute.

These inconsistences in examinee classification may be due to reasons such as misspecifications in the provisional Q-matrix or model misfit. Since the provisional Q-matrix was developed through a process of retrofitting based on the judgement of content experts, the question of possible misspecifications in the Q-matrix arises. One way to resolve this issue is to use a Q-matrix validation procedure to correct potential misspecifications in the provisional Q-matrix or to use a Q-matrix estimation method to establish the Q-matrix from scratch. Although there are several existing methods of Q-matrix validation and estimation, only very few well-developed methods of Q-matrix validation are available for identifying misspecifications in a Q-matrix (de la Torre & Chiu, 2016). In the chapters that follow, one Q-matrix validation procedure and one Q-matrix estimation procedure are proposed that can be used to correct the misspecifications in the provisional Q-matrix and estimate a Q-matrix using examinees item response profiles,

respectively. The TIMSS data will be analyzed again in a later chapter using the modified Q-matrices obtained from the proposed methods. Examinees' updated estimated attribute profiles will then be re-evaluated with the special focus of investigating whether the inconsistencies discovered in the Part I analysis were remedied.

Chapter 4. Methods

In this chapter, the algorithms for the integrated Q-matrix validation (IQV) method and the two-step Q-matrix estimation (TSQE) method will be explained.

4.1 The Integrated Q-matrix validation (IQV) Method

The algorithm for the IQV method consists of a two-part procedure that integrates joint maximum likelihood estimation (JMLE) for cognitive diagnostic models (Chiu, Koehn, Zheng, & Henson, 2016) and the Q-matrix validation method (Chiu, 2013). The primary purpose of the JMLE algorithm for CDMs is to estimate examinees' attribute profiles using the provisional Q-matrix that was developed based on content expert knowledge. The examinee attribute profile estimates are then used in the Q-matrix validation method (Chiu, 2013) to recover the misspecified Q-matrix. The steps of the IQV method are as follows.

1) Examinees' attribute profiles, which will serve as the input into the JMLE function, are estimated using the nonparametric classification (NPC) method (Chiu & Douglas, 2013). The NPC method estimates examinees' attribute profiles by comparing their observed response profiles with each of the ideal response profiles of all possible proficiency classes as follows.
    a) For each examinee, determine the ideal response for item j, where $j = 1,\ldots, J$. Compare each of the ideal response profiles of the possible attribute classes with examinees' observed item response profiles.
    b) Determine the NPC estimator $\tilde{\alpha}$ by minimizing the distance between the ideal item response profiles and observed item response profile.
2) Examinees' attribute profiles, $\tilde{\alpha}$, obtained from step 1 then serve as input to initialize the JMLE algorithm. Because examinees' attribute profiles have now

been determined, item parameter estimates β_{j0}, β_{jk}, and $\beta_{jkk'}$ are obtained using Equation 18 in Chiu, Koehn, Zheng, and Henson (2016). Through a process of iterations, $\tilde{\alpha}$ is then updated.

3) The Q-matrix refinement method is then applied to obtain the corrected Q-matrix. The examinees' attribute profile estimates obtained at the completion of step 2 serves as input for the initial step of the Q-matrix refinement method. The mean residual sum of square (RSS) for each item across examinees are computed for each observed response and corresponding ideal response to each item. Starting with the item with the highest RSS, the q-vectors are updated through a series of steps and the correct q-vector for each item is obtained by minimizing the RSS for each item. The corrected Q-matrix is identified when the RSS of each item remains the same.

A primary advantage of the IQV method is that, it can be used with more general CDMs beyond the DINA and DINO models. In addition, compared to the general method of Q-matrix validation (de la Torre & Chiu, 2015) which can also be used with more general CDMs, the IQV method does not involve the process of determining an arbitrary cutoff in advance to stop the algorithm.

4.2 The Two-Step Q-matrix Estimation (TSQE) Method

The two-step Q-matrix estimation (TSQE) method is grounded in nonlinear factor analysis (FA) and the Q-matrix refinement method (Chiu, 2013). FA is a multivariate statistical approach that can be used to determine the variability among observed, correlated variables by reducing a large number of variables into a smaller set of variables, based on the loadings of the variables on the factors. FA determines both the number of

factors within a set of variables and the extent to which each variable is representative of the factors (i.e., the loading values). The loadings show the extent to which each of the variables contributes to the variance. Attributes that are required to answer each item correctly are identified based on a predetermined factor loading threshold value λ. The purpose of the threshold value is to determine variables that load onto a factor adequately and can be considered as being associated to the factor. The threshold value also establishes when variable loads onto too many factors in which case the variable can be marked as not being representative of any factors (Howard, 2016). For this study, λ values of 0.8 and 0.9 were used.

To estimate the Q-matrix, the matrix of tetrachoric correlations between variables are first computed. This pairwise correlation matrix *X*, is then used as the input for FA to obtain an initial Q-matrix. The Q-matrix refinement method is applied as a second procedure to estimate the true Q-matrix. Since the attributes are dichotomous, the tetrachoric correlation is appropriate for computing the pairwise correlation estimates for each pair of variables. The algorithm is described in the following steps.

1) First the tetrachoric correlation matrix for all item pairs is computed.
2) FA is applied to obtain a provisional Q-matrix.
    a) Maximum likelihood estimation is used to find the common factors of the data, where the common factors represent the attributes in the Q-matrix.
    b) The communality, the amount of variance explained by FA, is computed as one minus the uniqueness for item *j*, is estimated.

$$1 - e_j = \sum_{k=1}^{K} l_{jk}^2$$

where e_j is the uniqueness of item *j* and l_{jk}^2 is the squared loading for the *jth* item and the *kth* attribute.

c) To determine which attributes load on item *j*, first the squared loadings of each item is ranked from largest to smallest: $l_{j(K)}^2, \ldots, l_{j(1)}^2$.

d) The cumulative sums of the ordered squared loadings (s_j) are calculated.

$$s_{j1} = l_{j(K)}^2, \;\; s_{j2} = l_{j(K)}^2 + l_{j(K-1)}^2, \ldots, \;\; s_{j(K)} = \sum_{k=1}^{K} l_{j(k)}^2$$

e) The cumulative sums are divided by the commonality to determine the proportion of the variance explained by the addition of each attribute.

$$A_{jk} = \frac{s_{jk}}{1 - e_j}$$

where: A_{jk}denotes the squared loadings $l_{j(K)}^2, \ldots, l_{j(1)}^2$ that explains at least λ percent of the explained variance.

f) For each item, the cumulative sum for which $A_{jk} \geq \lambda$ represents the sum of variances of the attributes that are required to answer the item correctly. Specifically,

$$q_{jk} = \begin{cases} 1 \text{ for } A_{jk} \geq \lambda \\ 0 \text{ for } A_{jk} < \lambda \end{cases}$$

g) All attributes that contribute to the squared loadings for which $A_{jk} \geq \lambda$ are designate in the Q-matrix as 1 (required), while attributes that did not contribute are designated as 0 (not required).

3) The initial Q-matrix obtained at the end of step 2 is used as the input for step 0 of the Q-matrix refinement method to estimate examinees' attribute profiles. Through a series of steps, the mean residual sum of square (RSS) across examinees for each observed response and corresponding ideal responses to each item are computed, and the q-vector with the lowest RSS is identified for each item.

It is important to note that when FA is applied, the factors are ordered in decreasing order of loadings and includes both factors that are considered relevant and those that are assumed to reflect measurement error or noise. Therefore, a rotation is required to simplify the interpretation of the factors that are considered relevant. After a varimax orthogonal rotation, an attribute or attributes that contribute to the specified threshold value (λ) of the explained variance of an item are considered as the attribute(s) required to correctly answer the item. Although FA is primarily used for analyzing latent variables with continuous distributions and CDMs analyze discrete variables, this compatibility issue is resolved by using tetrachoric correlation instead of Pearson correlation as the input for FA. The calculation of tetrachoric correlation is based on the assumption that the variables are dichotomous.

There is currently limited research on the use of FA for Q-matrix estimation and the criteria for determining the threshold value. However, a review of research using FA in the context of cyberpsychology and human-computer interaction (Hinkin, 1995, 1998; Costello and Osborne, 2005; Tabachnik and Fidel, 2001, 2007; Hair, Black, Babin, Anderson, and Tatham, 2006) provides several recommendations for determining λ, the threshold value, with suggested values ranging from 0.32 to above 0.45, (as cited in Howard 2016). Howard (2016) also recommends that satisfactory variables should have

loadings of above 0.4. Since preliminary analysis using the proposed TSQE method showed no difference between high λ values, 0.8 and 0.9 were used as the threshold for this study to ensure that attributes identified as required for each item are highly correlated to the items.

Chapter 5. Simulation Studies

To evaluate the performance of the proposed methods, simulation studies were designed across a variety of conditions and with different CDMs. The first two studies evaluate the effectiveness of the IQV method for validating a provisional Q-matrix that is developed based on the judgment of content-area experts, by comparing this Q-matrix with a modified Q-matrix obtained through the validation process. The third study evaluates the effectiveness of the TSQE method for estimating a Q-matrix from scratch by using only the information from examinees' responses. The fourth study was designed to compare the performance of the proposed methods under the same conditions.

5.1 Simulation Studies for the Integrated Q-Matrix Validation (IQV) Method

The performance of the IQV method is evaluated with two simulation studies using the DINA and G-DINA models. Study 1 evaluates the effect of the percentage of misspecifications on the recovery of the Q-matrix, while study 2 evaluates the effect of the type of misspecification on recovery of the Q-matrix.

5.1.1 Simulation Study 1

5.1.1.1 Simulation Study Design

The simulation studies were conducted with four J by K Q-matrices with J = 20 or 30, and K = 3 or 5. These four Q-matrices were used to generate data for the study. For each item, the number of required attributes were between 1 and 5, while for each dataset the number of examinees N ranged between 1000 to 3000. The misspecified Q-matrices were constructed by randomly changing 10% or 20% of q-entries in a correct Q-matrix from 0 to 1 or from 1 to 0. Data generation and model specification were carried out using

R programming language (R Core Team, 2016), a freely available software. Each condition consisted of 25 replications. Examinee responses were generated from the DINA or G-DINA models. The slipping (s_j), and guessing (g_j), parameters ranged between 0.1 and 0.3. Examinees' attribute profiles were generated based on the multivariate normal threshold model, in which the attributes are correlated instead of being independent, and attribute patterns do not have equal probabilities of occurrence. Variances and covariances were set to 1.0 and 0.5 respectively

Table 5.1. Q-matrices: J=20; K=3

	Correct Q-matrix			10% Misspecified Q-matrix		
Item #	K1	K2	K3	K1	K2	K3
1	1	0	0	1	0	0
2	0	1	0	0	1	0
3	0	0	1	1	0	1
4	1	1	0	0	1	0
5	1	0	1	1	0	1
6	0	1	1	0	0	1
7	1	1	1	1	1	1
8	1	0	0	1	0	0
9	0	1	0	0	1	0
10	0	0	1	0	1	1
11	1	1	0	1	1	0
12	1	0	1	1	0	1
13	0	1	1	0	1	1
14	1	1	1	1	1	0
15	0	0	1	1	0	1
16	0	1	0	0	1	0
17	0	1	1	0	1	1
18	0	1	1	0	1	1
19	1	1	1	1	1	1
20	1	1	1	1	1	1

.

The simplest of the correct Q-matrices, where J = 20 and K = 3, includes 8 one-attribute, 8 two-attribute, and 4 three-attribute items, as shown in the left panel of Table 5.1 The Q-matrix when J = 20 and K = 3, includes 5 each of one-attribute, two-attribute and three-attribute items, 4 four-attribute items, and 1 item with five attributes. The misspecified Q-matrix where J = 20 and K = 3 is shown on the right panel of Table 5.1. The misspecified Q-matrices served as the input for Step 0 of the Q-matrix refinement method algorithm.

5.1.1.2 Measures

The results of the analysis will be evaluated using the mean recovery rate (MRR), the sensitivity rate (SEN), and the specificity rate (SPE). The MRR is the average percentage of q-entries in the modified Q-matrix that are identical to the q-entries in the correct Q-matrix. High MRRs indicate the effectiveness of the method to recover the correct Q-matrix from a misspecified one.

$$MRR = \frac{\sum_{r=1}^{25}\sum_{k=1}^{K}\sum_{j=1}^{J} I\left[q_{jkr} = \hat{q}_{jkr}\right]}{K \times J \times 25}$$

where:
q_{jkr}= q-entries in the correct Q-matrix
$\hat{q}_{jkr}$= q-entries in the modified Q-matrix
The sensitivity rate (SEN), a measure of the proportion of misspecified q-entries that are corrected and the specificity rate (SPE), a measure of the proportion of correct q-entries that are retained, are computed as follows.

$$SEN = \frac{fp}{fp + tn}$$

$$SPE = \frac{tp}{tp + fn}$$

where:
tp represents true positive, the number of correctly specified q-entries that were retained
fp represents false positive, the number of misspecified q-entries that were corrected
fn represents false negative, the number of correctly specified q-entries that changed
tn represents true negative, the number of misspecified q-entries that were not corrected

MRR, SEN, and SPE close to or equal to 1 are desirable.

5.1.1.3 Results

The results for examinee responses generated from the DINA model indicates that all misspecified q-entries were corrected and all correct q-entries were retained, since MRR, SEN, and SPE are 1.0 for all conditions. The results for the G-DINA model are summarized in Table 5.2. MRRs for Q-matrices with 10% misspecification were between 0.82 and 0.87, and between 0.81 to 0.86 for Q-matrices with 20% misspecifications. A decrease in MRR, SEN, and SPE is observed as *K* increases from 3 to 5 for all values of *N*

Table 5.2. Simulation Study 1: Mean Recovery Rate (MMR): G-DINA Model

			10% Q-matrix Misspecification			20% Q-matrix Misspecification		
N	*J*	*K*	MRR	SEN	SPE	MRR	SEN	SPE
1000	20	3	0.87	1.00	0.86	0.85	0.83	0.86
	20	5	0.82	0.90	0.81	0.81	0.80	0.81
	30	3	0.87	0.89	0.87	0.85	0.89	0.84
	30	5	0.82	0.87	0.81	0.80	0.87	0.78
2000	20	3	0.85	0.83	0.85	0.84	0.83	0.84
	20	5	0.83	0.80	0.83	0.83	0.80	0.84
	30	3	0.87	1.00	0.86	0.84	0.89	0.83
	30	5	0.86	0.87	0.86	0.84	0.83	0.84
3000	20	3	0.88	0.83	0.89	0.85	0.83	0.86
	20	5	0.82	0.80	0.82	0.82	0.75	0.84
	30	3	0.87	1.00	0.86	0.86	0.83	0.87
	30	5	0.83	0.87	0.83	0.80	0.83	0.84

and *J*. The results also show that MRR increases as *J* increases, while *N* does not seem to influence MRR, SEN, or SPE. Higher SEN and SPE rates were also noted when $J = 30$. Similar trends were observed for simulation with 20% Q-matrix misspecifications. However, MRR and SEN were lower than for the simulations with 10% Q-matrix misspecifications. While SEN were observed to decrease as the percentage of misspecification increased, no distinct association was noted between SPE and the percentage of misspecifications.

5.1.2 Simulation Study 2

The goal of Study 2 is to determine the effect of q-entry misspecification and q-vector misspecification on the recovery of the Q-matrix.

5.1.2.1 Simulation Study Design

Two correct Q-matrices with $J = 30$ and $K = 3$ or 5 from Study 1 were modified by introducing misspecifications either by q-entry denoted as $Q_{mis.e}$, or by q-vector denoted as $Q_{mis.v}$. To create the Q-matrices with misspecified q-vectors, the number of misspecified q-entries in a q-vector was fixed to 1 and the number of misspecified q-vectors in each Q-matrix ranged from 1 to 10. To create Q-matrices with misspecified q-entries, the number of misspecified q-vectors in each Q-matrix was fixed at 10. The first Q-matrix had one misspecified q-entry in each q-vector. For each subsequent Q-matrix, an additional q-entry misspecification was included within one of the 10 misspecified q-vectors. In total, 21 Q-matrices with between 10 to 30 misspecifications were used for the study. Examinee responses were generated from the G-DINA models and examinees' estimated attribute profiles were generated based on the multivariate normal threshold model, with variances

and covariances set to 1.0 and 0.5 respectively. The R programming language (R Core Team, 2016) was used for data generation and model specification. Each condition consisted of 25 replications.

5.1.2.2 Results

The results for Study 2 are evaluated using the measures described in Study 1. The result for $Q_{mis.v}$ summarized in Table 5.3 shows that when K = 3, MRR reduced as the number of misspecifications increased. All misspecified q-entries were corrected for Q-matrices with up to 7 misspecifications, while Q-matrices with 8 to 10 misspecifications had 1 or 2 unrecovered q-entries. Similar results were noted for the Q-matrix with K = 5 however, only Q-matrices with up to 6 misspecifications had all misspecified q-entries recovered. Higher MRR, SEN, and SPE were obtained when K = 3. While N does not seem to influence MRR, SEN, or SPE, no clear association was noted between SPE and the number of misspecifications.

Table 5.3. Simulation Study 2: Mean Recovery Rates (MRR) for the G-DINA model (Q-matrix misspecified by q-vector, $Q_{mis.v}$)

# of Misspecified q-vectors	N	*J* = 30, *K* = 3			*J* = 30, *K* = 5		
		MRR	SEN	SPE	MRR	SEN	SPE
1	1000	0.88	1.00	0.89	0.87	1.00	0.86
	2000	0.88	1.00	0.89	0.87	1.00	0.86
	3000	0.88	1.00	0.89	0.86	1.00	0.85
2	1000	0.87	1.00	0.89	0.86	1.00	0.85
	2000	0.88	1.00	0.88	0.87	1.00	0.86
	3000	0.87	1.00	0.89	0.87	1.00	0.86
3	1000	0.88	1.00	0.89	0.86	1.00	0.85
	2000	0.87	1.00	0.89	0.87	1.00	0.86
	3000	0.87	1.00	0.89	0.87	1.00	0.86
4	1000	0.87	1.00	0.88	0.86	1.00	0.85
	2000	0.87	1.00	0.89	0.86	1.00	0.85
	3000	0.87	1.00	0.88	0.86	1.00	0.85
5	1000	0.86	1.00	0.88	0.85	0.80	0.84
	2000	0.87	1.00	0.88	0.86	1.00	0.85
	3000	0.87	1.00	0.88	0.86	1.00	0.85
6	1000	0.88	1.00	0.88	0.85	1.00	0.83
	2000	0.87	1.00	0.88	0.85	0.83	0.83
	3000	0.85	1.00	0.88	0.85	1.00	0.83
7	1000	0.87	1.00	0.88	0.84	0.86	0.84
	2000	0.86	1.00	0.86	0.84	0.86	0.84
	3000	0.86	1.00	0.86	0.83	0.86	0.83
8	1000	0.86	0.88	0.86	0.83	0.88	0.82
	2000	0.87	0.88	0.85	0.83	0.88	0.82
	3000	0.86	0.88	0.85	0.83	0.88	0.82
9	1000	0.86	0.89	0.85	0.83	0.89	0.82
	2000	0.85	0.89	0.85	0.83	0.89	0.82
	3000	0.85	0.89	0.84	0.83	0.89	0.82
10	1000	0.86	0.90	0.85	0.83	0.80	0.84
	2000	0.85	0.90	0.85	0.83	0.80	0.83
	3000	0.86	0.90	0.85	0.84	0.80	0.84

The results in Table 5.4 show the effect of q-entry misspecifications on the recovery of the Q-matrix. When K = 3, MRR decreased as the number of misspecifications increased, while the number of unrecovered q-entries increased as the number of misspecifications increased. All misspecified q-entries were corrected for Q-matrices with up to 12 misspecifications. In addition, sample size does not seem to affect MRR, SEN, or SPE. Likewise, when K = 5, MRR and SEN decreased as the number of misspecifications increased. However, there were more unrecovered q-entries when K = 5. SPE was not affected by N or the number of misspecifications. Based on these results, q-entry and q-vector misspecification showed similar effects on MRR, SEN, and SPE.

Table 5.4. Simulation Study 2: Mean Recovery Rates (MRR) for the G-DINA model (Q-matrix misspecified by q-entry, $Q_{mis.e}$)

Number of misspecified q-entries	$J = 30, K = 3$				$J = 30, K = 5$		
	N	MRR	SEN	SPE	MRR	SEN	SPE
10	1000	0.87	0.90	0.86	0.87	0.90	0.87
	2000	0.87	0.90	0.86	0.87	0.90	0.87
	3000	0.90	1.00	0.89	0.86	0.90	0.86
12	3000	0.88	1.00	0.86	0.87	0.92	0.87
	2000	0.87	0.92	0.86	0.85	0.92	0.84
	3000	0.88	0.92	0.87	0.86	0.83	0.86
14	1000	0.87	0.93	0.85	0.86	0.79	0.87
	2000	0.87	0.86	0.87	0.86	0.79	0.87
	3000	0.88	0.86	0.88	0.85	0.86	0.85
16	1000	0.87	0.88	0.86	0.85	0.81	0.85
	2000	0.87	0.88	0.87	0.85	0.88	0.84
	3000	0.87	0.88	0.87	0.85	0.88	0.84
18	1000	0.86	0.89	0.85	0.85	0.78	0.86
	2000	0.86	0.83	0.86	0.83	0.83	0.83
	3000	0.85	0.83	0.85	0.83	0.78	0.84
20	1000	0.85	0.85	0.85	0.84	0.80	0.85
	2000	0.85	0.85	0.85	0.83	0.85	0.83
	3000	0.86	0.80	0.88	0.81	0.85	0.81
22	1000	0.85	0.86	0.85	0.83	0.77	0.84
	2000	0.84	0.82	0.85	0.83	0.77	0.84
	3000	0.85	0.82	0.86	0.83	0.82	0.83
24	1000	0.84	0.83	0.84	0.83	0.79	0.83
	2000	0.84	0.83	0.85	0.83	0.83	0.82
	3000	0.84	0.83	0.85	0.82	0.79	0.82
26	1000	0.83	0.81	0.84	0.82	0.81	0.82
	2000	0.83	0.85	0.82	0.82	0.77	0.83
	3000	0.83	0.81	0.84	0.82	0.77	0.83
28	1000	0.82	0.82	0.82	0.82	0.79	0.83
	2000	0.83	0.79	0.85	0.81	0.79	0.81
	2000	0.83	0.79	0.85	0.82	0.79	0.83
30	1000	0.84	0.80	0.86	0.81	0.80	0.81
	2000	0.83	0.80	0.85	0.80	0.77	0.81
	3000	0.84	0.80	0.86	0.81	0.80	0.81

5.2 Simulation studies for the Two-step Q-Matrix Estimation (TSQE) Method

The performance of the TSQE method is evaluated with two simulation studies using the DINA model and RRUM. In addition to determining the effectiveness of the method, the impact of the percentage of misspecification on the estimation of the Q-matrix will be investigated in Study 3, while Study 4 will assess the performance of the TSQE method and the IQV method.

5.2.1 Simulation Study 3

The goal of Study 3 is to determine the effectiveness of the method in estimating a Q-matrix with the DINA model and RRUM.

5.2.1.1 Simulation Study Design

Study 3 was conducted with four correct Q-matrices with *J, K,* and *N* the same as the Q-matrices used for Study 1. Data generation and model specification was also carried out using R programming language (R Core Team, 2016). Examinees' responses were generated for the DINA model and RRUM, with the slipping (s_j), and guessing (g_j), parameters generated from the uniform distribution. Examinees' attribute profiles were generated based on the multivariate normal threshold model, with variances and covariances set to 1.0 and 0.5 respectively. To determine the effect of the threshold value (λ) on MRR, λ values of 0.7, 0.8, and 0.9 were considered for the studies. The effect of item quality on MRR was investigated by using slipping (s_j) and guessing (g_j) parameter estimates of 0.1, 0.2, and 0.3. For the RRUM, the baseline parameter (π_i^*) and the penalty and the penalty parameter (r_{ik}^*) were set to 0.9 and 0.6 respectively. The effect of sample

size was investigated for N = 1000, 3000, and 5000. Table 5.5 shows the two Q-matrices with 20 items. When K = 3, there are 9 one-attribute items, 9 two-attribute items, and 2 three-attribute items. When K= 5, the Q-matrix consists of 5 each for one-attribute and three attributes items, and 10 items with two attributes. Each condition included 25 replications.

Table 5.5. Q-matrices for Simulation Study 3

Attributes (K=3)				Attributes (K=5)					
Item	K1	K2	K3	Item	K1	K2	K3	K4	K5
1	1	0	0	1	1	0	0	0	0
2	0	1	0	2	0	1	0	0	0
3	0	0	1	3	0	0	1	0	0
4	1	1	0	4	0	0	0	1	0
5	0	1	1	5	0	0	0	0	1
6	1	0	1	6	1	1	0	0	0
7	1	0	0	7	1	0	1	0	0
9	0	0	1	9	1	0	0	0	1
10	1	1	0	10	0	1	1	0	0
11	0	1	1	11	0	1	0	1	0
12	1	0	1	12	0	1	0	0	1
13	1	1	1	13	0	0	1	1	0
14	1	0	0	14	0	0	1	0	1
15	0	1	0	15	0	0	0	1	1
16	0	0	1	16	1	1	1	0	0
17	1	1	0	17	1	1	0	1	0
18	0	1	1	18	1	1	0	0	1
19	1	0	1	19	0	1	1	1	0
20	1	1	1	20	0	1	1	0	1

5.2.1.2 Results

Tables 5.6 and 5.7 summarize the results from study 3. The results show that the TSQE method is efficient in estimating a Q-matrix, yielding MRR estimates as high as 100% by q-entry, and for both CDM model. The results in Table 5.6 also show that for the

DINA model, MRR was the highest when s_j and g_j are fixed at 0.1 for all conditions, which is as expected since accuracy rate should improve as item quality increases (i.e., low slipping and guessing parameters). There is a drop in MRR as *K* increases and an increase in MRR as *J* increases. The results show that MRR is higher when $\lambda = 0.7$ or 0.8, with little difference between MRR for both λ values, especially when item quality is good. Sample size did not seem to have any impact on MRR as similar results were obtained across the different values of *N*.

For data generated based on RRUM shown in Table 5.7, the results show similar MRR for all values of λ. MRR increases as *J* increases but decreases as *K* increases. However, there seems to be no effect on MRR as sample size increases.

Table 5.6. Study 3: Mean Recovery Rate (MRR): DINA Model

			N = 1000			*N* = 3000			*N* = 5000		
J	*K*	*s*, *g*	λ = 0.7	λ = 0.8	λ = 0.9	λ = 0.7	λ = 0.8	λ = 0.9	λ = 0.7	λ = 0.8	λ = 0.9
20	3	0.1	0.91	0.91	0.84	0.90	0.90	0.87	0.91	0.92	0.87
		0.2	0.91	0.91	0.81	0.91	0.92	0.80	0.92	0.92	0.81
		0.3	0.88	0.86	0.83	0.89	0.89	0.81	0.88	0.89	0.82
	5	0.1	0.85	0.88	0.86	0.86	0.87	0.86	0.87	0.86	0.86
		0.2	0.84	0.85	0.85	0.85	0.85	0.86	0.84	0.88	0.85
		0.3	0.81	0.76	0.80	0.78	0.76	0.79	0.77	0.78	0.84
30	3	0.1	0.98	0.99	0.98	0.97	0.98	0.98	0.98	0.99	0.98
		0.2	0.97	0.99	0.96	0.97	0.98	0.96	0.97	0.98	0.95
		0.3	0.93	0.96	0.97	0.94	0.97	0.95	0.94	0.98	0.95
	5	0.1	0.90	0.90	0.87	0.88	0.89	0.87	0.90	0.89	0.87
		0.2	0.90	0.89	0.86	0.87	0.89	0.86	0.87	0.89	0.86
		0.3	0.82	0.84	0.83	0.86	0.85	0.86	0.80	0.85	0.86

Table 5.7. Study 3: Mean Recovery Rate (MRR): RRUM Model

		N = 1000			*N* = 3000			*N* = 5000		
J	*K*	λ = 0.7	λ = 0.8	λ = 0.9	λ = 0.7	λ = 0.8	λ = 0.9	λ = 0.7	λ = 0.8	λ = 0.9
20	*3*	0.87	0.87	0.86	0.87	0.88	0.87	0.86	0.87	0.86
	5	0.84	0.84	0.84	0.87	0.88	0.86	0.84	0.86	0.84
30	*3*	0.88	0.89	0.88	0.87	0.87	0.86	0.88	0.88	0.88
	5	0.86	0.86	0.86	0.84	0.85	0.84	0.86	0.86	0.86

5.2.2 Simulation Study 4

The goal of study 4 is to compare the effectiveness of the IQV and TSQE in validating and estimating Q-matrix respectively.

5.2.2.1 Simulation Study Design

Study 4 was conducted with a provisional Q-matrix with J = 30, and K = *5*. Item responses conforming to the RRUM were generated based on the provisional Q-matrix for N = 1000, 3000, and 5000 examinees. The baseline parameter (π_i^*) and the penalty parameter (r_{ik}^*) were set to 0.9 and 0.6 respectively. To evaluate the performance of the methods with different CDMs, the data were fitted with the DINA and G-DINA models. To compare the effect of Q-matrix misspecification on the performance of the proposed methods, two Q-matrices with 10% and 20% misspecifications were included in the study. Since study 3 showed that MRR is best when λ = 0.8, this threshold value was used for the TSQE method.

5.2.2.2 Results

Table 5.8 summarizes the MRR for Q-mod IQV and Q-mod TSQE, and the proportion of identical q-entries between the two modified Q-matrices. The results for the DINA model from analysis using the IQV method shows that all misspecified q-entries were recovered for all conditions, while MRR decreased as the percentage of misspecification increased for analysis with the TSQE method. For the G-DINA model, there is also a decrease in MRR as the percentage of misspecification increased, while MRR for Q-mod IQV is generally higher than for Q-mod TSQE. A comparison of the MRR by model shows that higher MRR were obtained for the DINA model and for both proposed methods. A review of the proportion of identical q-entries between the two modified Q-matrices indicates a higher number of matches for analysis with the DINA model. The proportion of identical q-entries between the two modified Q-matrices does not seem to be influenced by *N* or by the percentage of misspecification.

Table 5.8. Study 4: Comparison of modified Q-matrices obtained from the IQV and TSQE methods

		DINA Model			G-DINA Model		
		Q-mod IQV	Q-mod TSQE	IQV vs TSQE	Q-mod IQV	Q-mod TSQE	IQV vs TSQE
Percentage of Misspecification	N	MRR	MRR	Proportion of identical q-entries	MRR	MRR	Proportion of identical q-entries
	1000	1.00	0.85	0.85	0.87	0.83	0.84
	3000	1.00	0.85	0.85	0.87	0.84	0.85
10%	5000	1.00	0.83	0.83	0.85	0.82	0.83
	1000	1.00	0.81	0.81	0.84	0.79	0.77
	2000	1.00	0.80	0.80	0.83	0.76	0.82
20%	3000	1.00	0.80	0.80	0.84	0.78	0.79

Chapter 6. Analysis of the TIMSS Dataset: Part II

In this chapter, the provisional Q-matrix used for the analysis in Chapter 3 will be validated using the IQV method and the Q-matrix will be estimated from scratch using the TSQE method developed in the dissertation. The subset of the TIMSS data will then be re-analyzed with the G-DINA model using the Q-matrices modified by the two proposed methods. In addition to verifying the effectiveness of both methods, the modified Q-matrices will be evaluated for appropriateness from a content perspective, and the updated estimated attribute profiles will be reviewed to determine if the inconsistencies in classification identified during the first part of the analysis have improved.

6.1 Procedure for the TIMSS Data Analysis

To correct any possible misspecifications in the provisional Q-matrix (Q-prov) obtained through retrofitting, two separate analyses were carried out for each group using the IQV and TSQE methods. In addition to the group analysis, an analysis combining all data was completed using both proposed methods. The purpose of including the combined analysis is to determine if there are any significant differences in the modified Q-matrices (Q-mod) obtained from the group analysis and from the combined analysis. Each Q-mod is evaluated from a content perspective and compared with Q-prov and the other Q-mods. For each method, the modified Q-matrix that is most interpretable will be used for the rest of the analysis. The data is then fitted with the G-DINA model using the selected modified Q-matrices to determine examinees' updated estimated attribute profiles. Finally, the estimated attribute profiles are examined to establish if the inconsistencies in classification that were previously identified have improved.

6.2 Results

The results for the analysis include a review of the factor loadings and provisional Q-matrix obtained from the first step of the TSQE method, a comparison of the proportion of identical q-entries in the provisional and modified Q-matrices, SEN and SPE rates by method, an evaluation of the modified Q-matrices from a content perspective, and an evaluation of examinees' estimated attribute profiles.

6.2.1 Review of the Provisional Q-matrix obtained from the factor analysis step of the TSQE method

As outlined in the methods section (page 40), in the first part of the TSQE procedure, FA is applied to the correlation matrix to obtain the provisional Q-matrix (Q-prov FA). In Q-prov FA, factors that contribute to the cumulative sum of the loading equal to or greater than the threshold value, $\lambda = 0.8$, are specified as the attributes required to answer each item correctly. Table 6.1 shows the loadings for each factor after rotation and

Table 6.1. Review of the Provisional Q-matrix obtained from analysis with the TSQE method

	Factor Loadings						Cumulative loadings	Q-prov FA					
J								*K*					
	1	2	3	4	5	6		*K1*	*2*	*K3*	*K4*	*K5*	*K6*
1	0.06	0.19	0.14	0.05	0.13	0.53	0.86	0	1	1	0	0	1
2	0.33	0.66	0.07	0.25	-0.02	0.12	0.98	1	1	0	0	0	0
3	0.14	0.97	0.04	0.01	0.16	0.06	0.97	0	1	0	0	0	0
4	0.16	0.09	0.11	0.11	0.31	0.95	0.95	0	0	0	0	0	1
5	0.63	0.23	0.08	0.07	0.18	0.21	0.86	1	1	0	0	0	0
6	0.14	0.19	0.10	0.96	0.08	0.02	0.96	0	0	0	1	0	0
7	0.25	0.25	0.21	0.46	0.02	0.29	0.81	1	1	0	0	1	1
8	0.13	0.12	0.10	0.12	0.74	0.03	0.87	1	0	0	0	1	0
9	0.31	0.29	0.16	0.18	0.50	0.04	0.81	1	0	0	0	1	0
10	0.27	0.12	0.20	0.04	0.58	0.17	0.85	1	0	0	0	1	0
11	0.06	0.16	0.46	0.22	0.36	-0.05	0.82	0	0	1	0	1	0
12	0.30	0.06	0.36	0.09	0.31	-0.09	0.98	1	0	1	0	1	0
13	0.30	0.30	0.89	0.09	0.05	0.06	0.89	0	0	1	0	0	0
14	0.02	0.70	0.10	0.14	0.02	0.32	1.02	0	1	0	0	0	1

15	0.36	0.18	0.15	0.34	0.25	0.10	0.95	1	0	0	1	1	0

how they contribute to the cumulative loadings to determine the attributes required for each item. Q-prov FA is then used as input into the Q-matrix refinement method to obtain Q-mod TSQE shown in Table 6.3.

6.2.2 Mean Recovery Rates by q-entries, Sensitivity Rates, and Specificity Rates

Table 6.2 summarizes the proportion of q-entries in Q-mod IQV and Q-mod TSQE that are identical to q-entries in Q-prov, and the SEN and SPE rates for the modified Q-matrices. For each group, similar MRRs were observed in both modified Q-matrices. Higher SEN was observed for Q-mod IQV, and SPE was similar for both modified Q-matrices. While Q-mod TSQE had the lowest SEN for two groups, SPE was lowest in two groups for Q-mod IQV.

Table 6.2. Comparison of the Modified Q-matrices with the Provisional Q-matrix

	Q-mod IQV			Q-mod TSQE		
Groups	MRR by q-entries	SEN	SPE	MRR by q-entries	SEN	SPE
USA	0.90	0.78	0.91	0.90	0.67	0.92
High-performing	0.92	0.78	0.94	0.90	0.67	0.93
Mid-performing	0.91	0.89	0.91	0.93	0.89	0.94
Low-performing	0.88	0.78	0.89	0.88	0.78	0.89
Combined data	0.92	0.78	0.94	0.92	0.78	0.93

6.2.3 Evaluation of the Modified Q-matrices

An evaluation of the Q-matrices obtained by the IQV method showed differences across the four groups. For example, item 10 asks examinees to identify the missing number in a given pattern of numbers. Although the modified Q-matrices by groups are

not included in the results shown, for *J*10, Q-mod IQV for the mid-performing group indicates *K*2 (recognizing multiples and factors) as the required attribute, while Q-mod IQV for the low-performing group indicates *K*6 (describing relationships in patterns) as the required attribute. However, Q-prov indicates that both *K*2 and *K*6 are required to answer the item correctly. Generally, Q-mod TSQE showed more similarities across the groups than Q-mod IQV. For example, Q-mod TSQE for all groups specifies *K*6 as the required attribute for *J*14 while Q-mod IQV shows *K*2 as the required attribute for the high performing group, *K*2 and *K*6 for the low-performing group and USA, and *K*1, *K*2, and *K*6 for the mid-performing group.

In addition, some of the entries in the modified Q-matrices from the group analysis are not interpretable. For example, the Q-mod IQV for the low-performing group shows *K*1, *K*2, *K*3, and *K*4 as the required attribute for *J*3 (identify factors of 12). While skills noted in *K*1 (computing with whole numbers) and *K*2 (recognizing multiples and factors) may be used to answer the item, *K*3 (find the missing operation) and *K*4 (solve real-life problems) are both irrelevant in this situation. Also, Q-mod for USA indicates that *K*5 (computing with fractions) is required for *J*10 (find the next number in a pattern with 3,6,9,12). However, knowledge of computing with fractions is not relevant for identifying a given pattern. For both methods, the Q-mods for the individual analysis seem to have more errors than the Q-mods for the combined data.

Since the two Q-mods obtained using the combined data show more consistent and interpretable results compared to the Q-mods obtained from the group analysis, results based on analysis with these two Q-matrices will be used for the rest of this study. Table 6.3 provides a side-by-side comparison of the provisional Q-matrix and the two modified

Q-matrices. The attributes required to answer 5 items changed in Q-mod IQV. Of the 3 items (*J*7, *J*10, *J*14) in Q-prov that require both *K*2 and *K*6, only *J*14 requires both attributes in Q-mod IQV, while *J*7 and *J*10 both require only *K*6. Likewise, the required attributes

Table 6.3. Changes in the Q-matrix based on the analysis for the combined data

	Q-prov						Q-mod IQV						Q-mod TSQE					
J	*K1*	*K2*	*K3*	*K4*	*K5*	*K 6*	*K1*	*K2*	*K3*	*K4*	*K5*	*K 6*	*K1*	*K2*	*K3*	*K4*	*K5*	*K6*
1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
2	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
3	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
5	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0
6	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
7	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
8	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0
9	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0
10	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
11	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
12	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
13	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
14	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1
15	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0

for 6 items changed in Q-mod TSQE, with all 6 items, including the 3 items in Q-prov that require both *K*2 and *K*6, now requiring only *K*6. The changes in the required attribute for three items (*J4*, *J*7, *J*10) were the same in both modified Q-matrices, while 2 items (*J*1, *J11)*, have different required attributes specified in both modified Q-matrices.

Further investigation by a content analysis showed that while in some instances (*J*3, *J*7, *J*10, *J*14), the changes in the modified Q-matrices are meaningful due to reasons such as the possibility of using multiple strategies or methods for solving a problem, in other instances, such as for *J*1, the change noted in Q-mod IQV does not seem meaningful from a content perspective. This is because *J*1 asks examinees to identify the correct number sentence and so requires *K*3 (find the missing number or missing operation) as specified in

Q-prov. However, as shown in Table 6.3, Q-mod IQV indicates that *K*4 (solve real-life problems involving measurements, money, and time) is required, while Q-mod TSQE indicates that *K*6 (relationships and patterns) is required. Since *K*3 is the most relevant for *J*1, *K*3 will be assigned in place of *K*4 to *J*1 in Q-mod IQV.

6.2.4 Review of Examinee Attribute Profiles

In this section, the estimated attribute profiles obtained from the preliminary analysis will be compared by attribute with the estimated attribute profiles obtained from the analyses using the modified Q-matrices. In addition, estimated attribute profiles obtained from the analyses with the modified Q-matrices will be evaluated to determine the extent to which the inconsistencies identified during the preliminary analysis might have been resolved.

A comparison of the estimated attribute profiles from the analysis using the modified Q-matrices shows that 3454 of the 4102 examinees have identical estimated attribute profiles, while 2172 examinees from the analysis using the provisional Q-matrix and analysis based on the modified Q-matrices have identical estimated attribute profiles. Table 6.4 provides a summary of the proportions of identical classifications by attribute based on analysis with Q-prov and both Q-mods. A review of the proportion of examinees with identical classifications by attribute based on analyses with the modified Q-matrices shows that between 96% and 100% of examinees had identical classifications by attribute. This indicates that very similar results were obtained for analyses based on both methods. A comparison of examinee attribute patterns from the preliminary analysis and analysis using the modified Q-matrices shows similar results for all attributes expect *K*4. This

difference is most likely because mastery of *K*4 is required for two additional items in Q-mod IQV method. The proportion of q-entries that changed from 0 to 1 were similar for both

Table 6.4. Review of examinee attribute profiles

	Q-mod IQV vs. TSQE	Q-mod IQV vs. Q-prov			Q-mod TSQE vs. Q-prov		
	Proportion identical	Proportion identical	Proportion of q-entries changed from 0 to 1	Proportion of q-entries changed from 1 to 0	Proportion identical	Proportion of q-entries changed from 0 to 1	Proportion of q-entries changed from 1 to 0
*K*1	0.98	0.98	0.01	0.01	0.98	0.02	0.01
*K*2	0.99	0.78	0.01	0.20	0.78	0.02	0.20
*K*3	1.00	0.84	0.03	0.13	0.84	0.03	0.13
*K*4	0.99	0.90	0.08	0.01	0.99	0.00	0.01
*K*5	0.98	0.98	0.01	0.01	0.98	0.01	0.01
*K*6	0.96	0.95	0.02	0.03	0.94	0.03	0.03

proposed methods for all attributes except *K*4, which had 0.08 proportion changed for analysis based on Q-mod IQV and 0.00 for analysis based on Q-mod TSQE method. *K*2 and *K*3 had the highest proportion changed from 1 to 0 for analyses based on both modified Q-matrices. This high changes in proportion is expected because compared to the provisional Q-matrix, fewer items in the modified Q-matrices require mastery of *K*2 and *K*3.

Table 6.5 shows the item response profiles, estimated attribute profiles, and proportion correct by attribute for the sample of examinees from the preliminary analysis. The results indicate that the inconsistencies noted in the preliminary analysis have improved and examinees are more appropriately classified. For example, results from the preliminary analysis showed discrepancies in the classification for examinee 1065 and 3130. However, their estimated attribute profiles based on the modified Q-matrices are

more consistent with the similarities in their proportion correct by attribute estimates. In the estimated attribute profiles from both modified Q-matrices, the only difference in their

Table 6.5. Comparison of estimated attribute profiles and attribute prevalence estimates obtained from the analysis with Q-prov, Q-mod IQV, and Q-mod TSQE

		Q-prov		Q-mod IQV		Q-mod TSQE	
ID	Item Response Profile	$\hat{\alpha}$	Proportion correct by attribute	$\hat{\alpha}$	Proportion correct by attribute	$\hat{\alpha}$	Proportion correct by attribute
1065	11010011	111000	1.00 0.50 1.00	101011	1.00 0.00 1.00	101001	1.00 0.00 1.00
	1111100		0.00 0.67 0.60		0.67 0.67 0.60		0.00 0.66 0.75
3130	11010011	111111	1.00 0.50 1.00	101111	1.00 0.00 1.00	101101	1.00 0.00 1.00
	1111101		1.00 0.67 0.60		1.00 0.67 0.60		1.00 0.66 0.75
1500	11010101	100101	1.00 0.50 0.67	100111	1.00 0.50 0.00	100111	1.00 0.00 0.00
	1110111		1.00 0.67 0.80		1.00 0.67 0.80		1.00 0.33 0.88
2111	11010101	100101	1.00 0.50 0.67	100111	1.00 0.50 0.00	100111	1.00 0.00 0.00
	1110111		1.00 0.67 0.80		1.00 0.67 0.80		1.00 0.33 0.88
1060	11010101	111111	1.00 0.50 1.00	101111	1.00 0.50 1.00	101111	1.00 0.00 1.00
	1111111		1.00 0.67 0.80		1.00 0.67 0.80		1.00 0.33 0.88
163	11011000	000000	1.00 0.25 1.00	101100	1.00 0.00 1.00	101010	1.00 0.00 1.00
	0111100		0.00 0.33 0.40		0.67 0.33 0.40		0.00 0.67 0.63
226	11011000	111101	1.00 0.50 1.00	101101	1.00 0.50 1.00	101111	1.00 0.00 1.00
	0111111		1.00 0.33 0.60		1.00 0.33 0.60		1.00 0.67 0.75
4028	11011100	111100	1.00 0.25 1.00	101100	1.00 0.00 1.00	101100	1.00 0.00 1.00
	0111001		1.00 0.33 0.40		1.00 0.33 0.40		1.00 0.33 0.63
1235	11011100	101101	1.00 0.25 1.00	101111	1.00 0.00 1.00	101111	1.00 0.00 1.00
	0111101		1.00 0.33 0.60		1.00 0.33 0.60		1.00 0.33 0.75
3439	11011010	000001	1.00 0.75 0.67	100101	1.00 0.50 0.00	100001	1.00 0.00 0.00
	0110110		0.00 0.33 0.80		0.67 0.33 0.80		0.00 0.33 0.88
1092	11011010	111111	1.00 0.50 0.67	100111	1.00 0.00 0.00	100111	1.00 0.00 0.00
	1110101		1.00 0.67 0.60		1.00 0.67 0.60		1.00 0.67 0.75
3155	11011001	111110	1.00 0.25 0.67	100110	1.00 0.00 0.00	100111	1.00 0.00 0.00
	1110101		1.00 1.00 0.40		1.00 1.00 0.40		1.00 1.00 0.63
1931	11011001	100010	1.00 0.50 0.67	100110	1.00 0.50 0.00	100011	1.00 0.00 0.00
	1110110		0.00 1.00 0.60		0.67 1.00 0.60		0.00 0.67 0.75
3136	11011001	101010	1.00 0.25 1.00	101010	1.00 0.00 1.00	101010	1.00 0.00 1.00
	0111100		0.00 0.67 0.40		0.67 0.67 0.40		0.00 1.00 0.63
4075	11011001	111110	1.00 0.25 1.00	101110	1.00 0.00 1.00	101110	1.00 0.00 1.00
	0111101		1.00 0.67 0.40		1.00 0.67 0.40		1.00 0.67 0.63
3252	11011111	100011	1.00 0.50 0.67	100111	1.00 0.00 0.00	100011	1.00 0.00 0.00
	1110100		0.00 1.00 0.80		0.67 1.00 0.80		0.00 0.67 0.88

Attribute Prevalence Estimates

	*K*1	*K*2	*K*3	*K*4	*K*5	*K*6
Q-prov	0.68	0.58	0.61	0.62	0.62	0.73
Q-mod IQV	0.69	0.39	0.50	0.69	0.62	0.71
Q-mod TSQE	0.69	0.39	0.50	0.61	0.62	0.73

classification is that while examinee 1065 is classified as nonmastery for *K*4, examinee 3130 is classified as mastery. Likewise, the updated estimated profiles for examinee 2111 and 1060 are more consistent, with the similarities in their proportion correct by attribute estimates. The only difference in their classification being that examinee 1060 is shown to have mastered *K*3 in the estimated attribute profiles obtained using the modified Q-matrices. Again, the changes in the estimated attribute profile for examinee 163 are consistent with the similarities in the proportion correct by attribute estimates for examinee 163 and 226. The estimated attribute profile for examinee 163 from the preliminary analysis showed nonmastery of all six attributes. However, estimated attribute profiles from both modified Q-matrices now indicate that the examinee has indeed mastered some attributes.

Generally, examinees had similar estimated attribute profiles based on analysis with the modified Q-matrices. The differences in classification is accounted for by the differences in the specification of the Q-matrices. For example, examinee 3252 is classified as mastery for *K*4 based on Q-mod IQV method and nonmastery based on Q-mod TSQE for the same attribute. This is because as shown in Table 6.3, in Q-mod IQV, *K*4 is required to answer 3 items (*J*1, *J*11, *J*15) correctly, while in Q-mod TSQE, *K*4 is only required to answer *J*15. Therefore, an examinee who answers *J*1 and *J*11 correctly, and *J*15 incorrectly, will have a proportion correct by attribute estimate of 0.67 for *K*4 based on analysis with the Q-mod IQV, and 0.0 based on analysis with Q-mod TSQE.

The results from the above data analysis indicates that the IQV and TSQE methods are effective for validating and estimating a Q-matrix respectively. Although there are still some inconsistencies in examinee classification, examinee estimated attribute profiles

obtained from analysis using the modified Q-matrices show more clearly defined classifications compared to the estimated attribute profiles based on the provisional Q-matrix.

Chapter 7. Implications and Limitations

This study presents one method for validating a Q-matrix, one method for estimating a Q-matrix, and a detailed analysis of the TIMSS 2011 fourth-grade dataset. The Integrated Q-matrix Validation method (IQV) combines a technique that uses Joint maximum likelihood estimation (JMLE) procedure (Chiu et al, 2016) to estimate examinees' attribute profiles with the nonparametric Q-matrix refinement method (Chiu 2013) to obtain the corrected Q-matrix. In the two step Q-matrix estimation method (TSQE), the matrix of tetrachoric correlations is used as input for factor analysis to obtain an initial Q-matrix. As a second step, this initial Q-matrix is introduced into the Q-matrix refinement method to obtain an updated Q-matrix.

The performance of the methods was examined with four simulation studies and a comprehensive analysis of a subset of data from TIMSS. The first two simulation studies evaluated the effectiveness of the IQV method to recover a Q-matrix using data based on the DINA and G-DINA models. In study 1, the effect of misspecifications on the recovery of the Q-matrix was evaluated by using Q-matrices with 10% and 20% misspecifications. The result showed that the IQV method can correct up to 100% of the misspecifications in a Q-matrix and retain up to 100% of the correct q-entries. Mean recovery rates (MRR) obtained for analysis with the DINA model were higher than for the G-DINA model. While test length, number of attributes, and percentage of misspecification influenced MRR, sample size seemed not to have any effect on MRR. Study 2 evaluated the effect of q-entry and q-vector misspecifications on the recovery of a provisional Q-matrix. The results showed that MRR and SEN decreased as the number of misspecified q-vectors and q-entries increased, while SPE did not seem to be influenced by the number of

misspecifications. In addition, higher measures were obtained when *K* is small. All three measures were not affected by sample sizes. In study 3, the performance of the TSQE method was evaluated with the DINA model and RRUM. The results of the study showed MRR as high as 100% for analyses with both CDMs. As noted in studies 1 and 2, MRR was affected by *J* and *K*, while sample size showed no effect on MRR. In addition, similar MRRs were obtained for $\lambda = 0.7$ and 0.8. In study 4, the performance of the proposed methods was compared by using both methods to analyze data generated for RRUM and fitted with the DINA and G-DINA models. Higher MRR rates were obtained for analysis using the DINA model, while MRR for Q-mod IQV is generally higher than the MRR for Q-mod TSQE for both models.

The proposed methods were used for a detailed of analysis of the TIMSS data with a provisional Q-matrix developed by content experts through a process of retrofitting. A preliminary analysis of the data with the G-DINA model indicated the presence of inconsistencies in examinees' estimated attribute profiles. To resolve this issue, the data was first analyzed using the IQV method to validate the provisional Q-matrix and the TSQE method to estimate a Q-matrix without using the provisional Q-matrix. The modified Q-matrices obtained from these analyses were then used to fit the data with the G-DINA model. An evaluation of the updated estimated attribute profiles showed that the inconsistencies observed in examinees' estimated attribute profiles had reduced.

7.1 Implications of the study

Educational institutions seek to apply data-informed decision-making processes to make pedagogical decisions and to promote continuous improvements of their students. CDMs have become popular due to their ability to provide detailed information about

examinees' strengths and weaknesses in form of an attribute profile. Since the information obtained from the attribute profiles can be used to tailor instruction to meet students' needs, CDMs can be useful tools for supporting formative assessments and improving achievement. Most of the currently available assessments are however not compatible for use with CDMs, and to make CDMs more accessible, an adhoc Q-matrix is often created through a process of retrofitting. Developing Q-matrix validation and estimation techniques, like the IQV and TSQE methods, that do not require complex estimation processes can make CDMs more accessible for use.

The IQV method unlike other Q-matrix validation methods does not require an arbitrary cut-off, which is a benefit for practitioners who do not have the technical know-how to determine a cut-off. The performance of the methods with other reduced models can be predicted from the results obtained in this study since both methods were analyzed with a general model. In addition, the Q-matrices obtained from analysis using the proposed methods will enhance the work of content experts who develop Q-matrices. The results obtained from the analysis of the TIMSS data indicates that the proposed methods may be used as a model for future analysis of the assessment with CDMs.

7.2 Limitations of the Study

Although the proposed methods showed potential in their performance, some limitations have been identified. The Q-matrices that were used for the simulation studies consists of up to 30 items and 5 attributes and included all possible attribute patterns that require between 1 and 3 attributes. In practice, an assessment may have more than 30 items and require examinees to show mastery of more than 5 attributes. In addition, not all

attribute patterns may be represented in a Q-matrix and the attribute patterns may not be equally distributed across the test. For all three CDMs used for the simulation studies, the results showed a decrease in MRR as the number of attributes in a Q-matrix increased from 3 to 5. This might be of concern in practice, especially for assessments that tests examinees' knowledge of more than 5 skills. Although the presented methods performed well with sample sizes between 1,000 and 5,000, it is not certain if this same kind of result will be achieved with sample sizes that are less than 1000 or greater than 5,000. In practice, school teachers have class sizes lower than 50 as such it is yet to be determined if these methods will be of any benefit under such conditions.

Some of the attributes specified in the Q-matrix developed for the TIMSS analysis are general since in some instances, multiple objectives were consolidated into one attribute. For example, as shown in Table 3.2 (page 30), 5 objectives were consolidated into $K5$ (understanding of fractions, fraction equivalent, and solving problems involving simple fractions). This implies that an examinee whose attribute profile indicates mastery for $K5$ may not necessarily be proficient in all the consolidated objectives, while an examinee that is classified as nonmastery for the attribute may actually be proficient in one or more of the five objectives. The results of the TIMSS data analysis showed that the estimated attribute profiles obtained from the analysis with the modified Q-matrices showed more consistency than the estimated attribute profiles obtained using the provisional Q-matrix. However, the persistent presence of inconsistencies indicates that other issues, such as the presence of a hierarchical structure between attributes, may be responsible for the discrepancies in examinee classification. An assumption of a hierarchical structure between attributes implies that an attribute may be a prerequisite for

another attribute. For example, in the provisional Q-matrix designed for the analysis with TIMSS, *K*2 (recognizing multiples and factors) and *K*6 (describing relationship in patterns, their extension, and generating numbers based on a given rule) mostly appeared together, which could imply that one of the attributes may be essential for learning the other.

7.3 Possible Future Direction

Additional investigation to resolve the inconsistencies in the estimated attribute profile will be helpful in ensuring that the provisional Q-matrix is appropriately specified, thus improving the recovery rate of the IQV method. The results from a scree test conducted during the initial analysis of the TIMSS data indicated that the Q-matrix should be developed with four attributes. However, based on consultations with content experts the provisional Q-matrix was developed with six attributes. Although it is currently not possible to establish a Q-matrix with four attributes due to limited resources, it would be interesting in future to develop a Q-matrix with four attributes and then compare the results of the analysis using the Q-matrix with the results from this study. Given the fact that the provisional Q-matrix was developed by consolidating objectives from the TIMSS blueprint, a future consideration would be to develop a Q-matrix that would not include any consolidated objecitves, thereby making the attributes more specific. The proposed methods can also be modified for use with polytomous attributes since assessments based on polytomous attributes provide additional diagnostic information that can inform instruction and improve student learning. Further studies on the effect of sample size on the recovery of the Q-matrix is also required to determine how well the methods will perform with sample sizes less than 1,000 and greater than 5,000. Studies will also be

designed to compare the proposed methods with existing Q-matrix validation and estimation methods.

8. References

Carlson, J.E., von Davier, M. (2013). Item Response Theory ETS R&D Scientific and Policy Contributions Series ETS SPC-13-05

Chen , Y., Culpepper, S., Chen, Y., Douglas, J. (2018) Bayesian Estimation of the DINA Q-matrix Psychometrika. 83(1) 89-108.

Chiu, C. (2013). Statistical Refinement of the Q-Matrix in Cognitive Diagnosis. Applied Psychological Measurement 37(8) 598-618.

Chiu, C. Y., & Douglass, J.A. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal item response patterns. Journal of Classification 30, 225-250

Chiu CY., Köhn HF., Zheng Y., Henson R. (2015) Exploring Joint Maximum Likelihood Estimation for Diagnostic Classification Models. Psychometrika 81(4), 1-24

Close, C. N., Davison, M.L., Davenport, E.C. (2012). An Exploratory Technique for Finding the Q-matrix in Cognitive Diagnostic Assessment: Combining Theory with Data

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical Assessment, Research & Evaluation, 10, 1–9.

Cui, Y., Gierl, M.J., Chang, H. (2012) Estimating Classification Consistency and Accuracy for Cognitive Diagnostic Assessment, Journal of Educational Measurement, 49, 1, (19 -38)

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.

Cognitive Diagnostic Assessments for Education: Theory and Practice Cambridge, New York: University Press

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. (2009). DINA Model and Parameter Estimation: A Didactic. Journal of Educational and Behavioral Statistics, 34(1), 115-130.

de la Torre & Chiu. (2016). A General Method of Empirical Q-matrix Validation. Psychometrika, 81, 253-273.

Gierl, M.J., Wang, C., & Zhou, J. (2008). Using the Attribute Hierarchy Method to Make Diagnostic Inferences about Examinees' Cognitive Skills in Algebra on the SAT©. Journal of Technology, Learning, and Assessment, 6(6). Retrieved [date] from http://www.jtla.org

Haberman, Shelby J. (2004). Joint and Conditional Maximum Likelihood Estimation for The Rasch Model for Binary Responses. ETS Research Report Series 1k

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. Journal of Educational Measurement, 26,301–321.

Henson, R., & Douglas J. Test construction for cognitive diagnosis. Applied Psychological Measurement, 29 (4), 262-277, 2005.

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.

Howard, C. (2016). A Review of Exploratory Factor Analysis Decisions and Overview of Current Practices: What We Are Doing and How Can We Improve? International Journal of Human–Computer Interaction, 32, 51–62.

Junker B.W., Sijtsma, K. (2001). Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory Applied Psychological Measurement; 25; 258-272.

Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. International Journal of Testing, 11(2), 144-177. DOI: 10.1080/15305058.2010.534571.

Park, Y.S., & Lee, Y.S. (2011). Diagnostic cluster analysis of mathematics skills. IERI Monograph Series: Issues and Methodologies in Large-Scale Assessment, 4, 75-107.

Liu, J., Xu, G., Ying, Z. (2012). Data-Driven Learning of Q-Matrix. Applied Psychological Measurement, 36(7), 548-564.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2015). CDM: Cognitive diagnosis modeling. R package version 3.1-14. Retrieved from the Comprehensive R Archive Network [CRAN] website

Rupp, A., Templin, J. (2008). The Effects of Q-Matrix Misspecification on Parameter Estimates and Classification Accuracy in the DINA Model. Educational and Psychological Measurement 68(1), 78-96.

Rupp, A., Templin,J. & Henson,R. (2010). Diagnostic Measurement: Theory, Methods, and Applications. The Guilford Press, New York, London.

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

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. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), Cognitively diagnostic assessment (pp. 327–359). Hillsdale NJ: Erlbaum.

Tatsuoka, K, Corter, J., Tatsuoka, C (2004). Patterns of Diagnosed Mathematical Content and Process Skills in TIMSS-R across a Sample of 20 Countries American Educational Research Journal 41 (4).,901-926.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0