Q-MATRIX OPTIMIZATION FOR COGNITIVE DIAGNOSTIC ASSESSMENT

BY

CONG CHEN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Associate Professor Jinming Zhang, Chair
Professor Carolyn Anderson
Professor Hua-Hua Chang
Associate Professor Steven Culpepper

ABSTRACT

Cognitive diagnostic assessment is a growing area in psychological and educational measurement with the advantage of providing diagnostic profiles (mastery/non-mastery of measured attributes) for examinees, giving insights for classroom teaching and student learning. Central to the successful implementation of a cognitive diagnostic assessment is the Q-matrix, the structure that specifies item-attribute relationships. However, the Q-matrix is prone to misspecification, given that it is often constructed based solely on human opinions. This thesis uses three research studies to investigate key issues of Q-matrix optimization for cognitive diagnostic assessments.

The first study investigates the effects of Q-matrix misspecification on the classification accuracy and consistency of diagnostic results. The two types of Q-matrix misspecifications examined are Q-entry misspecification (which includes three levels of misspecification: 10%, 20% and 30%), and attribute misspecification (which includes attribute exclusion and attribute inclusion). The results of a simulation study show that both Q-entry and attribute misspecification significantly deteriorate the accuracy of classification and the consistency of diagnostic results. In addition, the two classification accuracy and consistency indices have the potential to be useful in identifying possible attribute misspecification (e.g., attribute inclusion) of Q-matrix in empirical analyses. The second study provides a systematic performance evaluation of the three most commonly used Q-matrix validation methods: the sequential EM based δ-method (de la Torre, 2008), the Bayesian estimation method (DeCarlo, 2012), and the nonparametric Q-matrix refinement method (Chiu, 2013), with both basic and complex assessment design factors. The results of two simulation studies reveal that the Bayesian estimation method outperforms the other two methods in terms of recovering the misspecified Q-entries across various conditions. The performance of the three Q-matrix validation methods is also affected to different degrees by various assessment design factors, among which the data generation model is the most critical. The third study proposes a two-stage

cross-validation method that combines the strengths of the nonparametric refinement method and Bayesian estimation techniques for improving Q-matrix validation accuracy and computation efficiency. The results show that the proposed method can effectively optimize Q-matrices that are possibly misspecified in both simulation and empirical data settings.

ACKNOWLEDGEMENTS

I would like to express my special appreciation and thanks to my advisor, Professor Jinming Zhang, who has been a tremendous mentor for me. I would like to thank him for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His advice for my research as well as for my career have been priceless. I cannot imagine a better advisor and mentor for my Ph.D. study.

I would also like to thank the rest of my committee members, Professor Carolyn Anderson, Professor Hua-Hua Chang, and Professor Steven Culpepper for their insightful advice and help on my dissertation. They have provided great teaching and strong support throughout my Ph.D. years.

Finally, my deep and sincere gratitude to my family for their continuous and unparalleled love, help and support: my dear parents who raised me with love and supported me in all my pursuits; my beloved husband Rui Li and adorable son Eason who have been a constant source of support and encouragement during the challenges of graduate school and life.

This work is dedicated to them.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Cognitive diagnostic assessments (CDAs) have gained a lot of attention in psychological and educational measurement over the past decades, particularly for their advantage in offering examinees fine-grained diagnostic information about their mastery of measured attributes rather than a simple overall proficiency test score as traditional measurement methods do (Rupp & Templin, 2008a; Rupp, Templin, & Henson, 2010). With such diagnostic information, teachers could take appropriate actions to remedy individual students' specific learning weaknesses (Huebner, 2010).

To implement CDAs, statistical techniques, often referred to as cognitive diagnostic models (CDMs) or diagnostic classification models (DCMs) in the measurement literature, have been devised (e.g., de la Torre & Douglas, 2004; Embretson, 1984; Hartz, 2002; Junker & Sijtsma, 2001; Templin & Henson, 2006a; von Davier, 2005). CDMs are psychometric models that aim to classify examinees according to their level of mastery of specified latent characteristics or attributes, which can be classified as either conjunctive or disjunctive, or similarly, compensatory or non-compensatory. Normally, conjunctive is interchangeable with non-compensatory, and disjunctive is interchangeable with compensatory. The commonly used conjunctive (or non-compensatory) CDMs are the deterministic-input, noisy-and-gate (DINA) model, the noisy inputs, deterministic-and-gate (NIDA) model (Junker & Sijtsma, 2001), and the re-parameterized unified model (RUM, or the fusion model, Hartz, 2002, Roussos et al., 2007). The well-known disjunctive (or compensatory) CDMs are the deterministic input, noisy-or-gate (DINO) model (Templin & Henson, 2006), the noisy inputs, deterministic-or-gate (NIDO) model (Templin & Henson, 2006b), and the compensatory RUM (Templin, 2006). Many CDMs can be considered as special cases of some more generalized models, such as the generalized DINA model (G-DINA, de la Torre, 2011),

the general diagnostic model (GDM, von Davier, 2005), and the log-linear cognitive diagnostic model (LCDM, Henson, Templin, & Willse, 2009).

Implementation of CDAs normally consists of four major steps: (a) an attribute definition step in which a set of measured attributes are defined by subject-matter experts; (b) a Q-matrix (Tatsuoka, 1990) construction step in which required attributes for answering each item correctly are specified for individual items; (c) a data analysis step in which a chosen CDM is applied to obtain the individual examinee's mastery estimation of each required attribute; and (d) a score reporting step in which diagnostic information about examinees' mastery of required attributes is reported to examinees and teachers (Lee & Sawaki, 2009; Sawaki, Kim, & Gentile, 2009). Among these four steps, Q-matrix construction is the most important step because the Q-matrix, which identifies the latent attributes measured by each item, is the core element for determining the accuracy of a student's knowledge assessment and the quality of the diagnostic information that is used for individual remediation (DeCarlo, 2012). Considering an assessment consisting of J items measuring on a domain of K attributes, a Q-matrix is represented as a $J \times K$ matrix and the element is denoted as q_{jk}, where q_{jk} =1 indicates that attribute k is required by item j, and q_{jk} = 0 indicates that attribute k is not required by item j.

Applications of CDAs normally assume that the Q-matrix is properly defined. However, a Q-matrix might be incorrectly specified and thus not reflect the true item-attribute alignment, because it is often constructed based solely on human opinions. If a Q-matrix is not specified correctly, inferences resulting from the application of the CDMs will not be valid. Previous studies (Baker, 1993; DeCarlo, 2011; Im & Corter, 2011; Kunina-Habenicht, Rupp, & Wilhelm, 2012; Rupp & Templin, 2008a) have shown that a misspecified Q-matrix would lead to undesirable consequences, including poor model fit, inaccurate model parameter estimation and respondent

classification, as well as incorrect interpretation of the measured attributes. This leads to the central theme of this dissertation: practical issues regarding the optimization of the Q-matrix in the context of CDA applications.

This dissertation contains one research study on Q-matrix misspecification and two research studies on Q-matrix validation in CDAs. Specifically, the first study investigates the impact of different types of Q-matrix misspecification on classification accuracy and consistency of CDA results. Given that the prime goal of CDAs is to provide detailed diagnostic information about students' knowledge status in specific aspects of learning, classification results produced by CDAs must be accurate. To achieve this goal, Cui, Gierl, and Chang (2012) introduced two new classification indices, classification accuracy and consistency indices, and investigated their performance across several factors (e.g., item discrimination power, number of attributes, and sample size) with the DINA model. However, as the authors noted, it is crucial that researchers and practitioners are aware of the impact of certain other important factors, such as the accuracy and the structure of the Q-matrix, on their performance. The first study extends the study conducted by Cui, Gierl, and Chang (2012) to investigate the degree to which diagnostic classification accuracy and consistency are affected by two types of Q-matrix misspecification, Q-entry misspecification and attribute misspecification across various conditions, in the context of CDAs.

The second study is motivated by the desire to obtain a correctly specified Q-matrix for successful implementation of CDAs, considering the statistical consequences of a misspecified Q-matrix. For decades, several methods have been developed to facilitate the process of obtaining an optimal Q-matrix (Barnes, 2003, 2010; Chiu, 2013; DeCarlo, 2012; de la Torre, 2008; Desmarais, 2011; Desmarais, Beheshti, & Naceur, 2012; Desmarais, & Naceur, 2013; Liu, Xu, & Ying, 2012, 2013; Templin & Henson, 2006a). The results of these efforts can be classified into

two broad categories of methods: Q-matrix validation and Q-matrix reconstruction. However, all previous studies of Q-matrix validation have used one single method for their analyses without any substantive comparative explanations about why a specific method was selected over the others. There is no study, to the best of our knowledge, that systematically evaluates the performance of these Q-matrix validation methods based on a methodological perspective and on metrics that allow meaningful comparison. To fill this critical gap in the literature, the second study compares the three most commonly used Q-matrix validation methods, including the sequential EM based δ-method (de la Torre, 2008), the Bayesian estimation method (DeCarlo, 2012), and the nonparametric Q-matrix refinement method (Chiu, 2013), with both basic and complex assessment design factors within the CDA framework.

The third study extends the second study on obtaining an optimal Q-matrix by proposing a two-stage cross-validation method allowing for more accurate and efficient computation within the CDA framework. Based on the assessment results of the three most commonly used methods, the Bayesian estimation method performs the best in validating a Q-matrix under almost all conditions. However, it has two major disadvantages that limit its application in real word: (1) it requires pre-assignment of the misspecified or uncertain Q-entries in advance, and (2) it requires an intensive computation load, especially when an assessment consists of a large number of items or measured attributes. To make the Q-matrix validation process more efficient, the third study proposes a two-stage cross-validation method that combines the strengths of the nonparametric refinement method and the Bayesian estimation techniques, and then investigates its effectiveness in both simulated and empirical data settings.

In sum, this dissertation aims to address the following research questions that are related to the Q-matrix of CDAs:

(1) What are the impacts of different types of Q matrix misspecification on classification accuracy and consistency of diagnostic results in CDAs?

(2) Among the three most commonly used Q-matrix validation methods, which method achieves the best performance on validating a misspecified Q-matrix? Is their performance affected by different assessment design factors?

(3) Compared to the three most commonly used Q-matrix validation methods, does the proposed two-stage cross-validation method identify and correct misspecified Q-entries more accurately and efficiently under a wide range of conditions? Does it still work well in empirical data settings?

The remainder of this dissertation is structured as follows: Chapter 2 reviews previous research regarding CDAs, Q-matrix misspecification, and Q-matrix optimization methods. Chapters 3, 4 and 5 address the three research questions, respectively. Chapter 6 makes conclusions and discusses the future development of the research.

CHAPTER 2: REVIEW OF LITERATURE

This dissertation is focused on applications of CDAs and issues in Q-matrix optimization. This chapter starts with the introduction of two popular CDMs, the DINA and DINO models, followed by a review of previous research on Q-matrix misspecification in the context of CDAs. In the last section of this chapter, research on Q-matrix optimization methods is extensively reviewed, including both Q-matrix validation and Q-matrix reconstruction methods. Comments and a general conclusion are also provided for each section.

Cognitive Diagnostic Models

There are several extensive reviews of CDMs in the literature, such as DiBello, Roussos, & Stout (2007), Fu and Li (2007), and Rupp and Templin (2008b). This section concentrates on two popular CDMs, the DINA and DINO models that will be used in the rest of this dissertation. The DINA model is a conjunctive model; it assumes that answering an item correctly requires the conjunction of all the required attributes. The DINO model is a disjunctive model, assuming that a correct response to an item may occur when one or more required attributes are mastered.

The DINA model. The DINA model is one of the simplest CDMs (Rupp & Templin, 2008b) and has been the foundation of several approaches for making cognitive diagnostic inferences. In the DINA model, for each item, certain attributes are required to answer the item correctly. An examinee must possess all the required attributes for an item to answer the item correctly.

Let Y_{ij} be the response of examinee i to item j where $Y_{ij} = 1$ means correct response and $Y_{ij} = 0$ means incorrect response; q_{jk} =1 (or 0) represents the entry of a Q-matrix that indicates attribute k is required (not required) by item j; α_{ik} =1 represents examinee i possesses attribute k and 0 otherwise, and $\alpha_{i.} = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iK})$ denotes the vector of total K attributes that are

required to answer item j correctly. Parameter η_{ij} indicates whether examinee i possesses all the required attributes for item j, where

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

Note that η_{ij}=1 if examinee i possesses all required attributes k, and η_{ij}=0 if examinee i lacks at least one required attribute. The vectors $\eta_{i.} = (\eta_{i1}, \eta_{i2}, \ldots, \eta_{ij})$ are called *ideal response patterns*, and the latent vectors $\alpha_{i.} = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik})$ are called *knowledge state* (Junker & Sijtsma, 2001; Tatsuoka, 1995). Both vectors represent the deterministic input part of the DINA model that indicates a deterministic prediction of task performance from each examinee's knowledge state (Rupp & Templin, 2008b).

In the DINA model, the relationship between the latent response variable η_{ij} and the observed item performances Y_{ij} is represented by two error probabilities:

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

and

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

where s_j and g_j are false negative and false positive rates, respectively. Specifically, s_j denotes the probabilities of slipping, which is getting an incorrect answer on item j when all required attributes are possessed, and $g_j$denotes the probabilities of guessing, which is getting a correct answer on item j when at least one required attribute is missing.

In the DINA model, each η_{ij} acts as an "and" gate combining the deterministic inputs $\alpha_{ik}{}^{q_{jk}}$, and each response Y_{ij} is modeled as a noisy observation of each η_{ij}. The item response function for obtaining $Y_{ij} = 1$ on item j is as follows:

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

Assuming local independence of items and independent examinees, the conditional likelihood for all of the responses under the DINA model is as follows:

$$P(Y_{ij} = y_{ij}|\boldsymbol{\alpha}, s, g) = \prod_{i=1}^{I} \prod_{j=1}^{J} ((1 - s_j)^{y_{ij}} s_j{}^{1-y_{ij}})^{\eta_{ij}} (g_j{}^{y_{ij}} (1 - g_j)^{1-y_{ij}})^{(1-\eta_{ij})} \quad (2)$$

The DINA model has enjoyed particular attention because it is one of the most parsimonious CDMs and is easy to interpret (de la Torre, 2008; Huebner, 2010; Rupp & Templin, 2008b). Specifically, this model requires only two parameters for each item, slipping and guessing parameters, regardless of the number of underlying latent attributes.

The DINO model. The deterministic inputs, noisy "or" gate (DINO) model is defined in a way similar to the DINA model. As in the DINA model, there is a gate component in the DINO model, which is determined by a disjunctive variable ω_{ij} instead of η_{ij},

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

where α_{ik} and q_{jk} are as previously defined in the DINA model, ω_{ij} = 1 if examinee i has mastered at least one required attribute of item j, and ω_{ij} = 0 when examinee i has mastered none of the required attributes of item j (Templin & Henson, 2006b). The DINO model also has slipping and guessing parameters:

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

and

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

The probability of correct response based on ω_{ij} is defined as:

$$P(Y_{ij} = 1|\alpha_{ij}) = (1 - s_j)^{\omega_{ij}} g_j{}^{(1-\omega_{ij})}. \quad (3)$$

Although the item response function of the DINO model is quite similar to the DINA model, the meanings of the slipping and guessing parameters are slightly different. In the DINO model, slipping s_j denotes the probability of failure on item j when one or more required attributes are possessed; guessing g_j denotes the probability of answering item j correctly when all required attributes are not mastered.

Research on Q-matrix Misspecification

One core component in the application of CDAs is the construction of a Q-matrix that identifies which attributes are required to successfully answer a specific item. Four commonly used Q-matrix construction methods include: One method is the think-aloud protocol method in which examinees are asked to verbalize their thinking and reactions as they do a test (Leighton & Gierl, 2007). A second method uses eye-tracking in which a headset with a camera is used to investigate different response patterns of examinees to specific tests (Rupp, Templin, & Henson, 2010). A third method uses subject-matter expert analysis in which panels of subject-matter experts (e.g., item developers from professional testing companies or school teachers) are asked to carefully inspect items and determine the required attributes for each item based on their professional experience. Their opinions are then collected and aggregated to form a Q-matrix (Aryadoust, 2011; Kim, 2011; Lee & Sawaki, 2009). The fourth method combines subject-matter expert analysis and think aloud protocol methods (Liu, You, Wang, Ding, & Chang, 2013).

None of the Q-matrix construction methods is perfect. Each method has its own disadvantages when applied to real-life situations. For instance, the think-aloud protocol method requires the participation of individual examinees, and the eye-tracking method requires external resource tools to record examinee' responses. These requirements limit their applications in large scale assessments. The item-attribute alignment for the subject-matter expert analysis method is

dependent only on the opinions of subject-matter experts; thus, their judgment might be subjective. In particular, different subject-matter experts may have different opinions on the specific relationship between the required attributes and the items. For this reason, even though so much care has been placed on constructing an initial Q-matrix, it is still possible that the Q-matrix is incorrectly identified or mis-specified.

Several studies have investigated the effects of a misspecified Q-matrix in various contexts of CDAs (Baker, 1993; DeCarlo, 2011; Im & Corter, 2011; Kunina-Habenicht, Rupp, & Wilhel, 2012; Rupp & Templin, 2008a). Baker (1993) investigated the sensitivity of the linear logistic test model (LLTM; Fischer, 1973) to the misspecification of a Q-matrix. Using six levels of misspecification and four sample sizes for two types of Q-matrices, sparse Q-matrix and dense Q-matrix, Baker found that even a small amount (1%-3%) of Q-matrix misspecification can have a considerable impact on parameter estimation in the LLTM, and 5% to 10% Q-matrix misspecification can seriously degrade the credibility of parameter estimation.

Rupp and Templin (2008a) conducted a simulation study to investigate the effects of different Q-matrix misspecification on item parameter estimation and respondent classification accuracy for the DINA model. The misspecification of the Q-matrix investigated in their study were of two sets: (a) underfitting, overfitting, or a balanced misfit of the Q-matrix for blocks of items that required a fixed number of attributes; and (b) incorrect dependency assumptions about two attributes. To investigate the impact of Q-matrix misspecification on item parameter estimates, Rupp and Templin examined the estimates of both slipping and guessing parameters across multiple conditions and their mean absolute deviation (MAD) values. To measure the effects of Q-matrix misspecification on respondent classifications, they investigated various global correlational measures (i.e., Kappa, lambda, Cramer's V and contingency coefficient) and the

individual cross-classification tables. The Q-matrix they used was composed of four attributes and 15 items. Results showed that the effects of Q-matrix misspecification on the estimation of slipping and guessing parameters for the DINA model are predominantly local effects that affect only items from which the Q-matrix was mis-specified. Specifically, when attributes are incorrectly deleted from the Q-matrix, the slipping parameter for a misspecified item is strongly overestimated; when attributes are unduly added in the Q-matrix, the guessing parameter for the misspecified item is strongly overestimated. Hence, large values of slipping and guessing parameters of sets of items with the same attribute specifications can provide empirical evidence for Q-matrix misspecification.

Im and Corter (2011) investigated the statistical consequences of attribute misspecification with the rule space method for cognitive diagnostic measurement. Two types of attribute misspecification examined in their study were (1) exclusion of an essential attribute that is required for answering an item correctly and (2) inclusion of a superfluous attribute that is not necessary for answering an item correctly. The Q-matrix they used was composed of seven attributes and 20 items. Their results focused on the estimation of examinees' characteristics, since results of the rule space method do not include estimation of item characteristics. Results showed that exclusion of an essential attribute resulted in underestimation of examinees' mastery probabilities for the remaining attributes, while the inclusion of a superfluous attribute yielded underestimation of examinees' mastery probabilities for the other attributes. In addition, when an attribute is in an order relation with a misspecified attribute, the order relationships affect the bias of the estimated attributes' mastery probabilities in systematic ways. These results underscored the importance of correct attribute specification for cognitive diagnostic assessment.

DeCarlo (2011) investigated the impact of Q-matrix misspecification on latent class sizes by applying the DINA model to the widely used fraction subtraction data. Results showed that classification of examinees obtained from CDMs and the latent class size were largely associated with the specification of the Q-matrix. If latent class size estimates for one or more attributes are close to unity, the possibility of Q-matrix misspecification, such as the inclusion of an irrelevant attribute, should be considered.

Kunina-Habenicht, Rupp, and Wilhelm (2012) investigated the effects of model misspecification due to Q-matrix misspecification in terms of item parameter estimation and respondent classification within a log-linear modeling framework. Similarly, two types of Q-matrix misspecification were considered in their study: (a) random permutations of 30% of all Q-entries while matching the marginal distributional properties across items and attributes to those of the correct Q-matrix as closely as possible; and (b) misspecification of the number of attributes in the Q-matrix so that a three-dimensional model was estimated for data generated from a five-dimensional model and vice versa. Results showed that Q-matrix misspecification led to notably decreased classification accuracy and had a dramatic effect on parameter recovery of latent class distributions, correlations and attribute proportions. Item-fit indexes, the MAD and the root mean error of approximation (RMSEA) were more strongly sensitive to over specification than to under specification of the Q-matrix. Information based fit indexes, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) were sensitive to both over- and under specification of the Q-matrix.

Researchers have investigated the effects of misspecified Q-matrices with various CDMs. Their results have been consistent with each other and shown that misspecification of a Q-matrix leads to undesirable consequences such as poor model fit, inaccurate model parameter estimation

and incorrect classification of examinees and interpretations of the required attributes. Thus, Q-matrix misspecification has become an important concern in the implementation of CDAs.

Research on Q-matrix Optimization

To address concerns regarding the possible effects of Q-matrix misspecification, researchers in the field of psychological and educational measurement began to recognize the importance of Q-matrix optimization. For instance, Jang (2009) pointed out that "a sound Q-matrix is a prerequisite for drawing valid diagnostic inferences about a learner's skill competencies" (p.211), and that the development of Q-matrix validation methods should be taken into consideration in all cognitive diagnostic applications. Intuitively, the most basic and straightforward approach is to fit a model with all possible alternative Q-matrices, and then to identify the optimal Q-matrix by checking model fit indices such as AIC and BIC. However, this approach involves intensive computation. Given an assessment with J items and K attributes, there are 2^{K*J} possible Q-matrices to be investigated. And as the number of J or K increases, the number of possible alternative Q-matrices will grow exponentially.

In the literature, several Q-matrix optimization methods have been developed to facilitate the process of obtaining an optimal Q-matrix. These can be classified into two broad categories. One comprises Q-matrix validation methods that aim to detect and correct possible misspecified Q-entries for items when some elements of an existing Q-matrix are assumed to be known. Examples of Q-matrix validation methods include the sequential EM based δ-method (de la Torre, 2008), the Bayesian estimation method (Templin & Henson, 2006a; DeCarlo, 2012), and the nonparametric Q-matrix refinement method (Chiu, 2013). The second group of optimization methods, Q-matrix reconstruction methods, aims to derive a Q-matrix that best fits the data when the whole Q-matrix is unknown. Examples include the self-learning Q-matrix method (Liu, Xu, &

Ying, 2012, 2013), the Non-negative matrix factorization method (Desmarais, 2011; Desmarais, Beheshti, & Naceur, 2012; Desmarais & Naceur, 2013), and the Hill-climbing algorithm (Barnes, 2003, 2010). This section provides detailed descriptions of each Q-matrix optimization method. Discussion and comparison of these approaches follow.

Q-matrix Validation Methods

Sequential EM Based δ-method. de la Torre (2008) proposed the sequential EM based δ-method to validate a Q-matrix based on information from responses in the DINA model. Parameter δ is denoted as the differences in the probabilities of a correct response to a specific item for examinees who possess all the required attributes and examinees who do not. In other words, it is an item discrimination index by which items with higher δ values can differentiate between examinees more effectively than those with low δ values. According to de la Torre's rationale, for the DINA model, the Q-vector for item *j* is said to be the correct Q-vector if it maximizes the item discrimination index δ. Specifically, given a set of K latent attributes and 2^K possible attribute patterns $\boldsymbol{\alpha_l}$, the correct Q-vector for item *j* can be obtained as follows (de la Torre, 2008):

$$\boldsymbol{q_j} = \arg\max[P(X_j = 1|\eta_{ll'} = 1) - P\big(X_j = 1|\eta_{ll'} = 0\big)\big] = \text{argmax}\left[\delta_{jl}\right], \qquad (4)$$

where $\eta_{ll'} = \prod_{k=1}^{K} \alpha_{l'k}^{\alpha_{lk}}$, and $l, l' = 1,2, \ldots, 2^K - 1$.

To put it briefly, maximizing the item discrimination index δ is equal to minimizing the sum of the average slip and guessing parameters, s_j and g_j. For this reason, de la Torre (2008) proposed using the size of the slipping and guessing parameters to establish goodness-of-fit of models to data.

To search for a valid Q-matrix that maximizes item discrimination index δ more efficiently, a sequential search algorithm is provided as an alternative to the exhaustive search algorithm. It is

worth noting that this algorithm is based on the premises that: 1) for a given number of required attributes, the Q-vector with the least number of misspecified Q-entries has the least amount of shrinkage with regards to the optimal δ; and 2) when the inference of adding an attribute in a Q-vector can be separated, the item discrimination index δ can help to decide whether this is a required attribute. Based on these premises, the sequential search algorithm starts from comparing δ with single attribute patterns to obtain the first required attribute, say $a^{(1)}$, which results in the highest $\delta^{(1)}$. Then, this process moves to two-attribute patterns, then three-attribute patterns until K-attribute patterns. In general, the stopping rules for this process are: 1) when δ^m in step m is less than $\delta^{(m-1)}$ in step $m-1$, or 2) when $m = K$ (de la Torre, 2008). The sequential search algorithm is more efficient than the exclusive search algorithm because the exact number of δ^* that need to be computed is $(K_j + 1)K - (K_j^2 - K_j)/2$, where K_j is the correct number of attributes required for item j, which is significantly lower than the $2^K - 1$ using the exhaustive search algorithm. However, when real items are involved, $\delta_j$cannot be computed directly because true guessing and slipping parameters and the distribution of the attribute pattern are unknown. In addition, a clear separation between the groups $\eta_{\mathrm{j}} = 0$ and $\eta_{\mathrm{j}} = 1$ cannot be expected. Hence, the δ method is impractical for finding the correct Q-vectors when J or K are large. To solve this problem, de la Torre implemented the EM algorithm with cut-off points along with the δ method to avoid frequent data recalibration and to reduce intensive computation of loads in δ estimation. The sequential EM based δ-method can be implemented in Ox (Doornik, 2003) or R-CDM package (Robitzsch, Kiefer, George & Uenlue, 2015).

de la Torre (2008) conducted a simulation study to investigate the effectiveness of the sequential EM based δ-method for the DINA model. In terms of the effect of Q-matrix misspecification, the results showed that the parameter estimates for items with misspecified Q-

vectors have large biases and shrunken δ, and some correctly specified items were also affected by the Q-vector misspecification. When five cut-points were used for selecting candidate Q-vectors in the sequential EM based δ-method, de la Torre found that this sequential EM based δ-method could identify and correctly replace the misspecified Q-vectors in the Q-matrix while simultaneously retaining the correctly specified Q-vectors. Hence, the sequential EM based δ-method could be used for evaluating the appropriateness of a Q-matrix and revising a misspecified Q-matrix.

Bayesian Estimation Method. Using the DINA model, Templin and Henson (2006a) introduced a Bayesian estimation procedure to estimate entries in the Q-matrix by allowing uncertain elements in the Q-matrix. Simulation studies indicate that probabilistic Q-matrix specification could recover true structure when some elements are not known with certainty. Based on their study, DeCarlo (2012) proposed a Bayesian-extension DINA model for Q-matrix validation that specifies some elements of the Q-matrix as being random rather than fixed and use the posterior distributions to determine whether an uncertain Q-entry should be zero or one. In particular, the Bayesian-extension DINA model is based on the reparametrized deterministic input noisy and gate (RDINA) model (DeCarlo, 2011), which is a simpler but the mathematically equivalent form of the DINA model. The RDINA model is:

$$\text{logit}\, p\left(Y_{ij} = 1|\boldsymbol{\alpha}\right) = f_j + d_j \prod_{k=1}^{K} \alpha_{ik}{}^{q_{jk}} \quad , \tag{6}$$

where f_j is the guessing parameter (false alarm rate) and d_j is a discrimination parameter that indicates how well the item discriminates between examinees who possess the required attribute set and those who do not. Parameters of the DINA model can be easily recovered from parameters of the RDINA model; specifically,

$$g_j = \frac{\exp(f_j)}{1+\exp(f_j)} \quad \text{and} \quad s_j = 1 - \frac{\exp(f_j+d_j)}{1+\exp(f_j+d_j)}.$$

The Bayesian extension of the DINA model can be specified as:

$$p_j = p\left(Y_{ij} = 1|\boldsymbol{\alpha}\right) = \text{expit}\left(f_j + d_j \prod_{k=1}^{K} \alpha_{ik}{}^{q_{jk}}\right);$$

That is, Y_{ij} is conditionally independent and Bernoulli distributed (i.e., $Y_{ij} \sim \text{Bernoulli}(p_j)$).

In the Bayesian extension of the DINA model, the parameters d_j and f_j are specified as random. Instead of fixing all q_{jk} to zero or one, the uncertain q_{jk} are treated as Bernoulli random variables $\tilde{q}_{jk}$ with probability p_{jk} (i.e., $\tilde{q}_{jk} \sim \text{Bernoulli}(p_{jk})$).

A Beta prior, with hyperparameters α and β, is used for p_{jk},

$$p_{jk} \sim \text{Beta}(\alpha, \beta).$$

The Beta distribution is the conjugate prior for the Bernoulli distribution; therefore, the posteriors of p_{jk} given $\tilde{q}_{jk}$ also have Beta distributions,

$$p_{jk}\left|\tilde{q}_{jk}\right. \sim \text{Beta}\ (\alpha + \tilde{q}_{jk}, \beta + 1 - \tilde{q}_{jk})$$

with the mean

$$E(p_{jk}\left|\tilde{q}_{jk}\right.) = \frac{\alpha + \tilde{q}_{jk}}{\alpha + \beta + 1}.$$

The posterior density of $\tilde{q}_{jk}$ or p_{jk} is used to determine whether an attribute should be included for an item; that is, whether $\tilde{q}_{jk}$ is one or zero. Specifically, the posterior mean of $\tilde{q}_{jk}$ is rounded by using a cut point of 0.5 for the posterior mean of p_{jk} (DeCarlo, 2012),

$$q_{jk} = \begin{cases} 0, & \text{if } p_{jk} < 0.5 \\ 1, & \text{otherwise} \end{cases}.$$

For Q-matrix validation, the Bayesian-extension DINA model based method requires that the possibly misspecified Q-entries are identified in advance. The possibly misspecified Q-entries are treated as random variables and estimated simultaneously with other parameters in the model using an MCMC estimation algorithm.

DeCarlo conducted a simulation study to exam whether the Bayesian estimation method helps recover the true Q-matrix elements when there are uncertain Q-entries in eight conditions. Results showed that the posterior distributions for the random Q-matrix elements are useful for providing information about which elements should or should not be included. Although results showed excellent recovery rates in many conditions, recovery rates were not always 100%, which was different from what was reported in Templin and Henson (2006a). In fact, the recovery rate can be adversely affected by some uncertain elements when others are not correctly specified.

Nonparametric Q-matrix Refinement Method. Chiu (2013) developed the nonparametric Q-matrix refinement method for identifying and correcting the misspecified Q-entries of a Q-matrix in the context of the DINA model. Using the weighted Hamming distance, the nonparametric Q-matrix refinement method operates by minimizing the residual sum of square (RSS) computed from the observed response and the ideal response to each test item. The use of RSS as a loss function to identify the Q-matrix misspecification is based on the idea that when all examinees are correctly classified, the correct Q-vector for a specific item is expected to have the lowest RSS among all possible Q-vectors. Thus, the correct Q-matrix is expected to minimize the overall RSS of the test, given the independence between the RSS of each item. Specifically, let Y_{ij} and η_{ij} be the observed and ideal item responses of examinee i to item j, respectively. The RSS of item j for examinee i equals

$$\mathrm{RSS}_{ij} = \left(Y_{ij} - \eta_{ij}\right)^2,$$

and the RSS of item j across all examinees equals

$$\mathrm{RSS}_{\mathrm{j}} = \sum_{i=1}^{I}(Y_{ij} - \eta_{ij})^2 = \sum_{m=1}^{2^K}\sum_{i \in C_m}(Y_{ij} - \eta_{jm})^2 \,, \tag{7}$$

where C_m is the latent proficiency-class m, and I is the number of examinees. Note that the Q-matrix refinement method adopts the nonparametric classification method (Chiu & Douglas, 2013)

to classify examinees, where ideal item responses are class-specific; that is, examinees classified as being in the same class have the same ideal item responses.

The algorithm of the nonparametric Q-matrix refinement method starts by identifying the item with the highest RSS, which is most likely to be a misspecified Q-vector. Next the algorithm searches over all $2^K - 1$ possible Q-vectors and replaces them with the one having the lowest RSS. The process is iterative, and the termination rule is: 1) all items have been visited; and 2) the RSS of each item no longer changes. More specifically, following Chiu (2013), the steps of the algorithm are as follows:

Step 0: Initialize the search item pool as $\boldsymbol{S^{(0)}} = \{1, \ldots, J\}$ and the input Q-matrix as $\boldsymbol{Q^{(0)}}$.

Step 1: Use the nonparametric classification method to estimate examinees' class membership based on $\boldsymbol{Q^{(0)}}$.

Step 2: Estimate the ideal item responses of all examinees based on $\boldsymbol{Q^{(0)}}$ and the class membership estimated for them in Step 1.

Step 3: Compute the mean RSS across examinees for each observed response and its corresponding ideal response for each item. Select the item in $\boldsymbol{S^{(0)}}$ with the highest RSS. If the highest RSS occurs for item j, then denote the Q-vector for the item is $q_j^{(1)}$, where the superscript (1) is the rank of the corresponding RSS among all items.

Step 4: Compute each remaining $2^K - 2$ RSS by replacing $q_j^{(1)}$ in $\boldsymbol{Q^{(0)}}$ with the other $2^K - 2$ Q-vectors, one at a time.

Step 5: Update $\boldsymbol{Q^{(0)}}$ by replacing $q_j^{(1)}$ with $q_j^{*(1)}$, the Q-vector with the lowest RSS among all the $2^K - 1$ possible Q-vectors. The updated Q-matrix is denoted as $\boldsymbol{Q^{(1)}}$.

Step 6: Omit item j out of the searching item pool. That is, $\boldsymbol{S^{(1)}} = \boldsymbol{S^{(0)}} \backslash \{j\}$.

Step 7: Replace $\boldsymbol{Q}^{(\mathbf{0})}$ and $\boldsymbol{S}^{(\mathbf{0})}$ with $\boldsymbol{Q}^{(\mathbf{1})}$ and $\boldsymbol{S}^{(\mathbf{1})}$, respectively, and repeat Step 1 to Step 6. Iterate until all items have been visited.

Step 8: Repeat Step 1 to Step 7 until the RSS of each item no longer changes.

This algorithm starts by identifying the item with the highest RSS and determining whether the Q-vector should be updated. Every update of the Q-matrix leads to the reclassification of examinees and the decrease of RSS of each item, which may cause additional updates to the Q-vectors. Therefore, all items must be visited several times until the RSS of each item no longer changes. This algorithm is very efficient because it requires only $(2^K - 1)$ computations to refine and validate a Q-matrix at step 0. The nonparametric Q-matrix refinement method can be implemented in R-NPCD package (Zheng & Chiu, 2015).

Chiu (2013) conducted three simulation studies to evaluate performance of the nonparametric Q-matrix refinement method: (a) the effectiveness, efficiency, and applicability of the Q-matrix refinement method; (b) the effects of the number of misspecified Q-vectors and the number of misspecified Q-entries on Q-matrix recovery; (c) effect of misspecification type on Q-matrix recovery. Results demonstrated that the nonparametric Q-matrix refinement method could recover the correct Q-matrix from a misspecified Q-matrix across various conditions effectively and efficiently. In addition, the number of misspecified Q-vectors and Q-entries, and the misspecification types had little effect on the performance of this method, which proved its general applicability.

Discussion. Despite the important role that a Q-matrix plays in CDAs, there is scant research devoted to validating a Q-matrix. The amount of research devoted to this topic is limited partly because (a) cognitive diagnosis is a relatively new area in psychometrics and (b) the Q-matrix is often treated as fixed in current applications of CDAs. Among all three methods reviewed

above, the sequential EM based δ-method proposed by de la Torre (2008) is the pioneer study on Q-matrix validation. Results demonstrated that this method might be used to correct a misspecified Q-matrix under two conditions: (a) the response data is modeled by a DINA model; and (b) the number of misspecified Q-vectors is small compared to the total number of items. However, this method is based on only one model fit index δ. As de la Torre pointed out, other statistics that are more appropriate or useful might exist for Q-matrix validation. The Bayesian model-based method proposed by DeCarlo (2012) appears to be useful for detecting which attribute should be included or excluded for each item when the Q-matrix and the number of attributes are at least generally correctly specified. However, this method is limited to the Beta prior and particular Q-matrix uncertainty structures. In addition, it requires that the possible misspecified entries in the Q-matrix are specified in advance, which limits its application for empirical data settings. The nonparametric Q-matrix refinement method proposed by Chiu (2013) enjoys three advantages: (a) it does not rely on the estimation and model parameters and makes no additional assumptions other than those made by the utilized CDM; (b) it does not require a large number of examinees; (c) nor does it require excessive computational time. Thus, it is best for small and medium-sized educational testing programs. However, this method is limited in terms of dealing with the misspecification of the total number of attributes and in detecting attributes that have been entirely missed or misspecified.

Although all three methods aim to validate the Q-matrix, it is still the consensus among researchers that a more comprehensive process of Q-matrix validation should consider both statistical information and substantive expertise. For instance, de la Torre (2008) concluded that "decisions based purely on statistical information can be misleading" and "for a successful

implementation of Q-matrix validation, and cognitive diagnostic modeling for that matter, a collaboration between experts from various fields cannot be overemphasized" (p.361).

The reviewed three Q-matrix validation methods are all related to the DINA model which assumes only two possible correct response probabilities for each item. Since numerous other CDMs are available for CDA applications, it is useful to validate the Q-matrix and evaluate its performance beyond the DINA model, such as the DINO model. There are cases where not all required attributes for an item have to be mastered for a correct answer. Thus, the development of Q-matrix validation for disjunctive CDMs is also needed.

Q-matrix Reconstruction Methods

Hill-climbing Algorithm. Barnes (2003, 2010) applied a hill-climbing algorithm to extract a matrix representing the relationship between concepts (attributes) and questions directly from student response data. This algorithm varies K, the number of concepts, and the values in the matrix to minimize the total error for all students for a given set of test questions. To avoid local minima, each hill-climbing search starts with a different random Q-matrix, and the best Q-matrix is saved.

This algorithm first sets the number of concepts to one and generates a random Q-matrix with values zero or one. 2^K concept states are also generated, and the ideal response vector (IDR) for each concept state is calculated. For example, as shown in Table 1, for the state *01*, and a 2 by 5 Q-matrix, the column for each question is examined to find the corresponding ideal response vector. For questions 1 and 2, concept 1 is not required for answering each correctly, while concept 2 is required. Since the concept state of the student is 01, the student should answer questions 1 and 2 correctly. However, questions 3, 4, and 5 all require understanding concept 1 to answer each

correctly. The student is more likely to answer questions 3, 4, and 5 incorrectly. Thus, the IDR for this concept state would be *11000*.

Table 1: Example of a 2 by 5 Q-matrix

	Questions				
	1	2	3	4	5
Concept 1	0	0	1	1	1
Concept 2	1	1	1	1	0

The next step is to compute the total error of Q-matrix over all students. For efficiency, an array of size 2^n of all possible response vectors is created, and the i^{th} element of the array contains the number of students with response vector i. Each observed student response vector is compared to each IDR and is assigned to the one that is the closest Hamming distance to it, which is the closest IDR. The distance from the response to the IDR is the error for that response vector. The total error of Q-matrix is computed by means of multiplying the individual errors for each response vector by the total number of students with that response and summing over all observed response vectors. If the overall Q-matrix error is improved, the change of the values in the Q-matrix is saved. This process is repeated several times for all values in the Q-matrix until the overall Q-matrix error is not changing significantly.

After a Q-matrix is obtained in this fashion, the hill-claiming algorithm runs several times again with a new random Q-matrix, and the Q-matrix with the smallest number of errors is saved for avoiding a local minimum. To determine the best number of concepts for the Q-matrix, this algorithm is repeated by increasing the number of concepts K until a stopping criterion is met: (a) the overall Q-matrix error is smaller than a pre-set threshold, such as less than 1 per student; or (b) there is a decrease in the marginal reduction of error by adding more concepts.

Barnes (2003, 2010) conducted several studies to investigate the effectiveness of the hill-claiming algorithm on student data from an online tutorial system. Results showed that, as a data mining method, this algorithm had several advantages over other data mining techniques, including factor analysis, cluster analysis, and discriminant analysis: (a) it can automatically build an interpretable model for data without a priori knowledge of concepts or clusters needed for a group of students; (b) It requires much less data for acceptable performance than other methods, such as factor analysis, that are based on a covariance analysis; and (c) it can determine the number of concepts automatically when a stopping criterion is met. By comparing the Q-matrix extracted by the hill-claiming algorithm with the one that was defined by experts with real student data, he found that the extracted Q-matrix and the expert-created Q-matrix often did not correspond, but the correspondence that did occur was usually on the most difficult or complex questions. Results also demonstrated that the extracted Q-matrix could understand the relationships shown in the Q-matrix; hence, the interpretation of the extracted Q-matrix could be used to both understand student data and determine which questions were most difficult.

Matrix Factorization Method. Desmarais (2011) used non-negative matrix factorization (NMF; Lee & Seung, 2001) to automatically map latent attributes to items for constructing a Q-matrix from data. Like principal component analysis, non-negative matrix factorization is often used for dimensionality reduction by decomposing a matrix into two smaller matrices:

$$\boldsymbol{V} \approx \boldsymbol{W}\,\boldsymbol{H} \tag{8}$$

Here $\boldsymbol{V}$ is a $J \times N$ matrix that represents the observed responses of N examinees to J items, $\boldsymbol{W}$ is a $J \times K$ Q-matrix with K attributes, and $\boldsymbol{H}$ is a $K \times N$ matrix that represents the attributes mastery for each of the N examinees. Each entry in $\boldsymbol{W}$, $\boldsymbol{H}$, and $\boldsymbol{V}$ is restricted to be non-negative, which implies that the K attributes are additive causes that contribute to the success of items, and

that they can only increase the probability of success and not decrease it. The product of $\boldsymbol{W}$ and $\boldsymbol{H}$ yields an estimated result matrix $\widehat{\boldsymbol{V}}$, and the goal of matrix factorization is to minimize$\left\| \widehat{\boldsymbol{V}} - \boldsymbol{V} \right\|$. To achieve this goal, Lee and Seung's (2001) multiplicative updating rule is often used due to the simplicity of implementation. Specifically, for each iteration, every entry in $\boldsymbol{W}$ and $\boldsymbol{H}$ is updated multiplicatively to reduce the component wise Euclidean distance between $\boldsymbol{V}$ and $\boldsymbol{WH}$.

Desmarais (2011) showed that for simulated data, the non-negative matrix factorization method was effective in deriving the Q-matrix when there was only one attribute per item. However, their study was based on the additive (compensatory) model of attributes, where each attribute increases the probability of success of an item. Also, the performance of the non-negative matrix factorization method degraded with real data or under different ratios of variance between subject performance, item difficulty and skill mastery.

Desmarais, Beheshti, and Naceur (2012) extended the non-negative matrix factorization method to construct a Q-matrix with multiple attributes per item from student response data. Their study was based on the conjunctive model of attributes, which requires that each attribute be mastered for the success of an item. Results showed that the non-negative matrix factorization method could successfully derive a conjunctive Q-matrix from simulated data if items involve one or two attributes from a set of six attributes, and the slip and guess factor of the data were below 0.2. In addition, they also noticed that performance of this method degraded rapidly with an increased number of slip and guess parameters.

Although the non-negative matrix factorization method showed its potential for automatically deriving a Q-matrix from response data, there are still some issues that need to be addressed. The first issue is the interpretation of the Q-matrix obtained. Although the number of attributes *K* is allowed to be specified with the non-negative matrix factorization method, attributes

in the resulting Q-matrix are in an unpredictable order, which may cause interpretation difficulties. The second issue is that the resulting Q-matrix may not be unique and its different manifestations may vary widely, which may worsen the problem of interpretation.

To address these issues, Desmarais and Naceur (2013) proposed the Alternating Least-square factorization (ALS) method that starts the factorization process with an initial Q-matrix set to the expert Q-matrix. An advantage of this method is that it can be used for comparing the Q-matrix derived from data with the expert-based Q-matrix and further enhance the expert-based Q-matrix. Specifically, the ALS method starts with the response matrix $\boldsymbol{V}$ and an initial expert defined Q-matrix $\boldsymbol{W_0}$, then a least-squares estimate of the attribute matrix $\widehat{\boldsymbol{H}}_{\boldsymbol{0}}$ can be obtained by

$$\widehat{\boldsymbol{H}}_{\boldsymbol{0}} = (\boldsymbol{Q}_{\boldsymbol{0}}^{\boldsymbol{T}}\boldsymbol{Q}_{\boldsymbol{0}})^{-\boldsymbol{1}}\boldsymbol{Q}_{\boldsymbol{0}}^{\boldsymbol{T}}\boldsymbol{V} \tag{9}$$

where the $\widehat{\boldsymbol{H}}_{\boldsymbol{0}}$, a new estimate of the Q-matrix $\widehat{\boldsymbol{W}}_{\boldsymbol{1}}$, is also obtained by the least-square estimate

$$\widehat{\boldsymbol{W}}_{\boldsymbol{1}} = \boldsymbol{V}\widehat{\boldsymbol{H}}_{\boldsymbol{0}}^{\boldsymbol{T}}\,(\,\widehat{\boldsymbol{H}}_{\boldsymbol{0}}\widehat{\boldsymbol{H}}_{\boldsymbol{0}}^{\boldsymbol{T}}\,)^{-\boldsymbol{1}}, \tag{10}$$

and the estimation process goes on for estimating $\widehat{\boldsymbol{H}}_{\boldsymbol{1}}$ and $\widehat{\boldsymbol{W}}_{\boldsymbol{2}}$, etc. This alternation between equations (8) and (9) provides progressive refinement of the matrices $\widehat{\boldsymbol{H}}_{\boldsymbol{1}}$ and $\widehat{\boldsymbol{W}}_{\boldsymbol{2}}$ that better approximate the result matrix $\boldsymbol{V}$.

Desmarais and Naceur (2013) conducted a visual analysis to compare the ALS Q-matrix and the expert defined Q-matrix using the Tatsuoka's fraction algebra data. Results showed these two Q-matrices were relatively similar. The ALS Q-matrix preformed slightly better than the expert defined Q-matrix for making accurate response outcome predictions. Thus, the ALS factorization method can be used for deriving Q-matrix from response data and improving the initial expert defined Q-matrix.

Self-learning Q-matrix Method. Liu et al. (2012, 2013) proposed the self-learning Q-matrix theory to estimate the Q-matrix in the context of the DINA model. In particular, the

estimator of the Q-matrix is built only on the information of observed responses, and the estimation of the Q-matrix is based on an assessment of how well a given matrix Q fits the data. To obtain an estimator of Q-matrix, the authors introduced an important quantity, the T-matrix, which connects the Q-matrix with the observed response and attribute distribution. The T-matrix, $\boldsymbol{T}(\boldsymbol{Q})$, has $2^K - 1$ columns, each of which corresponds to one nonzero attribute vector, $\alpha \in \{0,1\}^K \backslash \{(0, \dots, 0)\}$. Let I_j be the notation of positive response to item j, and let "$\wedge$" be the notation of "and" combination. Each row of $\boldsymbol{T}(\boldsymbol{Q})$ corresponds to one item or one "and" combination of items. For example, $I_{j1} \wedge I_{j2}$ indicates positive responses to both items j_1 and j_2. Given $\boldsymbol{J}$ is the total number of items, $\boldsymbol{T}(\boldsymbol{Q})$ is defined as saturated if $\boldsymbol{T}(\boldsymbol{Q})$ contains $2^J - 1$ rows that include all the single items and all "and" combinations. Each column of $\boldsymbol{T}(\boldsymbol{Q})$ indicates if an examinee with the attribute vector could get the positive responses to the item combination.

A column vector β, the length of which equals the number of rows of $\boldsymbol{T}(\boldsymbol{Q})$, is define with each element of β corresponding to the number of people who have positive responses to the item combination. Thus, the linear equation is

$$\boldsymbol{T}(\boldsymbol{Q})\hat{P} = \beta \,, \tag{11}$$

where $\hat{P} = (\hat{P}_\alpha : \alpha \in \{0,1\}^K \backslash \{(0, \dots, 0)\})$ contains the estimated proportions of examinees with each attribute profile. For each binary matrix Q', an objective function was introduced,

$$\boldsymbol{S}(\boldsymbol{Q}') = \left| \boldsymbol{T}(\boldsymbol{Q}')\hat{P} - \beta \right| , \tag{12}$$

where $|\cdot|$ is the Euclidean distance. An estimate of a Q-matrix can be obtained by minimizing the $\boldsymbol{S}(\boldsymbol{Q}')$

$$\hat{Q} = \arg\inf_{Q'} \boldsymbol{S}(\boldsymbol{Q}') \,.$$

Liu et al. (2012) conducted simulation studies to investigate the performance of the self-learning Q-matrix theory with the DINA model by comparing the estimated Q-matrix and the true

Q-matrix. Results showed that the estimated Q-matrix could recover the true Q-matrix accurately when attributes have no special structure (e.g., uniform distribution). If attributes are correlated, the recovery rate was degraded. Moreover, additional information about the Q-matrix or the parametric form of the attribute distribution could substantially improve the estimation efficiency and reduce the computational complexity.

Liu et al. (2013) provided detailed theoretical analysis on the self-learnability of the underlying Q-matrix. In particular, they established sufficient conditions to ensure that attributes required by each item are learnable from the data. They also proved the consistency of results in the DINA model with known or unknown slipping parameters and a known guessing parameter. The authors also claimed that this method could be adapted to cover a large class of cognitive diagnostic models besides the DINA model, such as the DINO model, the NIDA model, and the NIDO model.

Discussion. Compared with validating a Q-matrix based on an expert defined Q-matrix, the problem of generating the Q-matrix completely from response data is more difficult and is "largely an unexplored area" (Liu et al., 2012). In the field of psychometrics, the self-learning Q-matrix theory proposed by Liu et al. (2012) is a pioneer study on the empirical estimation of Q-matrix from response data. Their results demonstrate the usefulness of this method and its implementation in estimating Q-matrices using various CDMs. However, this Q-matrix estimation approach requires several assumptions: (a) a saturated T-matrix; (b) a complete true Q-matrix exists so that for each attribute, there exists an item that only requires this particular attribute; and (c) guessing parameters for all items are known in the DINA model. If these assumptions are violated, performance of this method will be negatively affected. In addition, optimization of the function $\boldsymbol{S(Q)}$ over the space of $J \times K$ binary matrices requires evaluating the function $\boldsymbol{S(Q)}$ $2^{J \times K}$

times, which leads to a substantial computational load, especially when the number of J and K is reasonably large. Limitations imposed by strong assumptions and high computational cost are major impediments for widely using this method in practice.

Because of the importance of the Q-matrix in providing personalized interaction in intelligent learning environments such as online tutoring systems, a means of automatically deriving the Q-matrix from response data is also highly desirable in the field of educational data mining. Given that the goal of designing an online tutoring system is to accurately assess students' attribute profiles and personalize learning content in a relatively short period of time, researchers in the field of educational data mining aim to develop methods that can not only waive the labor-intensive task of assigning required attributes to corresponding items, but also offer a more objective and replicable means of deriving the Q-matrix (Desmarais, 2011). To achieve this goal, Barnes (2003, 2010) developed the hill-climbing algorithm to derive the Q-matrix from response data. This method is fully automated and has shown to perform better than other types of data mining techniques, including factor analysis, cluster analysis, and discriminant analysis in terms of attribute cluster analysis. However, discrepancies between the Q-matrix derived by the hill-climbing algorithm and the expert defined Q-matrix still exist. In addition, the hill-climbing algorithm only works well for small data sets with as few as 25 responses and a Q-matrix with less than 20 items. Thus, its applicability in large-scale assessments is limited.

Another method that has often been used for deriving the Q-matrix directly from response data is matrix factorization. Desmarais and his colleagues (2011, 2012) applied the NMF method to derive a Q-matrix with multiple attributes per item from student response data. This method was successful for simulated data, but its performance degraded rapidly with increased slip and guess parameters. Moreover, the Q-matrix derived from the NMF method may not be a unique solution

and may contain numerical values of various signs and amplitude that may cause interpretation difficulties. Thus, the ALS method (Desmarais & Naceur, 2013) was proposed to address the interpretation issue. Using an expert defined Q-matrix as a start point, the ALS method has proved to be a promising way of deriving a Q-matrix from response data and helping improve the expert defined Q-matrix. However, the ALS method is limited to a single real data case by Desmarais and Naceur in 2013. Further studies about the generalizability of this method to different dimensions of Q-matrix are needed.

The reviewed Q-matrix reconstruction methods are based solely on response data to derive all elements of the Q-matrix without considering the one designed by experts. However, it may cause problems in the real world because sometimes the automatically derived Q-matrices are not interpretable. Thus, it is not appropriate to ignore expert opinions completely. Desmarais and Naceur's study in 2013 showed this very clearly that researchers in the field of educational data mining began to recognize the importance of expert input. And they would like to improve the automatically derived Q-matrix by considering experts' opinions at the same time. This is consistent with the goals of the Q-matrix validation methods that are discussed in the previous sections. In the future, research that integrates techniques from both psychometrics and educational data mining would be highly useful to optimize the Q-matrix and further improve student learning in both classroom and intelligent learning environments.

Some general issues are raised in the review of Q-matrix optimization methods. First, the number of attributes is pre-defined in the current Q-matrix optimization methods. That means the attributes are fixed. However, in the real world, the number of attributes is not clear, even the specification of attributes to items is quite different from different experts. A possible solution is to use the goodness of model-data fit, such as AIC and BIC, to determine the best number of

attributes for each different model. Second, when the number of items or attributes becomes large, the computational load for optimizing the Q-matrix increases substantially. It is one of the most important impediments to applying them in the analysis of real data. Thus, it would be very useful to develop an innovative algorithm can be both quick and accurate. Third, the importance of expert opinion should not be ignored during the process of Q-matrix optimization because decisions based purely on statistical information can be misleading (de la Torre, 2008). A more effective Q-matrix construction process should be considered in which subject-matter experts and psychometricians work together to develop an initial Q-matrix, and then refine it utilizing both statistical information from Q-matrix optimization methods and substantive knowledge from subject-matter experts.

CHAPTER 3: THE EFFECTS OF Q-MATRIX MISSPECIFICATION ON CLASSIFICATION ACCURACY AND CONSISTENCY

CDAs have the potential to help teachers differentiate instructional needs for individual students by providing detailed feedback about students' knowledge and skill mastery status in specific aspects of learning. To fulfil such potential, the classification results produced by the CDAs must be valid and reliable. Standards 2.1 and 6.5 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME. 2014) ask for proof of adequate reliability and validity of any reported scores, but the reliability of diagnostic scores is often not reported (Templin & Bradshaw, 2013). To fill the gap in the literature on how to examine the classification accuracy and consistency of CDAs, Cui, Gierl, and Chang (2012) introduced two new classification indices, accuracy and consistency, as important indicators of the reliability and validity of classification results in CDAs. Their performance with the DINA model across several factors (e.g., item discrimination power, number of attributes, attribute dependency and sample size) has been investigated. However, the impact of some other factors, for example Q-matrix accuracy, on their performance remains unexplored. This study extends Cui et al.'s (2012) study by further investigating the impact of two types of Q-matrix misspecification, across multiple factors, on the classification accuracy and consistency of CDA results.

Method

Simulation Design

A simulation study was conducted to examine the impacts of two types of Q-matrix misspecification by manipulating two key factors: attribute dependency and data generation model.

The first type of misspecification studied was Q-entry misspecification. A review of the literature on Q-matrix misspecification showed that most application examples used a small

amount (1%-20%) of Q-entry misspecification (Baker, 1993; DeCarlo, 2011; Im & Corter, 2011; Kunina-Habenicht, Rupp, & Wilhel, 2012; Rupp & Templin, 2008a) and suggested a higher percentage of Q-entry misspecification for further investigation. Thus, three levels of Q-entry misspecification rates (10%, 20%, and 30%) were considered in this study. To generate the misspecified Q-matrices with different percentages of Q-entry misspecification, Q-entries in the correct Q-matrix were randomly selected at a specific probability (i.e., 10%, 20%, and 30%), and then replaced with opposite values (1s are replaced as 0s, and vice versa). Results of 0% Q-entry misspecification were provided as the baseline for comparison.

The second type of misspecification studied was attribute misspecification, including attribute exclusion and inclusion. Attribute exclusion means an essential attribute required for answering the test items correctly is excluded from the Q-matrix while attribute inclusion refers to an unnecessary attribute added to the Q-matrix that should not be involved for solving the test items. Both types of attribute misspecification are illustrated by the two Q-matrices (Q1 and Q2) shown in Table 2. Assuming that Q1 with measured attributes A1, A2, A3, A4, and A5 is the correct Q-matrix, then Q2 is a misspecified Q-matrix because an unnecessary attribute A6 is included. Conversely, if Q2 with measured attributes A1, A2, A3, A4, A5 and A6 is assumed as the correct Q-matrix, then Q1 is a misspecified Q-matrix because an essential attribute A6 is excluded.

Table 2: Q-Matrices for Illustrating Attribute Misspecification

Item	Q1					Q2					
No	A1	A2	A3	A4	A5	A1	A2	A3	A4	A5	***A6***
1	1	0	0	0	0	1	0	0	0	0	***1***
2	0	1	0	0	0	0	1	0	0	0	***1***
3	0	0	1	0	0	0	0	1	0	0	***1***
4	0	0	0	1	0	0	0	0	1	0	***1***
5	0	0	0	0	1	0	0	0	0	1	***1***
6	1	1	0	0	0	1	1	0	0	0	***1***
7	0	1	0	1	0	0	1	0	1	0	***1***
8	1	0	0	0	1	1	0	0	0	1	***1***
9	0	0	0	1	1	0	0	0	1	1	***1***
10	0	0	1	1	0	0	0	1	1	0	***1***
11	0	1	1	1	0	0	1	1	1	0	***1***
12	1	0	1	0	1	1	0	1	0	1	***1***
13	0	1	0	1	1	0	1	0	1	1	***1***
14	0	0	1	1	1	0	0	1	1	1	***1***
15	1	1	1	0	0	1	1	1	0	0	***1***
16	1	1	1	0	1	1	1	1	0	1	***1***
17	0	1	1	1	1	0	1	1	1	1	***1***
18	1	1	0	1	1	1	1	0	1	1	***1***
19	1	0	1	1	1	1	0	1	1	1	***1***
20	1	1	1	1	1	1	1	1	1	1	***1***
21	1	0	0	0	0	1	0	0	0	0	***1***
22	0	1	0	0	0	0	1	0	0	0	***1***
23	0	0	1	0	0	0	0	1	0	0	***1***
24	0	0	0	1	0	0	0	0	1	0	***1***
25	0	0	0	0	1	0	0	0	0	1	***1***
26	1	1	0	0	0	1	1	0	0	0	***1***
27	0	1	0	1	0	0	1	0	1	0	***1***
28	1	0	0	0	1	1	0	0	0	1	***1***
29	0	0	0	1	1	0	0	0	1	1	***1***
30	0	0	1	1	0	0	0	1	1	0	***1***
31	0	1	1	1	0	0	1	1	1	0	***1***
32	1	0	1	0	1	1	0	1	0	1	***1***
33	0	1	0	1	1	0	1	0	1	1	***1***
34	0	0	1	1	1	0	0	1	1	1	***1***
35	1	1	1	0	0	1	1	1	0	0	***1***
36	1	1	1	0	1	1	1	1	0	1	***1***
37	0	1	1	1	1	0	1	1	1	1	***1***
38	1	1	0	1	1	1	1	0	1	1	***1***
39	1	0	1	1	1	1	0	1	1	1	***1***
40	1	1	1	1	1	1	1	1	1	1	***0***

In addition to the two types of Q-matrix misspecification, two other factors were also manipulated in the simulation study. The first manipulated factor is attribute dependency. Two levels of attribute dependency were considered: independent and correlated. When attributes are independent, the mastery of one attribute does not correlate with the mastery of other attributes. Hence, students' true attribute patterns are assumed to follow a discrete uniform distribution of equal probabilities. When attributes are correlated, the mastery of one attribute most likely affects the mastery of another attribute. Thus, examinees' attribute patterns are assumed to follow a multivariate normal distribution *MVN* (0_K,$\boldsymbol{\Sigma}$), with mean vector zero and all correlations between the attributes in the correlation matrix $\boldsymbol{\Sigma}$ equal to 0.5. Let $\theta_{ik} = (\theta_{i1}, \ldots, \theta_{iK})$ be the underlying ability of examinees. The examinees' attribute pattern $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})$ is determined by

$$\alpha_{ik} = \begin{cases} 1, & if\ \theta_{ik} \geq \Phi^{-1}\left(\frac{k}{K+1}\right), \\ 0, & otherwise \end{cases}, \tag{13}$$

where $k = 1, \ldots, K$.

The second manipulated factor was the data generation model. Most previous research adopted the conjunctive DINA model to generate examinees' response data (Chiu, 2013; DeCarlo, 2011, 2012; de la Toree, 2008; Rupp & Templin, 2008a). To investigate the effects of Q-matrix misspecification in the context of both conjunctive and disjunctive ideal response patterns, both DINA and DINO models were used to generate data and estimate parameters.

We did not manipulate levels of the following factors: number of attributes, test length or sample size. The number of attributes was fixed at 5, test length set at 40 items, and sample size set to 2000. The [3 (Q-entry misspecification rate) + 2 (attribute misspecification)] × 2 (attribute dependency) × 2 (data generation model) design resulted in a total of 20 conditions, which can be

generally classified into eight groups (C1-C8), as shown in Table 3. For each condition, 100 data sets were generated to provide stable results.

Table 3. Simulation Conditions

Name	Data generation Model	Attribute Dependency	% of Q-entry misspecification	Attribute misspecification
C1	DINA	Independent	QM=10% QM=20% QM=30%	
C2	DINA	Correlated	QM=10% QM=20% QM=30%	
C3	DINO	Independent	QM=10% QM=20% QM=30%	
C4	DINO	Correlated	QM=10% QM=20% QM=30%	
C5	DINA	Independent		Attribute Exclusion Attribute Inclusion
C6	DINA	Correlated		Attribute Exclusion Attribute Inclusion
C7	DINO	Independent		Attribute Exclusion Attribute Inclusion
C8	DINO	Correlated		Attribute Exclusion Attribute Inclusion

Evaluation Criteria

Two classification indices, accuracy and consistency (Cui, Gierl, & Chang, 2012), were used as the evaluation criteria. Classification accuracy refers to the degree to which the classification of student latent classes based on observed item response patterns agrees with students' true latent classes. It is evaluated by the classification accuracy index P_a, defined as

$$P_a = P(X \in C_t) = \sum_{\alpha \in \Omega} \sum_{x \in \pi_t} P(X = x|\alpha) r_\alpha \quad (14)$$

Classification consistency refers to the degree to which classifications agree on the basis of two independent administrations or two parallel forms of the test. It is evaluated by the classification consistency index P_C, defined as

$$P_C = \sum_{\alpha \in \Omega} \left[\sum_{h=1}^{H} \left(\sum_{x \in \pi_h} P(X = x|\alpha) \right)^2 \right] r_\alpha \quad (15)$$

where H is the total number of attribute patterns, Ω is the countable space, r_α is the relative frequency of student attribute pattern α; C_t is a student's true latent class, C_h is the h^{th} latent class, π_h is the set of all possible item response patterns that would be classified into C_h, and $t = 1,2, \ldots, H$. Both indices are computed for each condition.

Results

Results of the simulation study are discussed focusing on the primary research inquiry: the impact of two types of Q-matrix misspecification on classification accuracy and consistency of CDA results. Note that classification results under the condition of 0% of Q-entry misspecification are considered as the baseline rate. For each condition, both pattern and attribute-level classification consistency and accuracy results are examined. Results of this study are presented in two sections, respectively. Specifically, the effects of different percentages of Q-entry misspecification are summarized in the first section, and the effects of attribute misspecification are presented in the subsequent section.

Effects of Different Percentage of Q-Entry Misspecification

Effect on classification accuracy. Table 3 presents a summary of both the pattern and attribute-level classification accuracy across conditions. On the whole, results suggest that the domain effect of Q-entry misspecification on both the pattern-level and attribute-level classification accuracy of CDA results is clearly visible. First, pattern-level classification accuracy index P_a decreases significantly as the percentage of Q-entry misspecification increases. For

example, in C1, when the percentages of Q-entry misspecification are 10%, 20% and 30%, P_a decreases sharply to 0.88, 0.67 and 0.63, respectively. Second, similar to the pattern-level results, classification accuracy indices for each attribute decrease as the percentage of Q-entry misspecification increases. However, at the attribute level, P_a decreases in various ways. For example, in C1, there is a steady decrease in the values of P_a for attributes 1, 4 and 5 while there is a sharp decrease in P_a for attributes 2 and 3. When the percentages of Q-entry misspecification are 10%, 20% and 30%, P_{a2} of attribute 2 decreases to 0.94, 0.56 and 0.36, respectively, while P_{a4} of attribute 2 decreases to 0.91, 0.95 and 0.88, respectively.

Table 4. Classification Accuracy Results with Q-entry Misspecification

Simulation Condition	% of Q-entry misspecification	Classification Accuracy					
		P_a	P_{a1}	P_{a2}	P_{a3}	P_{a4}	P_{a5}
C1	QM=0%	1.00	0.98	0.99	1.00	0.99	1.00
	QM=10%	0.88	1.00	0.94	0.95	0.91	0.99
	QM=20%	**0.67**	**0.69**	**0.56**	**0.48**	0.95	0.94
	QM=30%	**0.63**	**0.79**	**0.36**	**0.28**	0.88	**0.73**
C2	QM=0%	0.99	0.99	0.99	0.98	0.94	0.99
	QM=10%	0.92	0.91	1.00	0.97	0.95	**0.64**
	QM=20%	**0.79**	0.90	0.99	**0.74**	0.97	**0.70**
	QM=30%	**0.63**	**0.70**	**0.67**	**0.63**	0.90	**0.57**
C3	QM=0%	1.00	0.98	0.99	1.00	0.99	1.00
	QM=10%	0.88	0.87	1.00	0.95	0.95	0.92
	QM=20%	**0.67**	**0.68**	**0.54**	**0.54**	0.88	0.94
	QM=30%	**0.69**	0.86	**0.24**	**0.25**	0.82	**0.66**
C4	QM=0%	1.00	0.97	1.00	1.00	0.96	1.00
	QM=10%	0.92	0.92	1.00	0.98	0.98	0.98
	QM=20%	**0.76**	**0.82**	**0.47**	**0.42**	0.88	0.81
	QM=30%	**0.63**	0.86	**0.57**	**0.60**	0.93	0.89

Note. Bold values are smaller than 0.8.

Results in Table 4 also show that the two manipulated factors, attribute dependency and data generation model, have only a slight impact on both pattern-level and attribute-level classification accuracy. First, the values of the classification accuracy indices are very close to each other, no matter which model (i.e., the DINA or the DINO) is used for data generation. For

example, in C1, P_a for all three levels of Q-entry misspecification are 0.88, 0.67 and 0.63, respectively. And in C3, P_a for all three levels of Q-entry misspecification are 0.88, 0.67 and 0.69, respectively. Second, classification accuracy indices perform slightly better in the case of correlated attributes than for independent attributes. When percentages of Q-entry misspecification are 10% and 20%, P_a in C2 are 0.04 and 0.12 higher than that in C1, respectively.

Table 5. Classification Consistency Results with Q-entry Misspecification

Simulation Condition	% of Q-entry misspecification	Classification Consistency					
		P_c	P_{c1}	P_{c2}	P_{c3}	P_{c4}	P_{c5}
C1	QM=0%	0.93	1.00	0.96	0.99	0.99	0.99
	QM=10%	0.81	0.81	1.00	0.89	0.91	0.85
	QM=20%	**0.79**	0.92	0.98	**0.67**	0.93	**0.67**
	QM=30%	**0.65**	0.86	0.89	0.90	0.82	**0.72**
C2	QM=0%	0.95	0.99	0.98	0.98	0.96	0.89
	QM=10%	0.90	0.92	0.99	0.95	0.90	**0.67**
	QM=20%	**0.79**	0.92	0.98	**0.67**	0.93	**0.67**
	QM=30%	**0.60**	**0.78**	**0.62**	**0.74**	0.84	**0.65**
C3	QM=0%	0.93	1.00	0.96	0.99	0.99	0.99
	QM=10%	0.88	0.89	1.00	0.96	0.96	0.97
	QM=20%	**0.60**	**0.69**	**0.78**	**0.77**	0.83	0.90
	QM=30%	**0.63**	0.86	0.93	0.92	0.80	**0.79**
C4	QM=0%	0.94	1.00	0.95	0.99	0.99	0.93
	QM=10%	0.88	0.89	1.00	0.96	0.96	0.97
	QM=20%	**0.73**	0.87	**0.64**	0.81	0.83	**0.76**
	QM=30%	**0.61**	0.92	**0.65**	**0.65**	0.89	0.83

Note. Bold values are smaller than 0.8.

Effect on classification consistency. Table 5 presents a summary of both pattern and attribute-level classification consistency results across conditions. In general, the effects of Q-entry misspecification on classification consistency follow a similar pattern to that of classification accuracy. First, both pattern-level and attribute-level classification consistency indices decrease significantly as the percentage of Q-entry misspecification increases. For example, in C1, when the percentages of Q-entry misspecification are 10%, 20% and 30%, values of the pattern-level index P_c decrease to 0.81, 0.79 and 0.65, respectively. Second, the two factors, attribute

dependency and data generation model, have no noticeable effect on the classification accuracy results. For example, in the conditions of C2 and C4, P_c for all three levels of Q-entry misspecification are (0.90, 0.79, 0.60) and (0.88, 0.73, 0.61), respectively.

Effects of Different Types of Attribute Misspecification

Effect on classification accuracy. Table 6 summarizes the results of both pattern and attribute-level classification accuracy indices across conditions. First, the impacts of both types of Q-matrix misspecification on classification accuracy is examined. In general, both types of attribute misspecification have negative effects on classification accuracy. When an essential attribute is excluded, pattern-level classification accuracy index P_a across C5 and C8 drops to 0.83, 0.94, 0.86 and 0.92, respectively. When an unnecessary attribute is included, pattern-level classification accuracy index P_a across C5 and C8 declines to 0.95, 0.92, 0.95 and 0.90, respectively. Similarly, the attribute-level classification accuracy index P_{ai} for both types of attribute misspecification shows different degrees of decrease, ranging from 0.49 to 0.99. However, the decreasing characteristics show that attribute inclusion has a comparatively larger impact on classification accuracy than exclusion in almost all misspecification conditions. For example, in C8, P_a of attribute exclusion is 0.02 higher than that of attribute inclusion, and P_{ai} of the five attributes of attribute exclusion are 0.12, 0.25, 0.11, 0.10 and 0.02 higher than that of attribute inclusion, respectively. Second, results also show that when an unnecessary attribute 6 is included, P_{a6} is extremely low at 0.01, 0.18, 0.02 and 0.07 for the conditions of C5-C8, respectively. This finding indicates that attribute-level classification accuracy indices may be useful to provide researchers a way to identify possible attribute misspecification of a Q-matrix. Third, unsurprisingly, attribute dependency and data generation model have very limited impact on either the pattern or attribute-level classification accuracy. Results of C5 and C6 are very similar to those of C7 and C8.

Table 6. Classification Accuracy Results with Attribute Misspecification

Simulation Condition	Attribute Misspecification	Classification Accuracy						
		P_a	P_{a1}	P_{a2}	P_{a3}	P_{a4}	P_{a5}	P_{a6}
C5	Attribute Exclusion	0.83	0.90	0.94	0.89	0.94	0.98	
	Attribute Inclusion	0.95	0.80	**0.77**	0.84	0.81	0.99	**0.01**
C6	Attribute Exclusion	0.94	0.92	0.97	0.94	0.94	0.99	
	Attribute Inclusion	0.92	0.92	0.83	**0.76**	**0.57**	**0.75**	**0.18**
C7	Attribute Exclusion	0.86	0.89	0.96	0.89	0.93	0.98	
	Attribute Inclusion	0.95	0.81	0.80	0.86	0.84	0.99	**0.02**
C8	Attribute Exclusion	0.92	**0.61**	0.93	0.93	0.97	0.98	
	Attribute Inclusion	0.90	**0.49**	**0.68**	0.82	0.87	0.96	**0.07**

Note. Bold values are smaller than 0.8.

Effect on classification consistency. Table 7 presents a summary of both pattern and attribute-level classification consistency across conditions. Generally, the effects of attribute misspecification on classification consistency follow those observed in classification accuracy. Both pattern and attribute-level classification consistency indices decrease markedly when attribute misspecification is present but vary with groups of different levels. A moderate decrease is observed in pattern-level classification consistency index P_c, ranging from 0.83 to 0.95 for C1-C8. And a remarkable decrease is observed in attribute-level classification consistency indices P_{ci}, ranging from 0.59 to 1. Classification consistency of attribute exclusion, in general, is lower at the pattern-level, but higher at the attribute-level than for that of attribute inclusion. For example, in C6, P_c of attribute exclusion is 0.01 lower than that of attribute inclusion, while P_{ci} of the first five attributes of attribute exclusion are 0.04, 0.13, 0.13, 0.18 and 0.30 higher than that of attribute inclusion, respectively. In addition, attribute dependency and data generation model have no obvious influence on either pattern or attribute-level classification accuracy results, which is consistent with the results of classification accuracy.

Table 7. Classification Consistency Results with Attribute Misspecification

Simulation Condition	Attribute Misspecification	Classification Consistency						
		P_c	P_{c1}	P_{c2}	P_{c3}	P_{c4}	P_{c5}	P_{c6}
C5	Attribute Exclusion	0.83	0.83	0.90	0.82	0.89	0.97	
	Attribute Inclusion	0.93	**0.77**	**0.74**	**0.78**	**0.77**	0.97	1.00
C6	Attribute Exclusion	0.94	0.86	0.95	0.88	0.88	0.97	
	Attribute Inclusion	0.95	0.90	0.82	**0.75**	**0.70**	**0.67**	0.96
C7	Attribute Exclusion	0.82	0.82	0.92	0.82	0.88	0.97	
	Attribute Inclusion	0.93	**0.77**	**0.73**	0.80	**0.78**	0.98	1.00
C8	Attribute Exclusion	0.89	**0.59**	0.88	0.87	0.94	0.97	
	Attribute Inclusion	0.94	**0.64**	**0.68**	**0.78**	0.84	0.93	0.99

Note. Bold values are smaller than 0.8.

Summary and Discussion

The first study examined one of the most fundamental questions in the area of CDAs: *What are the impacts of different types of Q matrix misspecification on classification accuracy and consistency of diagnostic results?* The importance of this question cannot be overstated because it is directly related to the quality of diagnostic results and students' learning processes. To answer this research question, a simulation study was conducted to investigate the degree to which classification accuracy and consistency of diagnostic results are affected by two types of Q matrix misspecification: (1) Q-entry misspecification, and (2) attribute misspecification. In general, results have shown that both types of Q-matrix misspecification influence the classification accuracy and consistency of diagnostic results, but how and the degree of its influence vary. First, the values of the two classification indices decrease markedly as the percentage of Q-entry misspecification increases. When there is more than 10% of Q-entry misspecification, both classification indices fall to less than .80, which could adversely affect the quality of diagnostic results. Second, both types of attribute misspecification, attribute exclusion and attribute inclusion,

have negative effects on classification accuracy and consistency of diagnostic results, but vary with groups of different levels. Generally, classification accuracy and consistency for attribute exclusion is higher it is for attribute inclusion across various conditions. The values of pattern-level classification indices are generally high (>.80), while the values of attribute-level classification indices vary greatly, ranging from 0.49 to 1. In conclusion, Q-matrix misspecification has significant effects on the classification accuracy and consistency of diagnostic results. This is not surprising since the Q-matrix reflects the relationship between items and attributes in CDAs.

Results of this study also reveal that the two manipulated factors, attribute dependency and data generation model, have only trivial impacts on both pattern and attribute-level classification results. Another important finding is that when an unnecessary attribute is included in the Q-matrix, its attribute-level classification accuracy falls significantly lower than other attributes. For example, in C5-C8, the values of attribute-level classification accuracy of attribute 6 range from 0.01 to 0.18, and the attribute-level classification consistency is extremely high, the highest among all attributes. In C5-C8 the values of attribute-level classification consistency of attribute 6 range from 0.96 to 1. This indicates that the two classification indices may be useful to provide researchers a way to identify possible attribute misspecification (e.g., attribute inclusion) in empirical analyses.

This first study contributes to a better understanding of the effects of different types of Q-matrix misspecification on classification accuracy and consistency of diagnostic results, which can be used as a guide for researchers or practitioners who seek to design diagnostic assessments from a CDA framework and make diagnostic inferences for improved student learning. Based on our results, any misspecification of the Q-matrix, either Q-entry misspecification or attribute misspecification, can significantly deteriorate the classification accuracy and consistency of

diagnostic results. This may result in not only misinterpretation of students' skill profiles but also may adversely affect students' learning process. Thus, specifying the Q-matrix in the development of CDAs must be done with great care.

Based on this study, several directions for further research are suggested. First, the same Q-matrix setup can be used with other CDMs such as the NIDA and RUM models, or with more complex attribute dependencies, such as correlations between the attributes set to 0.7 or 0.8. Second, a more comprehensive simulation study could be conducted to investigate the impact of other factors on the performance of the two classification indices. For example, attribute structure (e.g., hierarchical structure) as well as the interplay among types of Q-matrix misspecification and Q-matrix design. Third, research studies that investigate how to identify and correct possible Q-matrix misspecification could be conducted to improve the quality of diagnostic results and guide diagnostic decision making for students' learning.

CHAPTER 4: COMPARISON OF THREE Q-MATRIX VALIDATION METHODS FOR COGNITIVE DIAGNOSTIC ASSESSMENT

Although the statistical consequences of misspecified Q-matrices have been recognized (as indicated in the previous chapter), only a few well-developed methods (e.g., Chiu, 2013; DeCarlo, 2012; de la Torre, 2008, Liu, Xu, and Ying, 2012) are available in the field of educational measurement to validate a misspecified Q-matrix for CDAs. In addition, all previous studies, to the best of my knowledge, on Q-matrix validation have focused on only a single method without any substantive comparative explanations about why a specific method was selected over the others. This is one of the major limitations in the current literature on Q-matrix validation in CDAs. No research has been conducted to systematically compare the performance of different Q-matrix validation methods. Hence, it appears not only a matter of curiosity but also of necessity to look into the methods that have been most used to evaluate their performance across various conditions.

This study investigates and compares the performance of three selected Q-matrix validation methods with both basic and complex assessment design factors under the CDA framework. The three methods include the sequential EM based δ-method (de la Torre, 2008), the Bayesian estimation method (DeCarlo, 2012), and the nonparametric Q-matrix refinement method (Chiu, 2013). The reasons for choosing these three methods are twofold. First, they are such methods that aim to refine or validate a specified Q-matrix in which the elements are pre-defined or some elements are assumed to be known. This scenario comprises the most common cases in educational assessments. Second, all three methods are introduced in the context of the DINA model, the most commonly used CDM for Q-matrix validation, which makes the comparison fair and reasonable. In addition to performance comparisons, the impact of different assessment design factors on their performance is also examined.

Method

The performance of three Q-matrix validation methods evaluated in this study can be grouped into two different sets: (a) performance with basic assessment design factors (Q-matrix misspecification rate, test length, number of attributes and sample size); and (b) performance with complex assessment design factors (attribute dependency, item parameter specification and the data generation model).

Simulation study 1: performance with basic assessment design factors

Simulation Design. Simulation study 1 was designed to compare the performance of the three Q-matrix validation methods with basic assessment design factors, including: a) Q-matrix misspecification rate (*QM*=10%, 20%, and 30%), b) number of attributes (*K*=3, 4 and 5), c) test length (*J*=20, 40 and 80), and d) sample size (*N*=500, 1000 and 2000).

The correct Q-matrices that correspond to tests of 20 items with 3, 4, and 5 attributes respectively are the same as those used in Chiu (2013), as shown in Table 8. For tests with 40 and 80 items, the corresponding Q-matrices were generated by doubling and quadrupling the correct Q-matrices for 20 items, respectively. Each Q-matrix is complete, containing at least one item devoted solely to each attribute. The correct Q-matrices are identifiable since all identifiability conditions, developed by Chen, Liu, Xue, and Ying in 2015, are satisfied. Three levels of sample sizes (*N*=500, 1000 and 2000) are used because the intention of the study is to maintain consistency with the previous studies (Chiu, 2013; DeCarlo, 2012; de la Torre, 2008) and to further examine their performance under various sample size conditions.

Table 8. Correct Q-matrices for tests of 20 items

Number of attributes											
3			4				5				
1	0	0	1	0	0	0	1	0	0	0	0
0	1	0	0	1	0	0	0	1	0	0	0
0	0	1	0	0	1	0	0	0	1	0	0
1	1	0	0	0	0	1	0	0	0	1	0
1	0	1	1	0	0	0	0	0	0	0	1
0	1	1	0	1	0	0	1	1	0	0	0
1	0	0	0	0	1	0	0	1	0	1	0
0	1	0	0	0	0	1	1	0	0	0	1
0	0	1	1	1	0	0	0	0	0	1	1
1	1	0	1	0	1	0	0	0	1	1	0
1	0	1	1	0	0	1	0	1	1	1	0
0	1	1	0	1	1	0	1	0	1	0	1
1	0	0	0	1	0	1	0	1	0	1	1
0	1	0	0	0	1	1	0	0	1	1	1
0	0	1	1	1	1	0	1	1	1	0	0
1	1	0	1	1	0	1	1	1	1	0	1
1	0	1	1	0	1	1	0	1	1	1	1
0	1	1	0	1	1	1	1	1	0	1	1
1	1	1	1	1	1	1	1	0	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1

In terms of the Q-matrix misspecification rate, de la Torre used about 3.4%, Chiu used 10% and 20%, and DeCarlo used about 20% as the proportion of misspecified Q-entries. As noted by Chiu (2013), a useful Q-matrix validation method should handle a much higher percentage of Q-matrix misspecification. Thus, we used 10%, 20%, and 30% of misspecified Q-entries to evaluate the effectiveness of the three methods. To generate the misspecified Q-matrices with different

percentages of misspecification, Q-entries in the correct Q-matrix were randomly selected at a specific probability (10%, 20% or 30%), and then replaced with opposite values.

Other factors were considered as fixed in simulation study 1. For example, the attribute dependency was set to be independent, the upper bound of both slipping and guessing parameters for all items was fixed at 0.2, and the DINA model was used to generate complete dichotomous responses. In sum, the simulation design yielded a total of 81 conditions: 3 (misspecification rate) × 3 (number of attributes) × 3 (test length) × 3 (sample size). Twenty datasets were generated for each condition.

Evaluation Criteria. Results were evaluated by two criteria: a) Q-matrix mean recovery rate (MRR) and b) the mean mis-recovery rate (MMR). MRR stands for the mean proportion of misspecified Q-entries that are successfully corrected using the Q-matrix validation method. MMR refers to the mean proportion of correct Q-entries in the original Q-matrix that are wrongly identified in the validation procedure. Note that in order to calculate MMR for the Bayesian estimation method, we double the number of uncertain Q-entries and treat them as random. That is, MMRs for 10%, 20% and 30% of misspecified Q-entries were calculated based on 20%, 40% and 60% of uncertain Q-entries, respectively.

Results. Because the results from different sample sizes (N = 500, 1000 and 2000) are quite similar, only results associated with a sample size 500 are presented and discussed here.

Table 9 presents the MRRs of three Q-matrix validation methods across basic assessment design factors. On the whole, results indicate that when the data generation model is the DINA model, the attributes are independent, and the upper bound of the slipping and guessing parameters is 0.2, the Bayesian estimation method achieves the best performance among the three Q-validation methods. Specifically, perfect MRRs are obtained across nearly all conditions with only one

exception. That is, in the condition of 20 items and 5 attributes, MRRs go down to 98% when the misspecification rate increases to 30%. Other factors such as misspecification rate and test length have very limited impact on Q-matrix recovery with the Bayesian estimation method.

The nonparametric Q-matrix refinement method performs the second best. Perfect or nearly perfect MRRs are obtained when the number of attributes is four or fewer, regardless of the number of items and misspecification rate. But MRRs decrease abruptly when there are fewer items, a larger number of attributes and a higher misspecification rate. For example, when there are 20 items and 5 attributes, MRRs go down from 97.61% to 59.83% as the misspecification rate increases from 20% to 30%. MRRs of the nonparametric Q-matrix refinement method are influenced slightly by test length. They are slightly higher in the condition of 40 items than that of 20 or 80 items, regardless of the number of attributes and misspecification rate.

Unlike the other two methods, MRRs of the sequential EM-based δ-method varies across conditions, ranging from 34.76% to 100%. They are still comparable to those of the other two methods when there are three attributes in the Q-matrix. But once the misspecification rate increases to 20% and 30%, MRRs decrease quickly, especially when there are more attributes and fewer items. In the condition of 20 items and 4 attributes, average MRRs decrease from 94.38% to 54.58% as the misspecification rate goes up from 10% to 30%. When the number of attributes reaches 5, average MRRs decrease sharply to 57%, 35 % and 47.83% for misspecification rates of 10%, 20% and 30%, respectively. However, adding more items can significantly improve the MRRs in the condition of a high misspecification rate and a relatively large number of attributes. For instance, in the condition of 5 attributes and 30% misspecification, MRRs increase from 47.83% (J=20) to 67.50% (J=40), and 74.71% (J=80).

Table 10 shows the MMRs of the three Q-matrix validation methods across basic assessment design factors. In general, MMRs of both the nonparametric refinement method and the Bayesian estimation method remain at a very low level, even close to zero, across all conditions. However, their MMRs increase slightly when there are fewer items, a larger number of attributes, and a higher misspecification rate. For example, in the condition of 20 items and 5 attributes, MMRs go up from 0 to 2.43% with the nonparametric refinement method and from 0% to 2.33% with the Bayesian estimation method, as the misspecification rate increases from 10% to 30%. Adding more items could offset this increase of MMRs.

On the other hand, MMRs of the sequential EM based δ-method change frequently and fall across a wide range from 0% to 15.93%. The number of attributes and misspecification rate significantly affect the performance of the sequential EM based δ-method. MMRs go up quickly as the number of attributes increases or the misspecification rate increases. For instance, in the condition of 20 items and 10% Q-entry misspecification, the corresponding MMRs are 0.09%, 0.83% and 13.94% for 3, 4 and 5 attributes, respectively. And in the condition of 40 items and 5 attributes, MMRs increase from 5.39% to 11.46% as the misspecification rate increases from 10% to 30%. Similarly, adding more items leads to lower MMRs, regardless of other factors.

Summary. In sum, the three Q-matrix validation methods perform differently across various basic assessment design factors. In terms of MRRs, the Bayesian estimation method outperforms the other two methods under almost all conditions with nearly perfect MRRs. The nonparametric Q-matrix refinement method performs as well as the Bayesian estimation method in most conditions, except for conditions that have fewer items, a larger number of attributes and higher misspecification rates. The performance of the sequential EM based δ-method is only comparable to the other two methods in limited conditions (3 attributes and less than 30%

misspecification). MRRs of all other conditions are significantly lower than the other two methods. Although the performance of the three methods is quite different, similar conclusions are found. As seen in Table 9, the sample size has minimal impact, but the number of attributes, test length and misspecification rate have various degrees of influence on the MRRs of the three methods. Specifically, the presence of a higher misspecification rate, a larger number of attributes and fewer items in a Q-matrix can result in lower MRRs.

In terms of MMRs, the performance of the nonparametric Q-matrix refinement method and the Bayesian estimation method are similar, and they both outperform the sequential EM based δ-method method across all conditions. All other assessment design factors have a small effect on the MMRs of both methods with only one exception (5 attributes, 20 items and 30% misspecification rate). However, their effects on MMRs of the sequential EM based δ-method method are different and large. MMRs change quickly as the number of items decreases, and the number of attributes and misspecification rate increase.

Table 9. MRRs of Three Q-matrix Validation Methods with Basic Assessment Design Factors

Sample Size & Number of Items		Number of Attributes	Sequential EM based δ-method			Nonparametric Q-matrix refinement method			Bayesian estimation method		
			QM=10%	*QM*=20%	*QM*=30%	*QM*=10 %	*QM*=20 %	*QM*=30%	*QM*=10 %	*QM*=20 %	*QM*=30 %
N=500	*J*=20	*K*=3	100.00	100.00	90.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	94.38	76.88	54.58	100.00	100.00	100.00	100.00	100.00	97.71
		K=5	57.00	35.00	47.83	100.00	97.61	59.83	100.00	100.00	98.00
	J=40	*K*=3	100.00	100.00	99.86	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	98.67	100.00	97.50	100.00	100.00	100.00	100.00	100.00	100.00
		K=5	91.00	77.75	67.50	100.00	99.75	99.08	99.75	99.88	99.58
	J=80	*K*=3	99.78	99.79	99.93	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	100.00	99.13	99.79	99.06	100.00	100.00	99.84	100.00	99.95
		K=5	96.13	94.56	74.71	99.75	96.56	98.25	99.75	100.00	99.88
N=1000	*J*=20	*K*=3	100.00	100.00	88.06	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	93.75	72.50	57.08	100.00	100.00	100.00	100.00	100.00	100.00
		K=5	65.00	35.24	44.00	100.00	99.78	52.83	100.00	100.00	100.00
	J=40	*K*=3	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	99.33	100.00	99.27	100.00	100.00	100.00	100.00	100.00	100.00
		K=5	96.50	84.50	72.17	100.00	100.00	99.83	100.00	100.00	100.00
	J=80	*K*=3	99.78	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	100.00	99.92	99.89	100.00	100.00	100.00	100.00	100.00	100.00
		K=5	97.88	98.31	78.88	100.00	99.19	99.96	100.00	100.00	100.00
N=2000	*J*=20	*K*=3	100.00	100.00	84.44	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	92.50	77.81	56.88	100.00	100.00	100.00	100.00	100.00	100.00
		K=5	64.00	34.76	45.50	100.00	100.00	66.67	100.00	100.00	100.00
	J=40	*K*=3	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	100.00	100.00	99.79	100.00	100.00	100.00	100.00	100.00	100.00
		K=5	96.75	82.63	77.00	100.00	100.00	100.00	100.00	100.00	100.00
	J=80	*K*=3	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	100.00	99.76	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=5	98.63	99.00	83.54	100.00	99.94	100.00	100.00	100.00	100.00

Table 10. MMRs of Three Q-matrix Validation Methods with Basic Assessment Design Factors

Sample Size & Number of Items		Number of Attributes	Sequential EM based δ-method			Nonparametric Q-matrix refinement method			Bayesian estimation method		
			QM=10%	*QM*=20%	*QM*=30%	*QM*=10%	*QM*=20%	*QM*=30%	*QM*=10%	*QM*=20%	*QM*=30%
N=500	*J*=20	*K*=3	0.09	0.00	0.71	0.00	0.00	0.00	0.00	0.00	0.00
		K=4	0.83	2.11	8.84	0.00	0.00	0.00	0.00	0.00	4.79
		K=5	13.94	16.90	15.93	0.00	0.32	2.43	0.00	0.00	2.33
	J=40	*K*=3	0.19	0.21	0.12	0.00	0.00	0.00	0.00	0.00	0.00
		K=4	0.41	0.55	0.71	0.00	0.00	0.00	0.00	0.00	0.00
		K=5	5.39	6.38	11.46	0.06	0.06	0.04	0.00	0.00	0.58
	J=80	*K*=3	0.09	0.08	0.12	0.00	0.00	0.00	0.00	0.00	0.00
		K=4	0.40	0.25	0.51	0.00	0.00	0.00	0.00	0.00	0.00
		K=5	4.38	5.27	11.45	0.03	0.06	0.07	0.00	0.06	0.04
N=1000	*J*=20	*K*=3	0.00	0.00	1.31	0.00	0.00	0.00	0.00	0.00	0.00
		K=4	0.07	1.64	7.77	0.00	0.00	0.00	0.00	0.00	0.00
		K=5	8.44	14.62	12.71	0.00	0.00	0.43	0.00	0.00	0.00
	J=40	*K*=3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		K=4	0.24	0.43	0.31	0.00	0.00	0.00	0.00	0.00	0.00
		K=5	2.14	4.34	6.21	0.00	0.00	0.00	0.00	0.00	0.00
	J=80	*K*=3	0.00	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00
		K=4	0.12	0.12	0.18	0.00	0.00	0.00	0.00	0.00	0.00
		K=5	1.68	1.84	6.13	0.00	0.00	0.00	0.00	0.00	0.00
N=2000	*J*=20	*K*=3	0.00	0.00	1.31	0.00	0.00	0.00	0.00	0.00	0.28
		K=4	0.00	0.86	7.05	0.00	0.00	0.00	0.00	0.00	0.00
		K=5	7.72	15.57	13.36	0.00	0.00	1.43	0.48	0.00	0.00
	J=40	*K*=3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		K=4	0.03	0.12	0.18	0.00	0.00	0.00	0.00	0.00	0.00
		K=5	1.08	4.06	4.04	0.00	0.00	0.00	0.00	0.00	0.08
	J=80	*K*=3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		K=4	0.07	0.02	0.11	0.00	0.00	0.00	0.00	0.00	0.00
		K=5	0.89	1.06	5.11	0.00	0.00	0.00	0.00	0.00	0.00

Simulation study 2: Performance with complex assessment design factors

Simulation design. Simulation study 2 compares the performance of the three Q-matrix validation methods with complex assessment design factors using the following manipulated factors: a) Q-matrix misspecification rate (QM=10%, 20% and 30%), b) attribute dependency (AD= independent and correlated), c) item parameter specification (upper bounds= 0.2, 0.3 and 0.4), and d) data generation model (DG= DINA and DINO). To better understand the impact of complex assessment design factors, the number of attributes is fixed at 5, test length at 40 items and sample size at 2000.

Specifically, three levels of Q-entry misspecification rate are consistent with the design of simulation study 1. And two levels of attribute dependency were considered: independent and correlated. That is, a total of 2000 vectors were generated from a multivariate normal distribution (i.e., $\theta \sim MVN(0, \rho)$). Here ρ represents a correlation matrix with equal off-diagonal elements. The off-diagonal elements are either all 0 or all 0.5 (Henson & Douglas, 2005), representing independent and correlated dependency, respectively. Thus, the examinees' attribute pattern $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iK})$ is simulated by equation (13).

In terms of item parameter specification, de la Torre (2008) set both slipping and guessing parameters at 0.2 for all items as the ideal condition; Chiu (2013) used four levels of item parameter specification ranging from 0.2 to 0.5 as the upper bounds for both parameters; and DeCarlo (2012) did not investigate the impact of different item parameter specifications. To make a fair comparison of the three methods, the upper bounds of the slipping and guessing parameters are set to 0.2, 0.3 and 0.4.

In terms of the data generation model, the majority of existing research has adopted the conjunctive CDMs to generate examinee response data. For example, the simulated examinee item responses used in Chiu (2013), DeCarlo (2012), and de la Torre's (2008) research were all generated to conform to the DINA model or the NIDA model, both of which are conjunctive CDMs. To investigate the performance of the three methods in a broader context, both the DINA and DINO models were used as data generation models.

A full cross design yields a total of 36 conditions: 3 (misspecification rate) × 2 (attribute dependency) × 3 (item parameter specification) × 2 (data generation model). Twenty datasets were generated for each of the 36 conditions. The Q-matrix validation procedures were implemented using R-CDM (Robitzsch, Kiefer, George, & Uenlue, 2015), R-NPCD (Zheng & Chiu, 2015), and R- R2OpenBUGS packages (Sturtz, Ligges, and Gelman, 2015), respectively. All other codes were written by the author.

Evaluation Criteria. Results were evaluated by the same evaluation criteria: MRRs and MMRs. For the Bayesian estimation method, the number of uncertain Q-entries was doubled and treated as random to calculate its MMRs.

Results. Table 11 presents the MRRs of three Q-matrix validation methods across complex design factors when the number of attributes is fixed at 5, test length at 40 items and sample size at 2000. On the whole, the Bayesian estimation method performs best, the nonparametric Q-matrix refinement method performs second, and the sequential EM based δ-method is the poorest. One noticeable finding is that the three methods do not differ significantly in their performance recovering a misspecified Q-matrix, and none of them have satisfactory results when the DINO model is adopted as the data generation model. More details regarding these points are discussed in the following paragraphs.

The Bayesian estimation method produces the highest MRRs across almost all conditions, but its performance is significantly better when using the DINA model as the data generation model than when using the DINO model for data generation. First, when using the DINA model for data generation, perfect MRRs are obtained when attributes are independent, regardless of the upper bound of item parameters or misspecification rate. But if attribute dependency is changed from independent to correlated, MRRs slightly decrease as the upper bound of item parameters and misspecification rate increase. For example, when attributes are correlated and the upper bound is 0.2, MRRs fall from 100% to 94.08% as the misspecification rate increases from 10% to 30%. Similarly, when attributes are correlated and the misspecification rate is 20%, MRRs fall from 97.13% to 92.13% as the upper bound increases from 0.2 to 0.4. However, when the DINO model instead of the DINA model is used for data generation, MRRs decrease substantially across all conditions, ranging from 39.75% to 65%. In such a case, all other assessment design factors have only limited impact on the MRRs associated with the Bayesian estimation method.

Compared to the Bayesian estimation method, MRRs of the nonparametric Q-matrix refinement method vary across conditions. On one hand, when the DINA model is used for data generation and attributes are independent, perfect MRRs are obtained across all conditions as long as the upper bound does not exceed 0.4. If the upper bound increases to 0.4, MRRs fall to 91.50%, 95.13% and 82.50% for misspecification rates of 10%, 20% and 30%, respectively. When attributes are correlated, however, MRRs decrease quickly as the upper bound and misspecification rate increase. For example, 98.75% of the misspecified Q-entries are successfully recovered when the misspecification rate is 10%

and the upper bound equals 0.2. However, as the upper bound increases to 0.4, only 58% of the misspecified Q-entries are successfully recovered; and as the misspecification rate increases to 30%, MRR drops as low as 51.33%, indicating that only about half of the misspecified Q-entries can be recovered via the nonparametric Q-matrix refinement method. On the other hand, when the DINO model is used for data generation, the nonparametric Q-matrix refinement method, similar to the Bayesian estimation method, could not effectively recover misspecified Q-entries for any condition.

The performance of the sequential EM based δ-method is not as good as the other two methods. It is largely affected by complex assessment design factors such as the upper bound of item parameters and misspecification rate. With data generated by the DINA model, MRRs decrease abruptly, especially for conditions with higher misspecification rates and larger upper bounds. For example, when attributes are independent and the upper bound equals 0.2, MRRs decrease sharply from 94.50% to 55.08% as the misspecification rate goes up from 10% to 30%. Once the upper bound increases to 0.4, MRRs further decrease to 58.50%, 54.75% and 45.25% for misspecification rates of 10%, 20% and 30%, respectively. Similar patterns are found when attributes are correlated. That is, attribute dependency only minimally affects the performance of the sequential EM based δ-method with close MRRs. Moreover, the sequential EM based δ-method yields the lowest MRRs, not too surprisingly, when data is generated by the DINO model. In such a case, less than half of the misspecified Q-entries could be recovered, regardless of all other complex assessment design factors, including the upper bound of item parameter, attribute dependency and misspecification rate.

Table 12 shows the MMRs of the three Q-matrix validation methods across complex assessment design factors when the number of attributes is fixed at 5, test length at 40 items and sample size at 2000. These results are also presented according to the data generation model used. First, in the condition using independent attributes and the DINA model for data generation, both the nonparametric refinement method and the Bayesian estimation method perform excellently with MMRs under nearly all conditions. Once attributes are considered as correlated, their MMRs are affected by the upper bound of item parameters and the misspecification rate. For instance, when attributes are correlated and the upper bound equals to 0.2, MMRs increase from 0.81% to 3% with the nonparametric refinement method, and from 0.50% to 2% with the Bayesian estimation method as the misspecification rate increases from 10% to 30%. And when attributes are correlated and the misspecification rate is 20%, MMRs increase from 1.59% to 5.53% with the nonparametric refinement method and from 0.88% to 6.05% with the Bayesian estimation method as the upper bound increases from 0.2 to 0.4. MMRs of the sequential EM based δ-method, however, change relatively dramatically compared with MMRs of the other two methods. As shown in Table 5, the upper bound of item parameters and misspecification rate have a large impact on the performance of the sequential EM based δ-method in terms of MMRs. The presence of a higher misspecification rate and larger upper bound of item parameters results in higher MMRs. For example, when attributes are independent and the upper bound equals to 0.2, MMRs of the sequential EM based δ-method apparently increase from 0.94% to 16.11% as the misspecification rate increases from 10% to 30%. Meanwhile, if the upper bound increases to 0.4, MRRs further increase to 4.92%, 10.88% and 13.5% for misspecification rates of 10%, 20% and 30%, respectively. Similarly,

attribute dependency only minimally affects MMRs of the sequential EM-based δ-method. Secondly, when the DINO model is used for data generation, MMRs increase rapidly for all three Q-matrix validation methods. Almost a third to a half of the correct Q-entries in the original Q-matrix could be wrongly identified in validation procedures. Other assessment design factors, including the upper bounds of item parameters, attribute dependency, and misspecification rates have little influence on the values of MMRs on the whole.

Summary. An important phenomenon observed from simulation study 2 is that when data is generated from the DINO model, the performance of all three methods generally deteriorates greatly, regardless of other assessment design factors. That is, when relying on a disjunctive CDM (such as the DINO model) as the data generation model, all three major Q-matrix validation methods can hardly recovery the misspecified Q-entries, as reflected by the low MRRs, and have a much greater chance of wrongly identifying correct Q-entries in the validation procedure, as reflected by the high MMRs.

However, as long as the DINA model is used for data generation, the most common case in the field of educational research, the Bayesian estimation method performs the best among the three methods reviewed. Specifically, the Bayesian estimation method produces the highest MRRs across all conditions. Its performance remains at a very robust level when attributes are independent, as reflected by the perfect MRRs. MRRs of the Bayesian estimation method show a slight drop with correlated attributes, ranging from 90.08% to 100%, but still are the highest among the three methods. The performance of the nonparametric Q-matrix refinement method is only comparable to the Bayesian estimation method when the upper bound does not exceed 0.3, attributes are independent, and the

misspecification rate does not exceed 10%. Once the upper bound reaches 0.4 or the misspecification rate reaches 20% or above, MRRs of the nonparametric Q-matrix refinement method decrease quickly, especially for the condition of correlated attributes and 30% misspecification rate. Only 51.33% of the misspecified Q-entries are successfully recovered. The performance of the sequential EM based δ-method is obviously less accurate than the other two methods and declines sharply as the upper bound and misspecification rate increase. Such decline becomes more serious when attributes are correlated. In addition, both the Bayesian estimation method and the nonparametric Q-matrix refinement method produce rather low MMRs across all conditions ranging from 0% to 6.75%. Attribute dependency has a noticeable negative influence on their performance. By contrast, MMRs of the sequential EM based δ-method vary greatly, ranging from 0.94% to 14.86%. In general, higher misspecification rates and larger upper bounds of item parameters yield higher MMRs, but the effect of attribute dependency does not show consistent patterns across different factors.

Table 11. MRRs of Three Q-matrix Validation Methods with Complex Assessment Design Factors

Data Generation Model & Attribute Dependency		Upper Bound (U)	Sequential EM based δ-method			Nonparametric Q-matrix refinement method			Bayesian estimation method		
			QM=10%	*QM*=20%	*QM*=30%	*QM*=10%	*QM*=20%	*QM*=30%	*QM*=10%	*QM*=20%	*QM*=30%
DINA	Independent	0.2	94.50	87.50	55.08	100.00	100.00	100.00	100.00	100.00	100.00
		0.3	89.50	77.25	57.08	100.00	100.00	100.00	100.00	100.00	100.00
		0.4	58.50	54.75	45.25	91.50	95.13	82.50	100.00	100.00	100.00
	Correlated	0.2	86.50	84.88	54.50	98.75	95.88	88.50	100.00	97.13	94.08
		0.3	78.75	81.00	65.17	93.25	93.63	85.67	99.50	97.50	92.83
		0.4	65.75	56.75	48.83	58.00	66.63	51.33	99.50	92.13	90.08
DINO	Independent	0.2	45.50	54.63	43.17	55.75	54.50	53.00	48.50	61.38	55.92
		0.3	45.25	54.13	41.92	56.25	55.50	51.25	48.00	62.25	57.67
		0.4	42.00	48.00	42.67	54.75	54.63	48.17	64.00	65.00	61.58
	Correlated	0.2	45.00	29.13	43.50	41.00	69.63	59.00	42.50	61.25	60.75
		0.3	38.75	39.00	39.25	43.25	69.63	61.50	39.75	61.63	54.67
		0.4	43.50	32.75	33.58	49.00	63.50	48.33	42.00	56.38	57.17

Table 12. MMRs of Three Q-matrix Validation Methods with Complex Assessment Design Factors

Data Generation Model & Attribute Dependency		Upper Bound (U)	Sequential EM based δ-method			Nonparametric Q-matrix refinement method			Bayesian estimation method		
			QM=10%	*QM*=20%	*QM*=30%	*QM*=10%	*QM*=20%	*QM*=30%	*QM*=10%	*QM*=20%	*QM*=30%
DINA	Independent	0.2	0.94	4.44	16.11	0.00	0.00	0.00	0.00	0.00	0.00
		0.3	2.03	6.50	12.64	0.00	0.00	0.00	0.00	0.00	0.00
		0.4	4.92	10.88	13.50	0.08	0.38	0.64	0.25	0.00	0.00
	Correlated	0.2	2.56	3.19	14.86	0.81	1.59	3.00	0.50	0.88	2.00
		0.3	3.94	4.69	6.82	1.83	2.81	4.61	2.25	2.88	1.83
		0.4	6.58	8.75	9.21	3.31	5.53	5.00	5.75	6.50	6.75
DINO	Independent	0.2	30.58	30.34	30.39	40.50	38.56	34.11	29.75	32.75	33.92
		0.3	28.47	29.56	32.46	39.67	36.75	35.00	27.75	31.25	32.17
		0.4	27.67	28.97	32.64	36.31	35.19	39.54	31.25	34.88	31.58
	Correlated	0.2	35.25	29.09	29.68	32.00	35.13	38.18	40.50	41.00	44.33
		0.3	33.44	27.06	25.61	31.33	34.91	35.54	39.50	48.13	51.83
		0.4	26.50	24.59	20.86	32.14	34.84	44.04	33.75	42.00	51.08

Conclusion

The second research study focused on a crucial issue in successful implementation of CDAs: *Among three most commonly used Q-matrix validation methods, which method achieves the best performance validating a misspecified Q-matrix? Is their performance affected by different assessment design factors?* To provide insight into these questions, the performance of the three Q-matrix validation methods, including the sequential EM based δ-method, the Bayesian estimation method, and the nonparametric Q-matrix refinement method, was investigated and compared with both basic assessment design factors (e.g., Q-matrix misspecification rate, test length, number of attributes, and sample size) and complex assessment design factors (e.g., attribute dependency, item parameter specification, and data generation model). Results of two simulation studies reveal that the Bayesian estimation method outperforms the other two methods in terms of both MRRs and MMRs, as long as the DINA model is used for data generation. Specifically, its performance maintains at a good level with the highest MRRs and lowest MMRs across almost all conditions. The presence of correlated attributes has negative influence on the performance of Bayesian estimation method, but it still performs much better than the others. Next, the nonparametric Q-matrix refinement method could perform as well as the Bayesian estimation method in most conditions. But it may not work well for cases with high misspecification rates and a short test, or with high upper bound of item parameters and correlated attributes. Under these conditions only about 50%~60% of misspecified Q-entries could be identified and corrected. In this case, there are risks in applying the nonparametric Q-matrix refinement method in the implementation of CDAs. Lastly, the sequential EM based δ-method, in general, lags behind the other two methods in validating a misspecified Q-matrix. Its performance is only comparable to the others in very limited

conditions when there are fairly few attributes (K=3) and a low misspecification rate (QM=10%).

Results also show that the performance of the three Q-matrix validation methods is affected to different degrees by various assessment design factors, among which the data generation model is the most critical. When a disjunctive CDM, such as the DINO model, is used for data generation, none of the methods effectively identify and correct misspecified Q-entries, as reflected by very low MRRs and high MMRs, regardless of other assessment design factors. This finding is not a surprise, since all three methods are introduced in the context of the DINA model, which is the most widely used model in cognitive diagnosis. Furthermore, although sample size has no obvious impact, the number of attributes, test length, and misspecification rate have different degrees of influence on the performance of the three methods. In particular, their performance degrades when there are more attributes, fewer items and a higher misspecification rate. The presence of correlated attributes also has a negative effect on their performance, and the size of this effect increases as the upper bound of item parameters and misspecification rate increase.

The contribution of this study is to provide a systematic performance evaluation of the three core Q-matrix validation methods based on a methodological perspective and on metrics that allow meaningful comparisons. Based on our findings, among the three methods, the Bayesian estimation method achieves the best performance under various conditions, and its performance was remarkably consistent, regardless of different assessment design factors. However, this method has two major limitations: 1) it requires the possible misspecified Q-matrix to be identified in advance; and 2) it requires a complex computational process, and the computing cost is typically high, especially when the

number of attributes or the number of items is large. Meanwhile, the nonparametric Q-matrix refinement method can perform almost as well as the Bayesian estimation method except for conditions with high misspecification rates in a short test or with high upper bound of item parameter and correlated attributes. Finally, the application of the sequential EM based δ-method has a limited range. It only works well for the condition of few attributes (K=3), a small number of Q-entry misspecifications (QM<20%), and independent attributes. Findings of this study provide useful information and practical guidance for educational researchers when they do further research in Q-matrix validation and cognitive diagnosis.

Future directions that might be informative are also provided. The first direction is to improve the current Q-matrix validation methods either in terms of validation accuracy or computational efficiency, or both. To make the Q-matrix validation methods more useful in practice, it would also be important to explore how to improve the accuracy of Q-matrix validation when data follow the disjunctive CDMs instead.

CHAPTER 5: A TWO-STAGE CROSS-VALIDATION METHOD FOR COGNITIVE DIAGNOSTIC ASSESSMENT

Based on the comparison results of the three most commonly used Q-matrix validation methods in Chapter 4, the Bayesian estimation method achieves the best performance among the three methods in terms of both MRRs and MMRs across almost all conditions. However, two of the most important shortcomings hinder its further development in practical applications: 1) the possible misspecified Q-entries in a provisional Q-matrix are required to be identified in advance, and 2) the prohibitively long computation time makes it computationally infeasible when the number of attributes or items is large. On the other hand, the nonparametric Q-matrix refinement method is computationally fairly fast and performs almost as well as the Bayesian estimation method in most conditions, except for the conditions with high misspecification rates in a short test, or with high upper bounds of item parameters and correlated attributes. This study proposes a two-stage cross-validation method, which incorporates the idea of minimizing the RSS based on the weighted Hamming distance and Bayesian estimation techniques, to improve Q-matrix validation accuracy and computational efficiency, and to work for complex conditions.

Method

At the first stage of the two-stage cross-validation method, the nonparametric Q-matrix refinement method was applied to an expert-defined Q-matrix Q^0 that might be misspecified in such a way that a refined Q-matrix Q^1 was obtained by minimizing the overall RSS computed from the observed response and the ideal responses to each test item. At the second stage, the refined Q-matrix Q^1 was compared with the expert-defined Q-

matrix Q^0, and the inconsistent Q-entries were identified as possible misspecified Q-entries. These were treated as random Bernoulli variables and estimated simultaneously with other parameters using an MCMC estimation algorithm. The posterior distributions from Bayesian estimation were then used to determine whether a possible misspecified Q-entry should be one or zero. An optimized Q-matrix Q^2 was obtained after all possible misspecfied Q-entries had been examined. The performance of the proposed method for validating Q-matrix was evaluated in both simulation and empirical data settings.

Simulation Study Design

Two simulation studies using settings similar to those of the second research study were conducted to provide comparable results. Because there is no expert-defined Q-matrix in simulation studies, Q-entries in the correct Q-matrix were randomly misspecified at specific probabilities (i.e., 10%, 20% and 30%), and then the misspecified Q-matrices were considered as expert-defined Q-matrices.

The first simulation study investigates the performance of the proposed method with basic assessment design factors: a) Q-matrix misspecification rate (*QM*=10%, 20%, 30%), b) number of attributes (*K*=3, 4 and 5), c) test length (*J*=20, 40 and 80), and d) sample size (*N*=500, 1000 and 2000). Other factors are considered as fixed: attribute independence; upper bounds of both slipping and guessing parameters for all items are fixed at 0.2; and the DINA model is adopted as the data generation model.

The second simulation study investigates the performance of the proposed method with complex assessment design factors: a) Q-matrix misspecification rate (QM=10%, 20% and 30%), b) attribute dependency (AD= independent and correlated), and c) item parameter specification (upper bounds=0.2, 0.3 and 0.4). The Q-matrix used for data

generation consisted of 40 items and 5 attributes, sample size was set to 2000 examinees, and the DINA model was adopted as the data generation model.

Empirical Study Design

In addition to simulated data, the proposed two-stage cross-validation method was also applied to two real datasets to evaluate its effectiveness. The first data set, a short version of the fraction subtraction data (Tatsuoka, 1990), consists of the binary responses of 536 middle school students to 15 fraction subtraction items that measure five attributes specified by experts. The five attributes are: (1) performing basic fraction-subtraction operation, (2) simplifying/reducing, (3) separating whole number from fraction, (4) borrowing one from whole number to fraction, and (5) converting whole number to fraction. The expert-designed Q-matrix of the fraction subtraction data is shown in Table 13.

The second data set is from the Test of Practical Chinese (CTEST), which is administered to non-native Chinese speakers as a second language test. The dataset includes the binary responses of 857 participants to 30 items of the 2007 CTEST reading section. All 30 items are dichotomously scored multiple-choice items. Four reading attributes are defined by subject-matter experts to fulfill the requirements of the assessment. These four attributes are: (1) selective attention, (2) semantic comprehension, (3) synthesis and organization of information, and (4) logical inference. Table 14 presents the expert-designed Q-matrix of the 2007 CTEST reading section.

Table 13: Expert-designed Q-matrix for the Fraction-Subtraction Data

No.	Item	A1	A2	A3	A4	A5
1	$\frac{3}{4}-\frac{3}{8}$	1	0	0	0	0
2	$3\frac{1}{2}-2\frac{3}{2}$	1	1	1	1	0
3	$\frac{6}{7}-\frac{4}{7}$	1	0	0	0	0
4	$3-2\frac{1}{5}$	1	1	1	1	1
5	$3\frac{7}{8}-2$	0	0	1	0	0
6	$4\frac{4}{12}-2\frac{7}{12}$	1	1	1	1	0
7	$4\frac{1}{3}-2\frac{2}{3}$	1	1	1	1	0
8	$\frac{11}{8}-\frac{1}{8}$	1	1	0	0	0
9	$3\frac{4}{5}-3\frac{2}{5}$	1	0	1	0	0
10	$2-\frac{1}{3}$	1	0	1	1	1
11	$4\frac{5}{7}-1\frac{4}{7}$	1	0	1	0	0
12	$7\frac{3}{5}-\frac{4}{5}$	1	0	1	1	0
13	$4\frac{1}{10}-2\frac{8}{10}$	1	1	1	1	0
14	$4-1\frac{4}{3}$	1	1	1	1	1
15	$4\frac{1}{3}-1\frac{5}{3}$	1	1	1	1	0

The performance of the proposed method in these empirical data settings was evaluated using a cross-validation sample to determine the adequacy of the optimized Q-matrix. Specifically, the original data sets were each partitioned into two equal sized subsamples. One was used for obtaining an optimized Q-matrix, and the other for testing the optimized Q-matrix against the expert-defined Q-matrix via evaluation criteria.

Evaluation Criteria

In terms of the two simulation studies, the MRR was considered as the evaluation criteria to measure the performance of the proposed method under each condition. And for the empirical studies, the optimized Q-matrix and expert-designed Q-matrix were compared in terms of their goodness of fit (e.g., the Log-like, AIC and BIC), classification accuracy and consistency (Cui, Gierl, Chang, 2012).

Table 14. Expert-designed Q-matrix of 2007 CTEST data

Item No.	A1	A2	A3	A4	Item No.	A1	A2	A3	A4
1	1	1	0	1	16	0	0	1	1
2	0	0	1	1	17	0	1	0	0
3	0	1	1	0	18	1	1	1	0
4	1	1	0	0	19	1	0	0	0
5	0	1	0	0	20	0	1	0	0
6	1	1	0	0	21	1	0	0	0
7	0	1	0	0	22	0	1	1	1
8	0	0	1	1	23	0	1	1	0
9	1	1	0	0	24	0	1	1	0
10	0	1	0	0	25	0	1	0	0
11	0	0	1	0	26	1	1	1	0
12	0	1	1	0	27	0	1	1	0
13	1	1	1	0	28	0	0	1	1
14	0	0	1	1	29	0	1	1	0
15	0	1	0	0	30	0	1	0	0

Results

The results from this research study are presented as two parts: the first compares the performance of the two-stage cross-validation method with the three most commonly used Q-matrix validation methods in two simulation studies; the second investigates the effectiveness of the proposed two-stage cross-validation method in two empirical studies.

Simulation Studies

The performance of the four Q-matrix validation methods in terms of MRRs with basic design factors is summarized in Table 15. Results indicate that the two-stage cross-validation method, on the whole, outperforms the other three Q-matrix validation methods if other factors (i.e., sample size, test length, number of attributes and data generation model) are held consistent. Specifically, perfect MRRs are obtained across almost all 81 conditions, except for only one condition when the sample size is 500, the number of items is 80, the number of attributes is 5 and misspecification rate is 30%. In this case the MRR slightly decreases to 99.89% but is still higher than the other three methods. Results also reveal that the basic assessment design factors, including sample size, test length, number of attributes and misspecification rate, have very limited impact on the performance of the proposed two-stage cross-validation method with perfect or nearly perfect MRRs for all conditions.

The performance of the four Q-matrix validation methods in terms of MRRs with complex assessment design factors is summarized in Table 16. It is shown that the two-stage cross-validation method is still the one that performs the best among these four methods, as reflected by the highest MRRs, when other factors are held consistent. Specifically, perfect MRRs are obtained when attributes are independent regardless of the upper bound of item parameters or the misspecification rate. Even when attributes are

correlated, the two-stage cross-validation method still produces perfect MRRs as long as the misspecification rate does not exceed 30% or the upper bound does not exceed 0.3. However, once the misspecification rate reaches 30%, MMRs from the two-stage cross-validation method degrade as the upper bound increases. For example, for the condition in which the attributes are correlated and the misspecification rate is 30%, the MMR goes down from 98.34% to 92.71% as the upper bound increases from 0.2 to 0.4. To sum up, the performance of the two-stage cross-validation method is slightly affected when there are correlated attributes, high misspecification rates and greater upper bounds of item parameters.

Table 15: MRRs of Four Q-matrix Validation Methods with Basic Assessment Design Factors

Sample Size & Number of Items		K	Sequential EM based δ -method			Nonparametric Q-matrix refinement method			Bayesian estimation method			Two-stage cross-validation method		
			10%	20%	30%	10%	20%	30%	10%	20%	30%	10%	20%	30%
N=500	*J*=20	*K*=3	100.00	100.00	90.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	94.38	76.88	54.58	100.00	100.00	100.00	100.00	100.00	97.71	100.00	100.00	100.00
		K=5	57.00	35.00	47.83	100.00	97.61	59.83	100.00	100.00	98.00	100.00	100.00	100.00
	J=40	*K*=3	100.00	100.00	99.86	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	98.67	100.00	97.50	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=5	91.00	77.75	67.50	100.00	99.75	99.08	99.75	99.88	99.58	100.00	100.00	100.00
	J=80	*K*=3	99.78	99.79	99.93	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	100.00	99.13	99.79	99.06	100.00	100.00	99.84	100.00	99.95	100.00	100.00	100.00
		K=5	96.13	94.56	74.71	99.75	96.56	98.25	99.75	100.00	99.88	100.00	100.00	99.89
N=1000	*J*=20	*K*=3	100.00	100.00	88.06	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	93.75	72.50	57.08	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=5	65.00	35.24	44.00	100.00	99.78	52.83	100.00	100.00	100.00	100.00	100.00	100.00
	J=40	*K*=3	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	99.33	100.00	99.27	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=5	96.50	84.50	72.17	100.00	100.00	99.83	100.00	100.00	100.00	100.00	100.00	100.00
	J=80	*K*=3	99.78	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	100.00	99.92	99.89	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=5	97.88	98.31	78.88	100.00	99.19	99.96	100.00	100.00	100.00	100.00	100.00	100.00
N=2000	*J*=20	*K*=3	100.00	100.00	84.44	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	92.50	77.81	56.88	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=5	64.00	34.76	45.50	100.00	100.00	66.67	100.00	100.00	100.00	100.00	100.00	100.00
	J=40	*K*=3	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	100.00	100.00	99.79	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=5	96.75	82.63	77.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	J=80	*K*=3	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=4	100.00	99.76	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		K=5	98.63	99.00	83.54	100.00	99.94	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 16: MMRs of Four Q-matrix Validation Methods with Complex Assessment Design Factors

Attribute Dependency	Upper Bund (U)	Sequential EM based δ -method			Nonparametric Q-matrix refinement method			Bayesian estimation method			Two-stage cross-validation method		
		10%	20%	30%	10%	20%	30%	10%	20%	30%	10%	20%	30%
Independent	0.2	94.50	87.50	55.08	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	0.3	89.50	77.25	57.08	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	0.4	58.50	54.75	45.25	91.50	95.13	82.50	100.00	100.00	100.00	100.00	100.00	100.00
Correlated	0.2	86.50	84.88	54.50	98.75	95.88	88.50	100.00	97.13	94.08	100.00	100.00	98.34
	0.3	78.75	81.00	65.17	93.25	93.63	85.67	99.50	97.50	92.83	100.00	100.00	95.23
	0.4	65.75	56.75	48.83	58.00	66.63	51.33	99.50	92.13	90.08	100.00	97.95	92.71

Empirical Studies

Tables 17 and 18 report the optimized Q-matrices generated by the two-stage cross-validation method for the Fraction and CTEST datasets, respectively. The proportions of consistency of Q-entry between the optimized and expert-designed Q-matrices are 97.3% for the Fraction dataset and 55% for the CTEST dataset. To determine the adequacy of the optimized Q-matrices, a cross-validation sample was used to evaluate their performance in terms of the goodness of fit, classification accuracy and consistency.

Table 19 summarizes the goodness of fit statistics for the Fraction data, including the Log-like, the AIC and the BIC, which indicate how well a Q-matrix fits the data. Results show that the Log-like values of the expert-designed Q-matrix and the optimized Q-matrix are -1752.15 and -1728.05, the AIC values are 3596.3 and 3548.11, and the BIC values are 3761.49 and 3713.29, respectively. The optimized Q-matrix generated by the two-stage cross-validation method appears to fit the fraction data better than the expert-design Q-matrix, as reflected by the larger Log-like value and smaller AIC and BIC values. Moreover, the classification accuracy and consistency statistics summarized in Table 20 also favor the optimized Q-matrix. Specifically, the classification accuracy index for the optimized Q-matrix is 0.679 compared with 0.623 for the expert-designed Q-matrix. Meanwhile, the classification consistency index for the optimized Q-matrix is 0.791 while for the expert-designed Q-matrix the index is 0.691. The optimized Q-matrix yields both higher classification accuracy and consistency than the expert-designed Q-matrix.

Similar to the Fraction dataset, Table 21 summarizes the goodness of fit statistics of the CTEST dataset. It is shown that the optimized Q-matrix fits the CTEST dataset better than the expert-designed Q-matrix due to its larger Log-like value and smaller AIC and BIC values. Specifically, the Log-like value of the optimized Q-matrix is -7843.41, which

is about 10 higher than the expert-designed Q-matrix. And the AIC and BIC values of the optimized Q-matrix are 15828.82 and 16117.18, about 20 lower than the expert-designed Q-matrix, respectively. What's more, the classification accuracy and consistency statistics shown in Table 22 also indicate that the optimized Q-matrix proposed by the two-stage cross-validation method outperforms the expert-designed Q-matrix in terms of classification results. In particular, the classification accuracy and consistency indices of the optimized Q-matrix are 0.530 and 0,669, which are 0.108 and 0.027 higher than those values for the expert-designed Q-matrix, respectively.

Table 17: Optimized Q-matrix for the Fraction-Subtraction Data

No.	Item	A1	A2	A3	A4	A5
1	$\frac{3}{4}-\frac{3}{8}$	1	0	0	1	0
2	$3\frac{1}{2}-2\frac{3}{2}$	1	1	1	1	0
3	$\frac{6}{7}-\frac{4}{7}$	1	0	0	0	0
4	$3-2\frac{1}{5}$	1	0	1	1	1
5	$3\frac{7}{8}-2$	0	0	1	0	0
6	$4\frac{4}{12}-2\frac{7}{12}$	1	1	1	1	0
7	$4\frac{1}{3}-2\frac{2}{3}$	1	1	1	1	0
8	$\frac{11}{8}-\frac{1}{8}$	1	1	0	0	0
9	$3\frac{4}{5}-3\frac{2}{5}$	1	0	1	0	0
10	$2-\frac{1}{3}$	1	0	1	1	1
11	$4\frac{5}{7}-1\frac{4}{7}$	1	0	1	0	0
12	$7\frac{3}{5}-\frac{4}{5}$	1	1	1	1	0
13	$4\frac{1}{10}-2\frac{8}{10}$	1	1	1	1	0
14	$4-1\frac{4}{3}$	1	1	1	1	1
15	$4\frac{1}{3}-1\frac{5}{3}$	1	1	1	1	0

Table 18: Optimized Q-matrix of 2007 CTEST data

Item No.	A1	A2	A3	A4	Item No.	A1	A2	A3	A4
1	1	1	0	0	16	1	0	1	1
2	1	1	1	1	17	1	1	0	0
3	1	1	0	1	18	1	1	1	0
4	1	1	0	0	19	0	1	0	0
5	1	1	0	0	20	0	1	1	0
6	0	0	1	1	21	1	0	0	0
7	1	0	1	1	22	0	1	0	1
8	1	1	0	0	23	0	1	0	1
9	1	0	1	1	24	0	1	0	1
10	1	1	0	0	25	0	0	1	1
11	0	1	0	0	26	1	1	0	0
12	1	1	1	1	27	1	1	0	0
13	0	1	0	0	28	1	1	1	1
14	1	1	1	1	29	1	1	0	1
15	1	1	0	0	30	1	1	0	0

Table 19: Goodness of Fit Summary of the Fraction Data

Q-matrix	Npara	Device	Log-like	AIC	BIC	Mean of RMSEA item fit
Expert	46	3504.30	-1752.15	3596.30	3761.49	0.10
Optimized	46	3456.11	-1728.05	3548.11	3713.29	0.09

Table 20: Summary of Classification Accuracy and Consistency for Fraction Data

Q-matrix	Classification Accuracy (P_a)	Classification Consistency (P_c)
Expert	0.623	0.679
Optimized	0.691	0.791

Note. Classification accuracy and consistency indices are provided by Cui, Gierl, Chang (2012)

Table 21: Goodness of Fit Statistics Summary for CTEST Data

Q-matrix	Npara	Device	Log-like	AIC	BIC	Mean of RMSEA item fit
Expert	71	15706.37	-7853.18	15848.37	16136.73	0.04
Optimized	71	15686.82	-7843.41	15828.82	16117.18	0.03

Table 22: Summary of Classification Accuracy and Consistency for CTEST Data

Q-matrix	Classification Accuracy (P_a)	Classification Consistency (P_c)
Expert	0.422	0.642
Optimized	0.530	0.669

Note. Classification accuracy and consistency indices are provided by Cui, Gierl, Chang (2012)

Summary and Discussion

By means of simulation and empirical studies, the third research study addresses these questions: *Compared to the three most commonly used Q-matrix validation methods, does the proposed two-stage cross-validation method identify and correct misspecified Q-entries more accurately and efficiently under a wide range of conditions? Does it still work well in empirical data settings?*

In general, the findings demonstrate the effectiveness of the proposed two-stage cross-validation method in both simulation and empirical data settings. On one hand, the two-stage cross-validation method performs more accurately and efficiently than the other three methods across all simulation conditions when the data generation model is the DINA model. Specifically, when the attributes are independent, and the upper bound of item parameter is fixed at 0.2, the two-stage cross-validation method can recover 100% of the misspecified Q-entries, regardless of sample size, test length, number of attributes and misspecification rate. However, when attributes are correlated, the performance of the two-stage cross-validation method is slightly affected by an increase in the upper bound of item parameter or the misspecification rate, but still is the best among the four methods.

The two-stage cross-validation method also demonstrates its effectiveness in two empirical studies. In particular, the optimized Q-matrix generated by the two-stage cross-validation method is evaluated and compared with the expert-designed Q-matrix in terms of its goodness of fit, classification accuracy and consistency. The results of the two empirical studies show that the optimized Q-matrix fits the data better, with larger maximized Log-likelihood values and smaller AIC and BIC values than the expert-designed Q-matrix. It is also shown that the optimized Q-matrix yields better classification

results than the expert-designed Q-matrix, as reflected by better classification accuracy and consistency results.

Although the proposed two-stage cross-validation method demonstrates its effectiveness in both simulation and empirical data settings, it is always helpful to consult with subject-matter experts to better understand the optimized Q-matrix, because some of the discrepancies between the optimized Q-matrix and the expert-designed Q-matrix do not make sense from the perspective of some experts. For example, in the Fraction data study, the first item, $\frac{3}{4} - \frac{3}{8}$, requires the fourth attribute (borrowing one from whole number to fraction) to answer it correctly in the optimized Q-matrix, but according to some subject-matter experts, this item only requires the first attribute (performing the basic fraction-subtraction operation).

It should be noted that the two-stage cross-validation method is a data based method designed to improve an existing Q-matrix, not to replace the current Q-matrix construction approaches. As a matter of fact, there is no single best way to ensure a sound Q-matrix. Depending either solely on statistical evidence or on practical evidence is insufficient, and more research is needed to provide complementary resources for Q-matrix optimization.

One possible future research direction is to combine the two-stage cross-validation method with traditional methods (such as the think-aloud verbal protocol approach) to cross-examine whether the item-attribute assignment is necessary and sufficient to account for the major attributes required for a CDA. Another direction would be to compare the performance of the optimized Q-matrix generated by the two-stage cross-validation method with the refined/estimated Q-matrix generated by other methods (e.g., Chen, Liu, Xu, and Ying, 2015) to further validate its effectiveness. And, since the proposed two-stage

cross-validation method is limited in its capability to provide useful statistical information for Q-matrix optimization by the use of the DINA model for data generation, more studies should be conducted to determine how to validate a misspecified Q-matrix in a broader context in order to conform to different CDMs, especially disjunctive CDMs.

CHAPTER 6: CONCLUSION AND FUTURE RESEARCH

The No Child Left Behind Act of 2001 emphasizes that assessments should "produce individual student interpretive, descriptive and diagnostic reports that include information regarding achievement on the academic assessments measured against the state's student academic achievement standards, which will help parents, teachers and principals to understand and address the specific academic needs of students" [section 1111(b) (3) (c) (xii)]. Motivated by this call for more formative assessments, there has been a high demand for CDAs to identify individual students' academic strengths and weaknesses in specific learning areas. The primary purpose of this dissertation is to investigate the practical issues of optimizing the Q-matrix in CDA applications.

Incorrect specification of the Q-matrix leads to undesirable statistical consequences such as poor model fit and inaccurate model parameter estimation. Despite its importance, there have been few studies about how Q-matrix misspecification affect the accuracy and consistency of classification results in CDAs. For this reason, the first study conducts a comprehensive simulation study to investigate the degree to which the classification accuracy and consistency of diagnostic results are affected by two types of Q-matrix misspecification: (1) Q-entry misspecification, and (2) attribute misspecification. Results indicate that any misspecification in the Q-matrix, either Q-entry misspecification or attribute misspecification, could significantly degrade to various degrees the classification accuracy and consistency of the diagnostic results. The two manipulated factors, attribute dependency and data generation model, had very limited influence on classification accuracy and consistency. Another important finding from this study is that the two classification indices can be used for identifying possible attribute misspecification in

empirical analyses, especially attribute inclusion, due to the extremely low attribute-level classification accuracy values associated with it.

To address the issue of Q-matrix misspecification, various Q-matrix optimization methods have been proposed by researchers in both psychometrics and educational data mining (Barnes, 2003,2010; Chiu, 2013; DeCarlo, 2012; de la Torre, 2008; Desmarais, 2011; Desmarais, Beheshti, & Naceur, 2012; Desmarais, & Naceur, 2013; Liu, Xu, & Ying, 2012, 2013; Templin & Henson, 2006a). However, these previous works are limited to showing the feasibility of their methods in validating an existing Q-matrix or in reconstructing a Q-matrix in very restricted contexts. The second study fills this critical gap in the literature by providing meaningful and fair performance assessments of the three most commonly used Q-matrix validation methods. Specifically, performance of the sequential EM based δ-method, the Bayesian estimation method, and the nonparametric Q-matrix refinement method is assessed and compared with both basic assessment design factors (e.g., Q-matrix misspecification rate, test length, number of attributes, and sample size) and complex assessment design factors (e.g., attribute dependency, item parameter specification, and data generation model). Results of the two simulation studies show that among the three methods, the Bayesian estimation method achieves the best performance on correcting the misspecified Q-entries across all conditions, as long as the DINA model is used as the data generation model. Meanwhile, although the nonparametric Q-matrix refinement method performs as well as the Bayesian estimation method in most conditions, its performance degrades badly for conditions under which the Q-matrix misspecification rate is high (30% or above) in a short test (20 items or less), or the upper bound of item parameter is fairly high (0.4 or above) while attributes are correlated. Under these

conditions, only about 50%~60% of misspecified Q-entries could be identified and corrected. The sequential EM based δ-method had very limited capability in recovering the misspecified Q-matrix. That method only works well for cases in which attributes are independent, the number of attributes is small (3 or less), and misspecification rate is fairly low (20% or less). Another critical finding from this second study is that the choice of data generation model can significantly affect the performance of all three methods. Once data is not generated using the conjunctive CDMs but rather using a disjunctive DINA model, none of the models could accurately and effectively validate the misspecified Q-entries with acceptable MRRs and MMRs. This is unsurprising because all three of the Q-matrix validation methods are defined in the context of the DINA model.

Following the second study, a two-stage cross-validation method is proposed to improve Q-matrix validation accuracy and computation efficiency for several complex conditions. Thus, the third study explores the performance in both simulation and empirical data settings of the two-stage cross-validation method, which incorporates the idea of minimizing the RSS based on the weighted Hamming distance and the Bayesian estimation technique. Results of the two simulation studies show that the two-stage cross-validation method performs more accurately and efficiently than the other three methods, as reflected by the highest MRRs and lowest MMRs under various combinations of different levels of both basic and complex assessment design factors. In addition, results of the two empirical studies also demonstrate the effectiveness of the two-stage cross-validation method in empirical data settings. Specifically, the optimized Q-matrix generated by the two-stage cross-validation method outperformed the expert-designed Q-matrix in terms of the goodness of fit, and the classification accuracy and consistency. But we should also be

aware that subject-matter experts are still needed to help better understand and interpret the optimized Q-matrix.

Q-matrix validation issues have not been thoroughly addressed in the literature, and a variety of studies could be conducted in the future. First, it is important to conduct a more comprehensive simulation study to investigate the impacts of complex factors on the classification accuracy and consistency indices. For example, in the first study, only two levels of attribute dependency are considered: independent and correlated. However, in practice, the attributes in a Q-matrix can have a hierarchical structure, and its impact should be further investigated. It would also be beneficial to explore how the interplay among types of Q-matrix misspecification influences classification accuracy and consistency. Secondly, the Q-matrix validation methods studied in this dissertation are all defined in the context of the DINA model. Once the examinee data does not conform to the DINA model, say to the DINO model instead, those methods are unable to recover a misspecified Q-matrix with fair accuracy and efficiency. In practice, it is possible that attributes in a Q-matrix are disjunctive in nature. Thus, future research on investigating how to validate the Q-matrix with a wider class of DCMs is necessary. Lastly, in this dissertation, the Q-matrix is developed after attributes are well defined; that is, attributes are assumed known due to comparison with a true and validated Q-matrix. However, it is also possible that the attribute dimension is not available in real situations, and the meanings of attributes are unclear. Therefore, future research on how to determine the number of attributes and how to interpret the Q-matrix accurately is also needed.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Aryadoust, V. (2011). Cognitive diagnostic assessment as an alternative measurement model. *SHIKEN: JALT Testing & Evaluation SIG Newsletter, 15(1),* 2-6.

Baker, F. B. (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement, 17*(3), 201-210.

Barnes, T.M. (2003). *The Q-matrix Method of Fault-tolerant Teaching in Knowledge Assessment and Data Mining* (Doctoral Dissertation). North Carolina State University.

Barnes, T. (2010). Novel derivation and application of skill matrices: The q-matrix method. In C. Ramero, S. Vemtora, M. Pechemizkiy, & R. S. J. de Baker (Eds.), *Handbook of educational data mining* (pp.159-172). Boca Raton, FL: Chapman & Hall.

Chen, Y., Liu, J., Xu, G. and Ying, Z. (2015). Statistical Analysis of Q-matrix Based Diagnostic Classification Models. *Journal of the American Statistical Association, 110,* 850-866.

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

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

Cui, Y., Gierl, M. J., & Chang, H. 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(1),* 8-26.

DeCarlo, T. L. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36(6),* 447-468.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69(3),* 333–353.

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

Desmarais, M. C. (2011). Mapping question items to skills with non-negative matrix factorization. *ACM SIGKDD Explorations Newsletter*, *13*(2), 30-36.

Desmarais, M. C., Beheshti, B., & Naceur, R. (2012). Item to skills mapping: deriving a conjunctive q-matrix from data. In *Intelligent tutoring systems* (pp. 454-463). Springer Berlin Heidelberg.

Desmarais, M. C., & Naceur, R. (2013). A matrix factorization method for mapping items to skills and for enhancing expert-based q-matrices. In *Artificial Intelligence in Education* (pp. 441-450). Springer Berlin Heidelberg.

DiBello, L. V., & Stout, W. F. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement, 44(4),* 285–291.

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

Doornik, J. A. (2003). *Object-oriented matrix programming using Ox (version 3.1)* [Computer software]. London: Timberlake Consultants Press.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175–186.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta psychologica, 37(6),* 359-374.

Fu, J., & Li, Y. (2007, April). An integrative review of cognitively diagnostic psychometric models. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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

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

Henson, R.A., Templin, J.L., & Willse, J.T. (2009). Defining a Family of Cognitive Diagnosis Models Using Log-Linear Models with Latent Variables. *Psychometrika, 74(2),* 191-210.

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research & Evaluation, 15(3),* 1-7.

Im, S. and Corter, J. E. (2011). Statistical consequences of attribute misspecification in the rule space method. *Educational and Psychological Measurement, 71(4),* 712-731.

Jang, E. E. (2009). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills, *Language Assessment Quarterly, 6(3),* 210-238.

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.

Kim, Y. (2011). Diagnosing EAP writing ability using the Reduced Reparameterized Unified Model. *Language Testing, 28(4),* 509-541.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The Impact of Model Misspecification on Parameter Estimation and Item-Fit Assessment in Log-Linear Diagnostic Classification Models. *Journal of Educational Measurement, 49(1),* 59-81.

Lee, D. D., & Seung, H. S. (2001). Algorithms for nonnegative matrix factorization. *In Advances in Neural Information Processing Systems, 13,* 556–562.

Lee, Y. W., & Sawaki, Y. (2009). Cognitive diagnostic approaches to language assessment: An overview. *Language Assessment Quarterly, 6(3),* 172-189.

Leighton, J.P., & Gierl, M.J. (2007). Verbal reports as data for cognitive diagnostic assessment. In J.P. Leighton & M.J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 146-172). Cambridge, MA: Cambridge University Press.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36(7),* 548-564.

Liu, J., Xu, G., & Ying, Z. (2013) Theory of self-learning Q-matrix. *Bernoulli, 19,* 1790-1817.

Liu, H., You, X., Wang, W., Ding, S. & Chang, H-H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification, 30,* 152-172.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25(3),* 258-272.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2015). CDM: Cognitive diagnosis modeling (R package version 4.6-0) [Computer software]. http://CRAN.Rproject.org/package=CDM.

Roussos, L. A, DiBello, L. V., Stout, W. F., Hartz, S. M., Henson, R., & Templin, J. (2007). The fusion model skills diagnosis system. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment in education* (pp. 275-318). New York, NY: Cambridge University Press.

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

Rupp, A., Templin J., & Henson R. A. (2010). Diagnostic measurement: Theory, methods, and applications. New York, NY: Guilford Press.

Sawaki, Y., Kim, H. J., & Gentile, C. (2009). Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessment. *Language Assessment Quarterly, 6(3),* 190-209.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. *Cognitively diagnostic assessment*, 327-359.

Templin, J. (2006). CDM: cognitive diagnosis modelling with mplus user guide. Unpublished manuscript.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30(2),* 251–275

Templin, J., & Henson, R. (2006a). A Bayesian method for incorporating uncertainty into Q-matrix estimation in skills assessment. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

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

von Davier, M. (2005). A general diagnostic model applied to language testing data. ETS Research Report: RR-05-16, Educational Testing Service, Princeton, NJ.

Zheng, Y., & Chiu, C.-Y. (2015). NPCD: Nonparametric methods for cognitive diagnosis (R package version 1.0-9) [Computer software]. http://CRAN.Rproject.org/package=NPCD