NEW Q-MATRIX VALIDATION PROCEDURES

BY RAGIP TERZI

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Education

Written under the direction of

Jimmy de la Torre

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2017

ABSTRACT OF THE DISSERTATION

New Q-Matrix Validation Procedures

by Ragip Terzi

Dissertation Director: Jimmy de la Torre

The primary purpose of cognitively diagnostic assessment (CDA) is to provide useful information about students' learning needs. The attributes (i.e., latent skills) possessed by examinees can be uncovered based on examinee responses to test items primarily in conjunction with cognitive diagnosis models (CDMs). Most, if not all, CDMs require a Q-matrix to specify the attributes measured by each item. When attributes are correctly specified, CDMs have been shown useful in identifying examinees' mastery or nonmastery of attributes in a domain of interest. However, conventional Q-matrix development process involves some degree of subjectivity, which can result in validity concerns due to inaccurate attribute specifications. Although some statistical procedures exist in the literature, additional work is still needed to address some concerns about validating attribute specifications in the Q-matrix.

Each of the three studies of this dissertation introduces new Q-matrix validation procedures. The first study presents an EM-based δ-method, namely, the *iterative modified sequential search algorithm* (IMSSA), to empirically validate the correctness of attribute specifications for the deterministic inputs, noisy "and" gate (DINA)

model. In this study, the performance of the IMSSA is compared to that of some existing parametric and nonparametric methods through simulated and real data analyses. The second study proposes new indices under the generalized DINA (G-DINA) model, namely, the *iterative* Jensen-Shannon Divergence (iJSD) index and *iterative* G-DINA model discrimination index (iGDI), to determine the correctness of attribute specifications in the Q-matrix. The iJSD is more general than the iGDI that can be applied under both dichotomous and nondichotomous models, whereas, the iGDI can only be used under dichotomous models. As with the iJSD, the main advantage of the iGDI is the inclusion of an iterative algorithm in the original GDI so that better results can be obtained. The feasibility of the iJSD and iGDI is investigated using simulated and real data. In the final study, the Wald-Q, an adaptation of the Wald statistical test to the Q-matrix validation context, is presented. The Wald-Q is applied under situations where the true underlying process is known or unknown. Using simulated and real data, the Wald-Q was compared to the IMSSA proposed in the first study and to iGDI proposed in the second study in conjunction with the DINA and G-DINA models, respectively.

Across the three simulation studies, different factors (i.e., sample sizes, test lengths, complexity of q-vectors, degrees of q-vector misspecifications, attribute structures, and item qualities) are varied to examine the performance of the new procedures. The new procedures are further applied to fraction-subtraction data. Practical applications of the proposed procedures can lead to the advancement of the use of CDAs in educational settings. Results leading to improvements in Q-matrix validation can also help other components of cognitive diagnosis modeling, such as the estimation of model parameters, model-data fit analyses, the accuracy of attribute classifications, and ultimately, validity of CDA inferences.

List of Tables

1.1. A Q-Matrix Sample for Fraction-Subtraction Test 3

2.1. Examples for Over- and Under-Specifications 15

2.2. True Q-matrix for the Simulated Data 18

2.3. Proportions of Recovery Comparisons 20

2.4. Proportions of Recovery Comparisons between Methods with and without Iterative Algorithm ($s = g = 0.2$) 21

2.5. Q-Matrix for Fraction-Subtraction Items 24

2.6. Single-attribute specifications and the corresponding $\delta-$values and $-$ratios 25

2.7. Suggested Q-Matrix by the IMSSA and QRM 26

3.1. True Q-matrix for the Simulated Data 46

3.2. False-Positive Rate of the iJSD and iGDI 48

3.3. True-Positive Rate of the iJSD and iGDI 49

3.4. Q-Matrix for Fraction-Subtraction Items 51

3.5. Suggested Q-Matrix by the iJSD and iGDI for the G-DINA model . . 52

4.1. Number of Parameters (i.e., Latent Groups) for Different Q-Vectors when $K = 3$ (DINA Model Assumed) 70

4.2. Restriction Matrix for the Q-Vector $(1,1,0)'$ ($K_j = 2$) (DINA Model Assumed) 71

4.3. Number of Parameters (i.e., Latent Groups) for Different Q-Vectors when $K = 3$ (No Model Assumed) 72

4.4. Restriction Matrix for the Q-Vector $(1,1,0)'$ ($K_j = 2$) (No Model Assumed) 72
4.5. Q-matrix for the Simulated Data 74
4.6. False-Positive Rates of the Wald-Q and IMSSA (DINA Model Assumed) 76
4.7. True-Positive Rate of the Wald-Q and IMSSA (DINA Model Assumed) 77
4.8. False-Positive Rate of the Wald-Q and iGDI (No Model Assumed) . . 78
4.9. True-Positive Rate of the Wald-Q and iGDI (No Model Assumed) . . 79
4.10. Q-Matrix for Fraction-Subtraction Items 80
4.11. Suggested Q-Matrix by the Wald-Q and IMSSA (DINA Model Assumed) 81
4.12. Suggested Q-Matrix by the Wald-Q and iGDI (No Model Assumed) . 82

Acknowledgements

First, I would like to express my deepest appreciation to my committee chair, my advisor, Dr. Jimmy de la Torre, who continually showed a spirit of professional adventure in his outstanding scholarship and mentoring. Without his persistent guidance and unwavering help, this dissertation would not have been possible. I hope to become a productive researcher and advisor like him.

Besides my advisor, I would like to thank my committee members, Dr. Youngsuk Suh, Dr. Chia-Yi Chiu, and Dr. Min-ge Xie, for their insightful comments, constructive criticisms, encouragement, and time.

I also would like to thank the Ministry of National Education of Turkey for the fellowship that provided me with such a great opportunity to learn from remarkable people in the field.

My wonderful fellows, Charlie Iaconangelo, Lokman Akbay, Mehmet Kaplan, Nathan Minchen, and Wenchao Ma, also deserve special thanks for their academic and personal friendship; I am lucky to have them in my life.

Last but not least, I would like to give special thanks to my parents, Nejat and Ayse, my sister, Senel, and my lovely son, Akif, for the joy, love, support, and sacrifices they have shown me.

Dedication

This dissertation is dedicated to my well-beloved love – Dr. Nuray Kivanc-Terzi

Table of Contents

Abstract ii

List of Tables iv

Acknowledgements vi

Dedication vii

1. Introduction and Objectives 1

1.1. Introduction 1

1.2. Objectives 4

1.3. References 7

2. Study I: An Iterative Method for Empirically-Based Q-Matrix Validation 9

2.1. Abstract 9

2.2. Introduction 10

2.3. Background 13

2.3.1. The DINA Model 13

2.3.2. An Iterative Method for Empirically-Based Q-Matrix Validation 14

2.4. Design and Analysis 17

2.4.1. Results 19

2.5. Implementation with Real Data 23

2.5.1. Results 23

2.6. Summary and Discussion 28

2.7. References 32

3. Study II: The Iterative Jensen-Shannon Divergence Index and Iterative GDI for Q-Matrix Validation 34

3.1. Abstract 34

3.2. Introduction 35

3.3. Background 39

3.3.1. The G-DINA Model 39

3.3.2. The G-DINA Model Discrimination Index 40

3.3.3. The Jensen-Shannon Divergence Index 41

3.3.4. The iJSD for Q-Matrix Validation 43

3.4. Design and Analysis 45

3.4.1. Results 46

3.5. Implementation with Real Data 50

3.5.1. Results 50

3.6. Summary and Discussion 53

3.7. References 55

4. Study III: The Wald Test for Empirical Q-Matrix Validation 58

4.1. Abstract 58

4.2. Introduction 59

4.3. Background 64

4.3.1. G-DINA and DINA Models 64

4.3.2. The G-DINA Model Discrimination Index 65

4.3.3. The Wald Test 66

4.3.4. The Wald Test for Q-Matrix Validation 68

4.4. Design and Analysis 73

4.4.1. Results 74

4.5. Implementation with Real Data 79

4.5.1. Results 80

4.6. Summary and Discussion 83

4.7. References 85

5. Summary 87

5.1. References 91

Chapter 1
Introduction and Objectives

1.1 Introduction

Traditional summative assessments describe examinees' proficiency as a unidimensional latent construct using single scores. The scores reported from summative assessments are based on traditional test theories, such as classical test theory (CTT) and item response theory (IRT). A main criticism of a single overall ability score is its inability to provide diagnostic information about specific skills in a target domain in which examinees should have proficiency (e.g., Leighton & Gierl, 2007). Because of this limitation, there has been increasing interest in cognitively diagnostic assessment (CDA), which partitions the latent construct into finer-grained and interrelated, but separable latent skills. Therefore, CDA is employed to uncover examinees' current knowledge, skill sets, and capabilities within a domain of interest. In this way, areas that may need specific academic support while learning is occurring can be identified (de la Torre, 2009). By providing specific feedback to teachers and students based on examinees' strengths and weaknesses, teaching and learning activities can be arranged accordingly (DiBello & Stout, 2007).

To date, a variety of CDAs have been introduced. For example, a subset of attributes used in Tatsuoka's (1984) widely-used fraction-subtraction data contain the following four attributes: (a) performing a basic fraction subtraction operation, (b) simplifying/reducing, (c) separating a whole number from fraction, and (d) borrowing one from whole numbers to fraction. Recently, another CDA has been designed for

middle school proportional reasoning content (Tjoe & de la Torre, 2014). A subset of the attributes defined in this domain are: (1) prerequisite skills and concepts required; (2a) comparing and (2b) ordering fractions; and (3) identifying a multiplicative. As indicated by the examples, the application of CDA can enrich the learning environment such that examinees can be classified based on their mastery or nonmastery of more specific skills.

For optimal results, CDAs need to be analyzed using certain cognitive diagnosis models (CDMs; de la Torre & Minchen, 2014). CDMs are psychometric models developed primarily for the purpose of identifying multiple finer-grained skills in a particular content area. Because of specific assumptions required for the relationship between task performance and attribute vectors (Junker & Sijtsma, 2001), a number of models have been developed. These CDMs can be classified into two types of models: reduced models that require fewer parameters (e.g., only slip and guessing parameters for the DINA model) and saturated models with more flexible parameterizations.

Regardless of the various assumptions underlying a particular model, CDMs have a common component called Q-matrix. The Q-matrix relates a subset of K attributes to each one of the J items. The row j and column k binary entry of the Q-matrix, q_{jk}, is 1 only if the kth attribute is required to correctly answer item j, or is 0 if the kth attribute is not. Each attribute vector can be considered a unique latent class among 2^K possible latent classes for K binary attributes. For example, a sample of three items from fraction-subtraction data can be seen in Table 1.1 (Tatsuoka, 1984). An examinee i who is classified into the latent class α_i = (1,0,1,0) has mastered the first and third attributes. That is, the student can perform "basic fraction subtraction operation" and "separate a whole number from fraction." A further outcome of this assessment could suggest teachers to focus on the other two attributes (i.e., simplifying/reducing and borrowing one from whole numbers to fraction) for the student in order to successfully become proficient in the fraction-subtraction domain.

Table 1.1: A Q-Matrix Sample for Fraction-Subtraction Test

Item		Attributes			
		α_1	α_2	α_3	α_4
1.	$\frac{3}{4}-\frac{3}{8}$	1	0	0	0
2.	$3\frac{7}{8}-2$	1	0	1	0
3.	$\frac{11}{8}-\frac{1}{8}$	1	1	0	0

Note. α_1 – performing a basic fraction subtraction operation; α_2 – simplifying/reducing; α_3 – separating whole number from fraction; and α_4 – borrowing one from whole numbers to fraction.

Nevertheless, the process of constructing the Q-matrix requires enormous effort especially from domain experts (Akbay, Terzi, Kaplan, & Karaaslan, 2017; Tatsuoka, 1984; Tjoe & de la Torre, 2014). Identifying required attributes via an extensive literature review, writing test items based on the identified attributes, and validating the attribute specifications by domain experts and statistical analyses are main challenges in the Q-matrix establishment. Conventional Q-matrix construction, which primarily relies on expert judgments, can be considered as an inherently subjective process. As such, this process can ultimately lead to attribute misclassifications as a result of the inaccurate estimation of model parameters (de la Torre & Chiu, 2016). The validity of inferences from CDAs due to attribute misclassifications has raised discussions among researchers (e.g., Chiu, 2013; de la Torre, 2008; Liu, Xu, & Ying, 2012; Rupp & Templin, 2008). Hence, any model-fit analysis without validating the attribute specifications may not be reliable. Unfortunately, it is usually assumed that the Q-matrix is correct once it has been specified by domain experts. This assumption is generally made due to the lack of well-established methods for Q-matrix validation (Chiu, 2013; DeCarlo, 2011; de la Torre, 2008; de la Torre & Chiu, 2016; Henson, Templin, & Willse, 2009; Liu et al., 2012).

1.2 Objectives

It was only recently that validity concerns about CDAs due to the involvement of human judgments in constructing the Q-matrix have received considerable attention from a number of researchers (e.g., Chiu, 2013; DeCarlo, 2011; de la Torre, 2008; de la Torre & Chiu, 2016; Henson et al., 2009; Liu et al., 2012). Even though existing methods have proposed Q-matrix validation procedures from various perspectives, there are still some remaining issues that need to be addressed. This dissertation has the following primary objectives: (1) to introduce an *iterative modified sequential* EM-based δ-method for the DINA model, (2) to present the *iterative* Jensen-Shannon Diverge (iJSD) index and the *iterative* G-DINA model discrimination index (iGDI) for the G-DINA model, and (3) to propose a Wald-based procedure for Q-matrix validation under the DINA and G-DINA models.

The first study proposed a new search algorithm based on the sequential EM-based δ-method (de la Torre, 2008), the *iterative modified sequential search algorithm* (IMSSA), to empirically validate attribute specifications. The IMSSA is an extension of the *sequential search algorithm* (SSA; de la Torre, 2008) that some of the limitations were addressed. First, the IMSSA can offer a more efficient procedure in that only the first K single-attribute q-vectors are examined so as to lessen complications associated with multiple search steps. Second, in the SSA, it was not clear what ε to use in practice because it could change under many conditions, such as changes in sample size, test length, item quality, and degree of misspecifications, all of which were fixed in de la Torre's (2008) study. In the IMSSA, the ε values were determined as a function of observable variables based on each level of estimated item qualities (i.e., high, medium, and low). Therefore, these values are more applicable across specific set of conditions. Third, the SSA was not implemented iteratively in that the validation procedure stops after one full iteration even if changes occur in the provisional Q-matrix. The algorithm

in the IMSSA was implemented iteratively, such that, if the attribute specifications in a q-vector are changed in the previous iteration, the algorithm is repeated using the updated Q-matrix as the provisional Q-matrix. In doing so, the iterative algorithm can alleviate negative effects of any misspecified attribute specifications given in the preceding iteration. In this study, the effectiveness of the IMSSA was compared to other iterative and noniterative Q-matrix validation methods.

The second study introduced two new indices for empirically-based Q-matrix validation purposes to verify the correctness of attribute specifications in the Q-matrix. In particular, the iJSD and iGDI were proposed under a wider class of CDMs so that assuming an underlying process would not be required. The main advantage of the iJSD is its applicability for dichotomous, as well as nondichotomous models, such as the continuous G-DINA model (Minchen & de la Torre, 2016) and MC-DINA model (de la Torre, 2009; Yigit, Sorrel, & de la Torre, 2016). Because the iGDI can be applied for dichotomous models only, the iJSD is more general than the iGDI. As with the iJSD, the main advantage of the iGDI is the inclusion of an iterative algorithm that could provide better results.

The third study presented a new Q-matrix validation procedure, namely the Wald-Q, for identifying attribute specifications for each q-entry by adapting the Wald test (Morrison, 1967) for multivariate hypothesis testing. The Wald-Q should eliminate some of the limitations of the existing methods. For example, although the method was an empirically-based procedure similar to that used by de la Torre (2008), Terzi and de la Torre (2015), and de la Torre and Chiu (2016), the calculation of a single optimal ε value was not required. Moreover, an iterative process was included in the Wald-Q so that any possible impact of misspecified entries can be eliminated at the succeeding steps. Furthermore, the Wald-Q can be designed for reduced and general CDMs based on the restriction matrix. Using simulated and real data, the performance of the Wald-Q was compared to the IMSSA proposed in the first study and to iGDI proposed in the

second study in conjunction with the DINA and G-DINA models, respectively.

1.3 References

Akbay, L., Terzi, R., Kaplan, M., & Karaaslan, K. G. (2017). Expert-based attribute identification and validation: An application of cognitively diagnostic assessment. *Journal on Mathematics Education*, *9*(1). Retrieved from `http://ejournal.unsri.ac.id/index.php/jme/article/view/4341`

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*, 598–618.

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

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

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*, 163–183.

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253–273.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, *20*, 89–97.

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

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*, 191–210.

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.

Leighton, J., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, UK: Cambridge University Press.

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

Minchen, N., & de la Torre, J. (2016). *The continuous G-DINA model and the Jensen-Shannon Divergence*. Paper presented at the international meeting of Psychometric Society, Asheville, NC.

Morrison, D. F. (1967). *Multivariate statistical methods*. New York, NY: McGraw-Hill.

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

Tatsuoka, K. K. (1984). *Analysis of errors in fraction addition and subtraction problems*. Computer-based Education Research Laboratory, University of Illinois.

Terzi, R., & de la Torre, J. (2015). *An iterative method of empirically-based Q-matrix validation*. Paper presented at the annual aeeting of the National Council on Measurement in Education, Chicago, IL.

Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal*, *26*, 237–255.

Yigit, H., Sorrel, M., & de la Torre, J. (2016). *Computerized adaptive testing for cognitively based multiple-choice data*. Paper presented at the VII European Congress of Methodology, Mallorca, Spain.

Chapter 2

Study I: An Iterative Method for Empirically-Based Q-Matrix Validation

2.1 Abstract

The primary aim of cognitive diagnosis modeling is to identify finer-grained skill mastery or nonmastery of examinees. The framework for specifying required attributes for each item in a test is called Q-matrix. The traditional way of constructing a Q-matrix based on expert opinion is inherently subjective, consequently resulting in serious validity concerns. Misspecifications in the Q-matrix can negatively affect parameter estimation, and ultimately, attribute classifications. The current study proposes a new validation method under the deterministic inputs, noisy "and" gate (DINA) model to empirically validate attribute specifications in the Q-matrix. Simulation studies are conducted, and results based on the proposed method are compared to those of other parametric and nonparametric methods. Results show that the new method outperforms the other methods across the board. Finally, the method is applied to a real data example using fraction-subtraction data.

2.2 Introduction

The purpose of cognitive diagnosis modeling is to discover latent skills possessed by examinees based on their responses to test items. Cognitive diagnosis models (CDMs) require a Q-matrix (Tatsuoka, 1983) to identify the specific subset of attributes measured by each item. The entry q_{jk} in row j and column k of the Q-matrix is 1 if the kth attribute is required to correctly answer for item j, and 0 otherwise.

Due to its nature, constructing a Q-matrix is usually subjective, which has raised serious validity concerns among researchers. For instance, the estimation of model parameters, and ultimately the accuracy of attribute classifications may be negatively affected by including or omitting multiple q-entries in the Q-matrix (Rupp & Templin, 2008). However, the Q-matrix is usually assumed to be correct once specified by domain experts. This assumption is generally made because until recently, few well-established methods have become available to detect misspecifications in the Q-matrix (Chiu, 2013; DeCarlo, 2011; de la Torre, 2008), particularly when general CDMs are involved (de la Torre & Chiu, 2016; Liu, Xu, & Ying, 2012). Any analysis, such as model-fit evaluation, that does not check the correctness of the Q-matrix, becomes questionable.

These concerns have led to developments of some statistical methods for validating the appropriateness of Q-matrix specifications. For instance, Chiu (2013) proposed a Q-matrix refinement method (QRM) based on a nonparametric classification procedure (Chiu & Douglas, 2013). This method aims to minimize the residual sum of squares (RRS) between the observed and ideal responses among all the possible q-vectors given a Q-matrix. The RSS of item j across all examinees is defined as:

$$RRS_j = \sum_{i=1}^{N} (X_{ij} - \eta_{ij})^2 = \sum_{m=1}^{2^K} \sum_{i \in C_m} (X_{ij} - \eta_{jm}), \tag{2.1}$$

where X_{ij} and η_{ij} are the observed and ideal item responses of examinee i to item j,

respectively, C_m is the latent proficiency class m, and N is the number of examinees. Note that the index j of η_{ij} in Equation 2.1 was replaced by m because item responses are class-specific, meaning that every examinee in the same latent class is assumed to have the same ideal response to an item (Chiu, 2013). The RSS is used to identify any misspecified q-entries for an item. In the algorithm, the item vector with the highest RSS (i.e., possibly the worst item vector) gets replaced by the one having the lowest RSS (i.e., possibly the best item vector). The process is repeated iteratively until convergence is met. Due to its nature as a nonparametric method, it neither relies on the estimation of model parameters nor makes any assumptions other than those made by the CDM itself (Chiu, 2013). However, if the underlying model is known, parametric methods should provide more powerful results particularly when N is large. Additionally, the method cannot identify skills that have not been included in the provisional Q-matrix (Chiu, 2013), which is actually a similar problem of model misfit caused by an incomplete set of the skills in the Q-matrix (de la Torre & Chiu, 2016).

Another method is the empirically based δ-method proposed by de la Torre (2008), where a *sequential search algorithm* was implemented for the deterministic inputs, noisy "and" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model. To define the correct q-vector among the other possibilities, the discrimination index of item j, δ_j, is computed. The index δ_j is the difference in the probabilities of correct responses between examinees who have the required attributes (i.e., $\eta_j = 1$) and those who do not (i.e., $\eta_j = 0$). The following two algorithms for the δ-method were explained in detail in de la Torre (2008).

Exhaustive Search Algorithm. The *exhaustive search algorithm* (ESA) for Q-matrix validation computes δ_{jl} for all $l = 2^K - 1$ possible q-vectors for each j item (de la Torre, 2008). The q-vector that gives the largest difference in the probabilities of correct response between $\eta_{jl} = 1$ and $\eta_{jl} = 0$ among all the possible attribute patterns

is chosen as the correct q-vector for the item. However, the algorithm becomes impractical when K is reasonably large. Additionally, unlike the sequential search algorithm (SSA; discussed below), the ESA has the tendency to choose over-specified q-vectors (de la Torre, 2008).

Sequential Search Algorithm. The SSA, in comparison to the ESA, is considered more efficient because it does not require the comparisons of δ_{jl} for all the possible attribute patterns. More specifically, δ_{jl} is computed for $(K_j+1)K-(K_j^2+K_j)/2$ q-vectors for item j, where K_j is the number of attributes required for item j (de la Torre, 2008).

The SSA starts by comparing $\delta_{jl}^{(1)}$ of single-attribute q-vectors with the superscript (1) referring to single-attribute q-vectors. Let $\delta_j^{(1)}$ be the largest of $\delta_{jl}^{(1)}$ from single-attribute q-vectors, and assume that this is due to α_1. The process continues by examining δ_{jl} of two-attribute q-vectors, $\delta_j^{(2)}$, where α_1 is one of the required attributes. If $\delta_{jl}^{(2)} > \delta_{jl}^{(1)}$, the single-attribute q-vector is replaced by a two-attribute q-vector. However, if $\delta_{jl}^{(1)} > \delta_{jl}^{(2)}$, the process is terminated choosing α_1 as the correct attribute specification for the q-vector. Otherwise, the process continues with such comparisons until a K-attribute q-vector is chosen as long as the difference of succeeding δ_{jl} values (i.e., $\hat{\delta}_j^{(K_j+1)} - \hat{\delta}_j^{(K_j)}$) is larger than a predetermined cut-off value.

As stated earlier, estimation that involves some misspecified q-vectors can affect the quality of parameter estimations (Rupp & Templin, 2008), and this in turn affects the accuracy of the validation method. Similarly, the noise due to the stochastic nature of the response process makes it possible to obtain a q-vector with more attributes than necessary. Especially using real data can cause $\hat{\delta}_j^{(K_j+1)} > \hat{\delta}_j^{(K_j)}$ or the reverse, resulting in over- or under-specifications. A suggested solution is to assign ε, which is a minimum increment in the discrimination index of the item before an additional attribute can be included, as shown in $\hat{\delta}_j^{(K_j+1)} - \hat{\delta}_j^{(K_j)} > \varepsilon$ (de la Torre, 2008).

The SSA has some limitations. As noted by de la Torre (2008), an incorrect Q-matrix because of over- and under-specifications of attributes can cause bias in parameter estimations. This issue cannot be completely addressed by the ESA and SSA because they usually choose q-vectors that have more than single-attribute specifications. It is also not clear what ε to use in practice because it could vary depending on many conditions, such as changes in sample sizes, test lengths, item qualities, and amount of misspecifications, all of which were fixed in de la Torre's (2008) paper. It should also be noted that the algorithm was not implemented iteratively, meaning that the validation method stops after one full iteration even if changes are made in the provisional Q-matrix.

2.3 Background

2.3.1 The DINA Model

The DINA model has been commonly used in many studies (e.g., de la Torre & Douglas, 2004, 2008; de la Torre, 2009a; DeCarlo, 2011; Kuo, Pai, & de la Torre, 2016; Liu, Ying, & Zhang, 2015; Park & Lee, 2014; Rupp & Templin, 2008; Terzi & de la Torre, 2015). This study focuses on the DINA model because of its more straightforward interpretations, a smaller sample size requirements for accurate parameter estimation, (Rojas, de la Torre, & Olea, 2012), and its flexibility for extension to more general CDMs. The DINA model is an example of a conjunctive model for dichotomously scored test items, where all required attributes of an item should be mastered by examinees before she can be expected to correctly answer the item. Nonmastery of one or more required attributes for an item is equivalent to nonmastery of all required attributes. Let examinee i's binary attribute vector be denoted by $\boldsymbol{\alpha}_i = \{\alpha_{ik}\}$. The item response function of the model is defined as:

$$P(X_{ij}=1|\boldsymbol{\alpha}_i)=(1-s_j)^{\eta_{ij}}g_j^{(1-\eta_{ij})}, \tag{2.2}$$

which is the probability of answering an item j correctly by examinees with the attribute pattern $\boldsymbol{\alpha}_i$, X_{ij} is the response of examinee i to item j, and η_{ij} is the ideal response computed as:

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

indicates whether or not all of the required attributes associated with item j are mastered by examinee i.

2.3.2 An Iterative Method for Empirically-Based Q-Matrix Validation

This study introduces an iterative procedure in conjunction with a modified version of the SSA, and is called *iterative modified* SSA (IMSSA). The IMSSA differs from the SSA in two respects. First, the IMSSA determines required attribute specifications based on only the single-attribute q-vectors. Similar to the empirically based δ-method (de la Torre, 2008), the IMSSA starts by estimating the item parameters via an empirical Bayesian implementation of the expected-maximization (EM) algorithm (de la Torre, 2009b) using a provisional Q-matrix. The K δs corresponding to the single-attribute q-vectors (i.e., $\delta_j^{(1)}$) are then computed and ordered from the highest to the lowest. The correct attribute specification is determined based on the size of $\delta_{jl^*}^{(1)}$ relative to the next smaller $\delta_{jl^*+1}^{(1)}$ (i.e., $\delta_{jl^*}^{(1)}/\delta_{jl^*+1}^{(1)}$, for $l^*=1,2,\ldots,K-1$) for item j. The noise due to the use of the estimated posterior distribution should be controlled so as to not cause any over- or under-specifications. That can be done by using a cut-off point denoted by $\varepsilon^{(1)}$, which represent the minimum ratio between ordered single-attribute q-vectors. Specifically, if $\delta_{j1}^{(1)}$ is considerably larger than $\delta_{j2}^{(1)}$ (i.e., $\delta_{j1}^{(1)}/\delta_{j2}^{(1)}>\varepsilon^{(1)}$),

the required attribute would be an attribute specified in the single-attribute q-vector corresponding to $\delta_{j1}^{(1)}$; if not, the attribute specifications in the first two q-vectors are chosen. It continues by checking the ratio $\delta_{j2}^{(1)}/\delta_{j3}^{(1)}$. If the ratio is smaller than $\varepsilon^{(1)}$, the attribute specification in the third q-vector is also chosen on the top of the previous two specifications, and it continues; otherwise, the process is terminated. The ratio between $\delta_{jl^*}^{(1)}$ and $\delta_{jl^*+1}^{(1)}$ was determined based on some preliminary findings. The values of $\varepsilon^{(1)}$, the cut-off point, were determined using simulation.

At this point, an example can be helpful to lay out the rationale as to how the study determines the correctness of attribute specifications on the relative size of ordered δs. For illustration purposes, we considered two items, each with a misspecified attribute specification. In practice, the provisional Q-matrix may not have entirely correct specifications. However, data based on parameter estimates using the provisional Q-matrix can be generated. The δ-computation for the simulated data can be monitored, which can allow us to define extreme changes in the relative size of δ.

Table 2.1: Examples for Over- and Under-Specifications

	$(1,0,0,0,0)' \rightarrow (1,0,1,0,0)'$								$(1,1,0,0,0)' \rightarrow (1,0,0,0,0)'$							
l^*	α_1	α_2	α_3	α_4	α_5	$\delta_{jl^*}^{(1)}$	$\delta_{jl^*}^{(1)}/\delta_{l^*+1}^{(1)}$		α_1	α_2	α_3	α_4	α_5	$\delta_{jl^*}^{(1)}$	$\delta_{jl^*}^{(1)}/\delta_{l^*+1}^{(1)}$	
1	**1**	0	0	0	0	.41	**6.86**	✓	**1**	0	0	0	0	.40	1.38	✓
2	0	0	1	0	0	.06	1.37		0	**1**	0	0	0	.29	**6.43**	✓
3	0	0	0	0	1	.04	1.02		0	0	0	0	1	.05	-3.39	
4	0	1	0	0	0	.04	-10.9		0	0	1	0	0	-.01	0.41	
5	0	0	0	1	0	-.00	–		0	0	0	0	1	-.03	–	

Note. The symbol ✓displays the chosen attributes based on the associated δ-ratio. $(1,0,0,0,0)' \rightarrow (1,0,1,0,0)'$: $(1,0,0,0,0)'$ is over-specified as in $(1,0,1,0,0)'$. $(1,1,0,0,0)' \rightarrow (1,0,0,0,0)'$: $(1,1,0,0,0)'$ is under-specified as in $(1,0,0,0,0)'$. Negative values in the ratio come from the negative δ. For example, .52 and .49 for the slip and guessing parameters, respectively, $\delta_{jl^*=4}^{(1)} = 1 - s_{jl^*=4} - g_{jl^*=4} = 1 - .52 - .49 = -.01$.

Examples of $\delta_{jl*}^{(1)}$ computations for the simulated data can help determine whether or not the algorithm could identify correct specifications. Assume that $K = 5$. Table 2.1 displays examples of items that have over- and under-specifications. In the first misspecification, the q-vector $(1,0,0,0,0)'$ is over-specified as in $(1,0,1,0,0)'$. The EM estimation is carried out with the latter q-vector, and δs of single-attribute q-vectors are computed and sorted from the highest to the lowest. The result suggests that the correct attribute specification is only α_1 ($\delta_{j1}^{(1)} = .41$) due to a large drop in $\delta_{j2}^{(1)}$ (i.e., $\delta_{j1}^{(1)}/\delta_{j2}^{(1)} = 6.86 > \varepsilon^{(1)}$). A similar result is also observed for an item that has been under-specified. The misspecification appears as $(1,0,0,0,0)'$ from the correct vector of $(1,1,0,0,0)'$ in the right-hand side of Table 2.1. The ratio of that highest $\delta_{j1}^{(1)}$ to the second highest $\delta_{j2}^{(1)}$ shows a small drop (i.e., $\delta_{j1}^{(1)}/\delta_{j2}^{(1)} = 1.38$); however, the next ratio is rather large (i.e., $\delta_{j2}^{(1)}/\delta_{j3}^{(1)} = 6.43$). Therefore, the attributes in the first two single-attribute q-vectors are accurately specified (i.e., α_1 and α_2).

Second, the IMSSA becomes more efficient than the original SSA because δ is not computed beyond single-attribute vectors. As such, the maximum number of comparisons for the new algorithm is K, which is considerably smaller than SSA (i.e., $(K_j+1)K - (K_j^2+K_j)/2$) and ESA (i.e., $2^K - 1$), where K is the total number of attributes and K_j is the number of attributes being measured by item j. For example, let $K = 10$ and $K_j = 3$. The maximum number of comparisons is 10 for the IMSSA, 34 for the SSA, and 1023 for the ESA. Thus, using the IMSSA can lessen complications associated with multiple search steps.

In summary, examining the relative size of δ using a provisional q-vector could suggest which attributes should be required – δ of required attributes are considerably larger compared to δ of other attributes. The new method aims to make two crucial contributions to the Q-matrix validation literature. First, using simulation, an approximation was made to generally define optimal $\varepsilon^{(1)}$ values applicable across specific

set of conditions. These values were determined based on the estimated item quality, which were divided into three levels (i.e., high, medium, and low) to define three ε values for the each level. Second, the algorithm was implemented iteratively, such that, if any q-vectors are changed in the previous iteration, a new calibration is carried out using the updated Q-matrix as the provisional Q-matrix. The iterative algorithm aims to alleviate negative effects of any misspecified attribute specification given in the preceding iteration. In this present study, iterative and non-iterative algorithms were compared to examine if an iterative algorithm can further identify and correct misspecifications in succeeding iterations.

2.4 Design and Analysis

To evaluate the viability of the proposed method, two simulation studies were conducted with the following goals: (1) to determine $\varepsilon^{(1)}$ values as a function of the estimated item quality, which could be generalized to various conditions; and (2) to compare the effectiveness of different validation methods with iterative and noniterative algorithms. The attribute profiles were generated with equal probabilities from a multinomial distribution. For each condition, 100 datasets were replicated using the DINA model with the following factors: sample sizes (N = 1,000 and 2,000), test lengths ($J = 15$ and 30), item qualities ($s_j = g_j = 0.1$, 0.2, and 0.3), and amount of misspecifications (5% and 10%). In this study, the three sets of item qualities were considered similar to Hou, de la Torre, and Nandakumar (2014). In each condition, 100 misspecified Q-matrices were generated, which contain 5% or 10% randomly misspecified q-entries. For example, if a Q-matrix has 10% misspecifications for $J = 30$ and $K = 5$, 15 of 150 entries were randomly altered by producing over- or under-specified q-vectors. In doing so, the study was able to focus on the impact of the amount of misspecifications rather than the type of misspecifications. Two constraints were imposed on altering the q-vectors such that misspecified q-vectors cannot have more than

two-attribute misspecified, and at least one attribute should be specified as 1. It should be noted that the true Q-matrices in Table 2.2 for $J = 15$ and 30 are related in two ways. Each attribute is measured six and 12 times when $J = 15$ and 30, respectively, and there are equal numbers of 1-, 2-, and 3-attribute q-vectors in the each Q-matrix.

Table 2.2: True Q-matrix for the Simulated Data

Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5
1*	1	0	0	0	0	11*	1	1	0	0	0	21*	1	1	1	0	0
2*	0	1	0	0	0	12*	1	0	1	0	0	22*	1	1	0	1	0
3*	0	0	1	0	0	13	1	0	0	1	0	23	1	1	0	0	1
4*	0	0	0	1	0	14	1	0	0	0	1	24	1	0	1	1	0
5*	0	0	0	0	1	15	0	1	1	0	0	25*	1	0	1	0	1
6	1	0	0	0	0	16	0	1	0	1	0	26	1	0	0	1	1
7	0	1	0	0	0	17*	0	1	0	0	1	27	0	1	1	1	0
8	0	0	1	0	0	18*	0	0	1	1	0	28	0	1	1	0	1
9	0	0	0	1	0	19	0	0	1	0	1	29*	0	1	0	1	1
10	0	0	0	0	1	20*	0	0	0	1	1	30*	0	0	1	1	1

Note. Items with * are used for $J = 15$.

To define approximate $\varepsilon_j^{(1)}$ values for the IMSSA, the estimated item quality was employed (i.e., $\hat{\delta}_j = 1 - \hat{s}_j - \hat{g}_j$). Three levels of ε values were chosen to define each item quality in accordance with Hou et al. (2014). If $\hat{\delta}_j \geq 0.70$, it was considered a high quality item, $0.50 \leq \hat{\delta}_j < 0.70$ a medium quality item, and $\hat{\delta}_j < 0.50$ a low quality item. Based on the results of a pilot study, the performance of the proposed method deteriorates when $\varepsilon^{(1)}$ is outside 1.5 and 2.5. Hence, for this study, determining the optimal $\varepsilon^{(1)}$ focused in the range 1.5 to 2.5, with an increment of 0.1. After defining optimal ε values corresponding to different item qualities, the second simulation study was conducted to compare the four validation procedures: IMSSA, MSSA, ESA, SSA, and QRM. These methods were compared based on the proportions of identifying attribute specifications at the vector level. The code to implement the IMSSA, MSSA, ESA, and SSA was written in Ox (Doornik, 2009), whereas, the NPCD R package (Zheng & Chiu, 2015) was used (R Core Team, 2014) for the QRM analyses.

2.4.1 Results

In the first simulation study, the performance of the IMSSA was observed under different conditions to define $\varepsilon^{(1)}$ values for each level of the item qualities. Focusing in the range 1.5 to 2.5, the $\varepsilon^{(1)}$ values were derived based on the highest proportions of recovery on average across all the conditions for each item quality. Table 2.3 displays the highest recovery.

For example, when $\varepsilon^{(1)} = 1.9$ for $0.50 \leq \hat{\delta}_j < 0.70$, 96% of the q-vectors were correctly identified on average, where response data were generated from the medium quality item (i.e., $s = g = 2$). The highest proportions of recovery were observed for $\hat{\delta}_j \geq 0.70$ and $\hat{\delta}_j < 0.50$ when the $\varepsilon^{(1)} = 2.2$ and 1.7, respectively. It shows that item qualities had a slightly different impact on the recovery rates. In short, $\varepsilon^{(1)}$ values 2.2, 1.9, and 1.7 were used for high, medium, and low quality items, respectively.

Given the cut-off values for the IMSSA and MSSA methods, the non-iterative (ESA, SSA, and MSSA) and iterative methods (QRM and IMSSA) were compared in the second simulation study. Table 2.4 shows results based on response data generated from the medium quality item. Findings are reported at the vector level, and, on average, six and 12 q-vectors were altered for $J = 15$ and 30, respectively.

Table 2.4 shows that among the non-iterative methods, the MSSA outperformed the others for each simulation condition considered in this study. In addition, the SSA provided better recovery than the ESA for Q-matrix validation. In general across the conditions, the average recovery of the MSSA was 92% of the q-vectors; whereas, the average recoveries of the SSA and ESA were 83% and 74%, respectively. The results showed that the modified version of the SSA (MSSA) improved the recovery in comparison to the SSA.

Table 2.3: Proportions of Recovery Comparisons

				ε										
Quality	*N*	*J*	%	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.5
H	1,000	15	5	.989	.995	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.997
			10	.993	.961	.964	.968	.972	.969	.969	.969	.965	.954	.940
		30	5	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			10	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.997
	2,000	15	5	.997	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			10	.928	.952	.963	.963	.969	.971	.977	.979	.968	.970	.975
		30	5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.996	.996	.996
		Average		.979	.987	.989	.990	.992	.991	.992	**.993**	.990	.988	.986
M	1,000	15	5	.926	.943	.947	.953	.960	.959	.957	.958	.946	.941	.936
			10	.816	.843	.861	.852	.853	.847	.833	.810	.787	.767	.749
		30	5	.957	.979	.988	.994	.994	.993	.992	.991	.989	.988	.986
			10	.949	.972	.985	.990	.992	.992	.990	.991	.990	.988	.983
	2,000	15	5	.949	.958	.963	.965	.969	.969	.971	.971	.973	.969	.963
			10	.836	.868	.874	.881	.874	.867	.858	.853	.839	.831	.806
		30	5	.992	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			10	.981	.995	.999	1.00	1.00	1.00	.999	.999	.999	.997	.997
		Average		.926	.944	.952	.954	**.955**	.953	.950	.947	.940	.935	.927
L	1,000	15	5	.740	.773	.803	.797	.795	.792	.777	.775	.775	.761	.760
			10	.597	.619	.628	.613	.619	.613	.601	.583	.595	.578	.565
		30	5	.825	.856	.869	.879	.877	.874	.868	.860	.850	.843	.832
			10	.764	.789	.799	.792	.790	.767	.753	.742	.722	.705	.685
	2,000	15	5	.841	.853	.850	.839	.845	.834	.820	.817	.817	.799	.801
			10	.681	.671	.679	.673	.647	.651	.630	.625	.611	.612	.595
		30	5	.912	.922	.930	.927	.926	.924	.914	.909	.899	.889	.873
			10	.855	.871	.869	.861	.840	.821	.805	.769	.748	.720	.705
		Average		.777	.794	**.803**	.798	.792	.785	.771	.760	.752	.739	.727

Table 2.4: Proportions of Recovery Comparisons between Methods with and without Iterative Algorithm ($s = g = 0.2$)

			Non-Iterative			Iterative	
N	J	%	ESA	SSA	MSSA	QRM	IMSSA
1,000	15	5	0.79	0.86	0.94	0.91	0.96
		10	0.61	0.74	0.76	0.73	0.83
	30	5	0.80	0.89	0.99	1.00	1.00
		10	0.67	0.89	0.97	0.99	0.99
2,000	15	5	0.83	0.79	0.95	0.87	0.97
		10	0.63	0.68	0.78	0.68	0.85
	30	5	0.85	0.89	1.00	1.00	1.00
		10	0.71	0.88	0.98	1.00	1.00
Average			0.74	0.83	0.92	0.90	0.95

Note. ESA: exhaustive search algorithm, SSA: sequential search algorithm with $\varepsilon = .01$, MSSA: non-iterative modified sequential search algorithm, QRM: Q-matrix refinement method with an iterative algorithm, IMSSA: iterative modified sequential search algorithm.

In general, the proposed MSSA and IMSSA worked much better than the other methods. Specifically, after averaging the proportions of recovery across the conditions (i.e., N, J, and amount of misspecifications), recovery based on the IMSSA (95%) and MSSA (92%) was 5% and 2% higher than that of the QRM (90%), respectively. Additionally, recovery based on the MSSA (92%) was 18% and 9% higher than that of the ESA (74%) and SSA (83%), respectively.

In comparing the iterative methods (i.e., IMSSA and QRM) under the item quality given, the IMSSA worked usually equally well as or better than the QRM. In continuing the comparison of iterative methods under other item qualities, both iterative methods had perfect performance under data generated from the high quality item (i.e., $s = g = 1$). Only one exception occurred for both methods where the recovery was above 97% when $J = 15$ with 10% misspecifications. When data were generated from the low quality item (i.e., $s = g = 3$), the IMSSA (80%) had 7% more recovery than the QRM

(73%).

It is interesting to report that the performance of the SSA and QRM was equally well or worse when the sample size was doubled. For example, under a condition where $N = 1,000$, $J = 15$ with 10% misspecifications, doubling the sample size to $N = 2,000$ resulted in the recovery dropping from 74% to 68% for the SSA and from 73% to 68% for the QRM. In contrast, considering the same conditions, the recovery improved from 61% to 63% for the ESA, from 76% to 78% for MSSA, and from 83% to 85% for the IMSSA. After doubling the test items from 15 to 30, the recovery increased from 74% to 89% for the SSA and from 73% to 99% for the QRM. The improvement was also substantial for the ESA, MSSA and IMSSA. Specifically, the recovery improved from 61% to 67% for the ESA, from 76% to 97% for SSA, and from 83% to 99% for the IMSSA. This finding indicates that doubling the test length can lead to better improvement in recovery more than doubling the sample size for the ESA, MSSA and IMSSA.

Similarly, with regards to the difference in recovery rates due to the amount of misspecifications within the same conditions (i.e., N and J), a larger test length provided a smaller gap than a larger sample size. That is, recovery differences between 5% and 10% misspecifications were higher with a small sample size and short length test. For example, among the non-iterative methods when $N = 1,000$ and $J = 15$, recovery differences between 5% and 10% misspecifications were 12%, 18%, and 18% for the SSA, ESA, and MSSA, respectively; it dropped to 0%, 13%, and 2% when $J = 30$ holding the sample size constant. Doubling the sample size with a fixed test length did not change the recovery differences, which was only 11%, 20%, and 17% for the SSA, ESA, and MSSA, respectively. In taking the amount of misspecifications into account for the non-iterative methods, doubling the test length had a considerably positive impact on the recovery for the SSA and MSSA, but it had a negative impact on the recovery for the ESA.

For the iterative methods, again, doubling the test length decreased the difference in recovery rates between 5% and 10% misspecified Q-matrices. Similarly, when $N = 1,000$ and $J = 15$, it was 18% for the QRM (i.e., $91 - 73 = 18$) and 13% for the IMSSA (i.e., $96 - 83 = 13$). However, that gap was smaller when $J = 30$ than $N = 2,000$. The difference substantially dropped for both methods after doubling the test length with a constant sample size. However, that gap had a 1% increase for the QRM and 1% decrease for the IMSSA after doubling the sample size with the constant test length (i.e., $J = 15$). Therefore, based on these findings, it can be stated that doubling the test length substantially improved the recovery for both iterative methods and decreased the recovery differences depending on the amount of misspecifications. For the QRM, doubling the test length had a positive impact but doubling the sample size had a negative impact on the recovery.

2.5 Implementation with Real Data

In addition to the simulation study, real data were analyzed to investigate the applicability of the method. The fraction-subtraction data (Tatsuoka, 1984) with 536 middle school students responses to 12 fraction subtraction problems were examined. The four attributes for this dataset are: (α_1) performing a basic fraction subtraction operation, (α_2) simplifying/reducing, (α_3) separating a whole number from fraction, and (α_4) borrowing one from a whole number to fraction. The 12 items with the corresponding attribute specifications and δ values are shown in Table 2.5. For the IMSSA, $\delta^{(1)}_{jl^*}$ statistic was computed and $\delta^{(1)}_{jl^*}/\delta^{(1)}_{jl^*+1}$ ratios were reported.

2.5.1 Results

Note that the data set of Tatsuoka (1984) has been one of the most commonly examined real data designed for cognitively diagnostic assessment (e.g., Chiu, 2013; Chiu

Table 2.5: Q-Matrix for Fraction-Subtraction Items

Item		Attributes				
		α_1	α_2	α_3	α_4	δ
1.	$\frac{3}{4}-\frac{3}{8}$	1	0	0	0	0.72
2.	$3\frac{1}{2}-2\frac{3}{2}$	1	1	1	1	0.66
3.	$\frac{6}{7}-\frac{4}{7}$	1	0	0	0	0.83
4.	$3\frac{7}{8}-2$	1	0	1	0	0.42
5.	$4\frac{4}{12}-2\frac{7}{12}$	1	1	1	1	0.74
6.	$4\frac{1}{3}-2\frac{4}{3}$	1	1	1	1	0.86
7.	$\frac{11}{8}-\frac{1}{8}$	1	1	0	0	0.80
8.	$3\frac{4}{5}-3\frac{2}{5}$	1	0	1	0	0.86
9.	$4\frac{5}{7}-1\frac{4}{7}$	1	0	1	0	0.80
10.	$7\frac{3}{5}-\frac{4}{5}$	1	0	1	1	0.84
11.	$4\frac{1}{10}-2\frac{8}{10}$	1	1	1	1	0.71
12.	$4\frac{1}{3}-1\frac{5}{3}$	1	1	1	1	0.82

Note. α_1 – performing a basic fraction subtraction operation; α_2 – simplifying/reducing; α_3 – separating a whole number from fraction; and α_4 – borrowing one from a whole number to fraction.

& Köhn, 2015; de la Torre, 2008; de la Torre & Chiu, 2016; DeCarlo, 2011). In CDM analyses, one of the main concerns is the completeness of the Q-matrix. Unfortunately, the fraction-subtraction data do not appear to have a complete Q-matrix. It was demonstrated by Chiu, Douglas, and Li (2009) that a complete Q-matrix should identify all possible attribute patterns and require each attribute to be represented by at least one single-attribute vector. This issue has been further discussed with the original data (see Table 4 on pp. 615, Chiu, 2013; DeCarlo, 2011) or subsets of it (see de la Torre, 2008; de la Torre & Chiu, 2016). The incompleteness of the Q-matrix in this dataset occurs because of the fact that only 58 of 256 ($K = 8$; Chiu, 2013) and 10 of 32 ($K = 5$; Chiu & Köhn, 2015) possible attribute patterns can be identified by the items, meaning that multiple classes may be merged (Chiu, 2013). Therefore, results of this data analysis should be interpreted with caution.

Table 2.6: Single-attribute specifications and the corresponding $\delta-$values and $-$ratios

j	α_1	α_2	α_3	α_4	δ_{jl^*}	$\delta_{jl^*}/\delta_{jl^*+1}$		j	α_1	α_2	α_3	α_4	δ_{jl^*}	$\delta_{jl^*}/\delta_{jl^*+1}$	
1	1*	0	0	0	0.72	1.61	✓	7	1*	0	0	0	0.73	1.03	✓
	0	0	1*	0	0.45	1.13	✓		0	1*	0	0	0.71	1.25	✓
	0	1*	0	0	0.40	1.16	✓		0	0	1*	0	0.56	3.67	✓
	0	0	0	1*	0.34		✓		0	0	0	1	0.15		
2	0	0	0	1*	0.55	1.60	✓	8	1*	0	0	0	0.82	1.09	✓
	1*	0	0	0	0.34	1.12	✓		0	0	1*	0	0.75	1.49	✓
	0	1*	0	0	0.30	1.01	✓		0	1*	0	0	0.51	3.92	✓
	0	0	1*	0	0.30		✓		0	0	0	1	0.13		
3	1*	0	0	0	0.83	1.84	✓	9	1*	0	0	0	0.75	1.07	✓
	0	0	1*	0	0.45	1.23	✓		0	0	1*	0	0.71	1.45	✓
	0	1*	0	0	0.37	5.21	✓		0	1*	0	0	0.49	3.34	✓
	0	0	0	1	0.07				0	0	0	1	0.15		
4	1*	0	0	0	0.39	1.04	✓	10	0	0	0	1*	0.66	1.27	✓
	0	0	1*	0	0.37	1.44	✓		1*	0	0	0	0.52	1.06	✓
	0	1*	0	0	0.26	3.35	✓		0	0	1*	0	0.49	1.06	✓
	0	0	0	1	0.08				0	1*	0	0	0.46		✓
5	0	0	0	1*	0.57	1.23	✓	11	1*	0	0	0	0.56	1.10	✓
	1*	0	0	0	0.47	1.12	✓		0	0	0	1*	0.51	1.01	✓
	0	1*	0	0	0.42	1.01	✓		0	0	1*	0	0.50	1.04	✓
	0	0	1*	0	0.41		✓		0	1*	0	0	0.48		✓
6	0	0	0	1*	0.67	1.26	✓	12	0	0	0	1*	0.64	1.33	✓
	1*	0	0	0	0.53	1.05	✓		1*	0	0	0	0.48	1.02	✓
	0	1*	0	0	0.51	1.04	✓		0	1*	0	0	0.47	1.06	✓
	0	0	1*	0	0.49		✓		0	0	1*	0	0.44		✓

Note. * indicates suggested attribute specifications.

Table 2.7: Suggested Q-Matrix by the IMSSA and QRM

	IMSSA				QRM			
Item	α_1	α_2	α_3	α_4	α_1	α_2	α_3	α_4
1.	1	1*	1*	1*	1	0	0	1*
2.	1	1	1	1	1	1	1	1
3.	1	1*	1*	0	1	0	0	0
4.	1	1*	1	0	1	0	1	0
5.	1	1	1	1	1	1	1	1
6.	1	1	1	1	1	1	1	1
7.	1	1	1*	0	1	1	0	0
8.	1	1*	1	0	1	0	1	0
9.	1	1*	1	0	1	0	1	0
10.	1	1*	1	1	1	1*	1	1
11.	1	1	1	1	1	0*	1	1
12.	1	1	1	1	1	1	1	1

Note. α_1 – performing a basic fraction subtraction operation; α_2 – simplifying/reducing; α_3 – separating a whole number from fraction; and α_4 – borrowing one from a whole number to fraction; * indicates suggested attribute specifications.

The changes in the relative size of δs corresponding to each single-attribute q-vector for 12 items are reported in Table 2.6. The suggested Q-matrix is also shown in Table 2.7. Given the results in the first simulation study, $\varepsilon^{(1)}$ values were set at 2.2, 1.9, and 1.7 when $\hat{\delta}_j \geq 0.70$, $0.50 \leq \hat{\delta}_j < 0.70$, and $\hat{\delta}_j < 0.50$, respectively. The results of the fraction-subtraction data obtained from the IMSSA were compared to the QRM. The IMSSA suggested attribute changes in seven items (i.e., items 1, 3, 4, 7, 8, 9, and 10); whereas, the QRM suggested attribute changes in three items (i.e., items 1, 10, and 11). Based on the IMSSA, the result indicated that item 1 (i.e., $\frac{3}{4} - \frac{3}{8}$) should require the other three attributes in addition to α_1. This suggestion may have occurred because this item requires more than just "performing a basic fraction subtraction problem" (i.e., α_1). Another suggestion was for item 3 (i.e., $\frac{6}{7} - \frac{4}{7}$), where α_2 and α_3 were

deemed required. Items 4 (i.e., $3\frac{7}{8}-2$), 8 (i.e., $3\frac{4}{5}-3\frac{2}{5}$), 9 (i.e., $4\frac{5}{7}-1\frac{4}{7}$), and 10 (i.e., $7\frac{3}{5}-\frac{4}{5}$) required α_2 in addition to α_1 and α_3. Note that another strategy for solving the problem in one of these four items – borrowing one from a whole number to fraction, performing a basic fraction, and simplifying/reducing – happens to give the correct answer. The following example shows another strategy to solve item 9:

$$\begin{aligned} 4\frac{5}{7}-1\frac{4}{7} &= \frac{(4\times 7)+5}{7}-\frac{(1\times 7)+4}{7} \\ &= \frac{33-11}{7}=\frac{22}{7}=3\frac{1}{7}. \end{aligned}$$

Another attribute suggestion (i.e., α_3) was for item 7 (i.e., $\frac{11}{8}-\frac{1}{8}$) on the top of α_1 and α_2. Similar to the preceding example, a different strategy – separating a whole number from fraction, performing a basic fraction subtraction operation, and simplifying/reducing – could also give the correct answer to item 7, as in,

$$\begin{aligned} \frac{11}{8}-\frac{1}{8} &= 1\frac{3}{8}-\frac{1}{8}=1\frac{3-1}{8} \\ &= 1\frac{2}{8}=1\frac{1}{4}. \end{aligned}$$

In applying the QRM, Chiu found that item 4, which appears as item 2 in this study, did not require the possession of the third attribute to be correctly answered (2013). In contrast, the QRM in this study suggested that the third attribute specification was necessary. An explanation could be because of the fact that Chiu used 20 items with 8 attributes (2013). Whereas, the IMSSA indicated that the mastery of the third attribute was required to answer item 2 correctly. The QRM also suggested to include and exclude α_2 in items 10 and 11, respectively.

As demonstrated by the examples, a deeper analysis is needed. The IMSSA has

more 1s than the QRM that can be controlled by adjusting the cut-offs. The cut-off values defined in the simulation study do not correspond to the real data analysis, which did not have a complete Q-matrix. The latter values were just approximations based on the item qualities defined in the simulation study. However, having more 1s can be a source of evidence about multiple strategies examinees can use. Further discussions about multiple strategies in cognitive diagnosis using the fraction subtraction data can be found in de la Torre and Douglas (2008), Huo and de la Torre (2014), and Mislevy (1996). Other reasons could be because the fraction subtraction data have fewer number of items and attributes than the simulation study.

As stated, the general purpose of this study was not to replace existing validation methods but rather to serve as a supplementary tool from a statistical point of view. Moreover, employing both existing validation methods and domain experts may be a more appropriate process for Q-matrix validation.

2.6 Summary and Discussion

CDMs aim to classify the attribute mastery or nonmastery of examinees. The Q-matrix is needed for specifying required attributes for each item in a test. The importance of revising attribute specifications in the Q-matrix should not be underestimated due to the inherent subjectivity of domain experts, consequently resulting in serious validity concerns.

The IMSSA for Q-matrix validation presented in this study aimed to empirically validate attribute specifications under a wide range of conditions with two degrees of misspecifications. This work extended the SSA (de la Torre, 2008) in several ways. First, it offered a more efficient solution that only the first K single-attribute q-vectors were examined. Attribute specifications were identified depending on changes in the

size of ordered δ-statistic. Second, in addition to less number of computational requirements, an iterative algorithm was included in the method to decrease negative effects of any misspecified attribute specification given in the previous iteration. Third, an approximation was made to generally define optimal $\varepsilon_j^{(1)}$ values applicable across the specific set of conditions. These values were arbitrarily determined based on the estimated item qualities as defined in Hou et al. (2014), and were divided into three levels (i.e., high, medium, and low) so that a different ε value for the each level can be defined. Choosing these values were not necessarily complete, but provides an approximation to the three levels of the item qualities.

Two simulation studies were carried out to compare results between iterative and non-iterative methods as well as between parametric and nonparametric methods. Therefore, iterative and non-iterative algorithms can be examined if an iterative algorithm can further identify and correct misspecifications in succeeding iterations. After fixing ε values based on the estimated item qualities in the first study, the second study compared three methods without iterative algorithms (SSA, ESA, and MSSA) to two methods with iterative algorithms (IMSSA and QRM). Among the noniterative methods, the MSSA reported better results, which had higher recovery than the QRM on average across all the factors.

Furthermore, as expected, the results showed that the IMSSA worked much better than the noniterative methods. Specifically, the iterative parametric (i.e., IMSSA) and nonparametric (i.e., QRM) methods were separately examined. The IMSSA resulted in higher recovery rates in a short test. In particular, a large sample size with a short test uncovered larger recovery discrepancy in favor of the IMSSA. However, when test length was long, both methods worked equally well. Furthermore, given large sample sizes and long tests, both methods had a perfect recovery regardless of the conditions and amount of misspecifications. On average, the IMSSA outperformed the QRM when data were generated from the medium quality item. When data were generated

from the high quality item, both methods had perfect recovery with a longer test, and a very high recovery rate with a short test; and that from the low quality item, the IMSSA still had a higher recovery rate than the QRM.

According to the simulation studies, the IMSSA showed promising improvements in Q-matrix validation that could enhance the estimation of model parameters, model-data fit analyses, and ultimately, the accuracy of attribute-classifications. It was further applied to the fraction-subtraction data. Based on the results suggested by the IMSSA, there were some attribute suggestions for seven items. Existence of a different strategy to solve the items could be a possible explanation for changes in attribute specifications. This study suggests that the existing methods should be complementary to each other so as to help domain experts for further investigation.

Using a 3.50-GHz I7 computer, it took the code the least amount of time to run the validation procedures for MSSA, followed by IMSSA, ESA, SSA, and QRM. For instance, it took 1.64, 3.11, 9.89, 24.35, and 30.00 minutes using MSSA, IMSSA, ESA, SSA, and QRM procedures, respectively, for 100 iterations under the condition where $N = 2,000$, $J = 30$, and medium quality items with 10% misspecifications in the Q-matrix. The number of iterations in the iterative procedures was usually between two and three, and did not go beyond four.

This present study had some limitations. For instance, the number of attributes was assumed to be known and fixed to $K = 5$. It would be interesting to investigate the method by relaxing this assumption. The findings of this study were based on the attribute structure generated from a uniform distribution. The performance of the methods should be investigated under a condition where attributes were generated from a higher order distribution (de la Torre & Douglas, 2004). Also, in addition to the δ-statistic used in this study, other statistics can be carried out for Q-matrix validation. This study should also be extended to make it applicable to a wider class of CDMs such as the G-DINA model (de la Torre, 2011). This will obviate the need to assume the

specific CDMs involved. Finally, this method should be applied to other real data sets with a complete Q-matrix so that further insights can be gained on how the proposed method could work in practice.

2.7 References

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*, 598–618.

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

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*, 633–665.

Chiu, C.-Y., & Köhn, H.-F. (2015). Consistency of cluster analysis for cognitive diagnosis the DINO model and the DINA model revisited. *Applied Psychological Measurement*, *39*, 465–479.

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

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

de la Torre, J. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*, 163–183.

de la Torre, J. (2009b). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253–273.

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

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*, 595–624.

Doornik, J. A. (2009). An object-oriented matrix programming language Ox 6. [Computer software]. London, UK: Timberlake Consultants Ltd.

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

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*, 98–125.

Huo, Y., & de la Torre, J. (2014). Estimating a cognitive diagnostic model for multiple strategies via the EM algorithm. *Applied Psychological Measurement*, *38*, 464–485.

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.

Kuo, B.-C., Pai, H.-S., & de la Torre, J. (2016). Modified cognitive diagnostic index and modified attribute-level discrimination index for test construction. *Applied Psychological Measurement*, *40*, 315–330.

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

Liu, J., Ying, Z., & Zhang, S. (2015). A rate function approach to computerized adaptive testing for cognitive diagnosis. *Psychometrika*, *80*, 468–490.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*, 379–416.

Park, Y. S., & Lee, Y.-S. (2014). An extension of the DINA model using covariates examining factors affecting response probability and latent classification. *Applied Psychological Measurement*, *38*, 376–390.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rojas, G., de la Torre, J., & Olea, J. (2012). Choosing between general and specific cognitive diagnosis models when the sample size is small. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

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

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

Tatsuoka, K. K. (1984). *Analysis of errors in fraction addition and subtraction problems*. Computer-based Education Research Laboratory, University of Illinois.

Terzi, R., & de la Torre, J. (2015). *An iterative method of empirically-based Q-matrix validation*. Paper presented at the annual aeeting of the National Council on Measurement in Education, Chicago, IL.

Zheng, Y., & Chiu, C.-Y. (2015). NPCD: The R package for nonparametric methods for cognitive diagnosis.

Chapter 3

Study II: The Iterative Jensen-Shannon Divergence Index and Iterative GDI for Q-Matrix Validation

3.1 Abstract

Interest in formative assessment has been growing in the educational measurement field. Specifically, cognitively diagnostic assessments (CDAs) are designed to provide more specific information to pinpoint teaching and learning deficiencies in classroom settings. Ultimately, CDAs aim to determine examinees' mastery or nonmastery of attributes in a particular content area. The Q-matrix is an important component of CDMs that plays a key role in specifying required attributes for each item. The Q-matrix is constructed by domain experts, which involves inherent subjectivity. Incorrect entries in the Q-matrix can have a negative impact on examinee classifications as a result of inaccurate item parameter estimation. Therefore, validity concerns have been brought to experts' attentions. This study proposes new indices, an *iterative* Jensen-Shannon divergence (iJSD) index and an *iterative* G-DINA model discrimination index (iGDI), to determine the correctness of attribute specifications in the Q-matrix for the DINA model. Simulation studies are implemented to investigate the false-positive and true-positive rates of both indices under a number of conditions. Results show that the indices can identify misspecified q-entries at a high rate, in particular, when attributes are correlated, and the false-negative rate is around the nominal level under favorable conditions. The paper concludes with discussions about the strengths and limitations of the indices followed by suggestions for future studies.

3.2 Introduction

Cognitive diagnosis models (CDMs) are multidimensional latent variable models that provide detailed feedback on students' learning progress. These models are used to classify examinees based on their mastery profiles in contrast to the traditional psychometric frameworks, such as classical test theory and item response theory, in which latent variables are reported on continuous scales. Attribute mastery profiles determined by CDMs specify membership in various latent groups. Each attribute pattern is generally represented by a binary vector with 1s and 0s, indicating mastery and nonmastery of each attribute being measured, respectively. Determining such mastery profiles aims to help instructors provide more targeted remedial instruction.

Over the past decade, a number of CDMs have been proposed in the literature (e.g., de la Torre & Douglas, 2004; de la Torre, 2011; Hartz, Roussos, Henson, & Templin, 2005; Henson, Templin, & Willse, 2009; Junker & Sijtsma, 2001; Templin & Henson, 2006; von Davier, 2008). CDMs can be classified into two categories, reduced and general, depending on whether or not cognitive processes underlying the responses can be assumed. Some of the commonly used reduced models are the deterministic inputs, noisy "and" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model, the deterministic input, noisy "or" gate (DINO; Templin & Henson, 2006) model, the compensatory and reduced reparameterized unified model (C-RUM and R-RUM; Hartz et al., 2005), the multiple-choice deterministic inputs, noisy "and" gate (MC-DINA; de la Torre, 2009) model, the *additive* CDM (*A*-CDM; de la Torre, 2011), and the linear logistic model (LLM; de la Torre & Douglas, 2004). Some of the commonly used general models are the general diagnosis model (GDM; von Davier, 2008), the log-linear CDM (Henson et al., 2009), and the generalized DINA (G-DINA; de la Torre, 2011) model. For the most part, general CDMs subsume reduced CDMs. Regardless of their generality, CDMs require a common component, called the Q-matrix (Tatsuoka, 1983).

The Q-matrix, generally a binary $J \times K$ matrix, is used to relate a specific subset of attributes to each item. Specifically, a q_{jk} entry, representing row j and column k of the Q-matrix, is 1 if the kth attribute is required to correctly answer item j, and is 0 if the kth attribute is not required. In the process of constructing the Q-matrix, domain experts need to spend a great deal of effort. However, inherently, this process is usually considered subjective due to the involvement of human judgments, and has raised serious validity discussions among researchers (e.g., Chiu, 2013; de la Torre, 2008; Liu, Xu, & Ying, 2012; Rupp & Templin, 2008). Entry misspecifications in the Q-matrix can have a negative impact on attribute classifications as a result of inaccurately estimating the model parameters (de la Torre & Chiu, 2016). Without validating the attribute specifications, it requires a big leap to assume that the Q-matrix entries have all been correctly specified by domain experts. To obviate the need to make such an assumption, researchers has proposed various methods to validate the attribute specifications in the Q-matrix (Chiu, 2013; DeCarlo, 2011; de la Torre, 2008; de la Torre & Chiu, 2016; Liu et al., 2012).

Until recently, only few methods have been developed to detect misspecifications in the Q-matrix. In 2008, de la Torre proposed an empirically based δ-method for provisional Q-matrix validation implemented through a sequential search algorithm under the DINA model. This method suggests the correct q-vector among all the possible $2^K - 1$ q-vectors based on the δ_{jl} value, which is the discrimination index of item j. The index for item j is the difference in the probabilities of correct responses between those who have mastered all the required attributes and those who have not, which are indicated by the latent variable, $\eta_{ij} = 1$ and $\eta_{ij} = 0$, respectively. The suggested attribute specifications of a q-vector are chosen so that the difference between success probabilities for the two groups is maximized. However, some questions remain opened. First, ε was used to prevent recovery from over- or under-corrections, but it remains uncertain how to determine ε values. Second, findings are not general enough,

which are limited to conditions fixed in de la Torre (2008).

Terzi and de la Torre (2015) introduced an empirically based iterative δ-method for Q-matrix validation by extending the empirically based δ-method (de la Torre, 2008). Some of the previous limitations were addressed. First, this method uses an iterative algorithm, which provided better results in identifying attribute specifications. Second, optimal values of ε were defined as a function of the estimated item quality. Third, the algorithm became more effective by focusing on single-attribute specifications. That is, the K single-attribute vectors were used to determine the attribute specifications. However, both empirically based δ-methods (de la Torre, 2008; Terzi & de la Torre, 2015) were limited to the DINA model only.

De la Torre and Chiu (2016) recently expanded the empirically based δ-method (de la Torre, 2008) to a wider class of CDMs, by developing a procedure based on the G-DINA model. Similar to de la Torre's (2008) study, they introduced a new discrimination index that has greater applicability. However, the generalizability of the findings was limited due to the fixed sample size, test length, and the number of attributes examined. Another concern of de la Torre and Chiu's (2016) work is that the GDI does not have a formal way for determining optimal ε values. The method is also not iterative in that it stops the recovery after suggesting attribute specifications at the first step.

A model-based approach was developed by DeCarlo (2011). In this method, after identifying possible misspecified Q-matrix entries in advance, these entries were considered random variables and estimated with the rest of the model parameters. However, in addition to identifying any misspecified q-vectors in advance, this method is also computationally time-consuming. Liu et al. (2012) have proposed a data-driven approach to identify misspecifications in the Q-matrix based on students' responses. This approach does not require any expert involvement. However, some limitations exist in that the identifiability of the Q-matrix may be weaker under the presence of unknown guessing parameters (Liu et al., 2012). This method was also only used with

specific CDMs, and requires further investigation under more complex methods.

Chiu (2013) developed a Q-matrix validation method (QRM) based on a nonparametric classification procedure (Chiu & Douglas, 2013). The QRM minimizes the residual sum of squares (RRS) between the observed and ideal responses among all the possible q-vectors given a provisional Q-matrix. The algorithm refines the attribute specifications based on an item vector with the highest RSS replaced by another q-vector with the lowest RSS, and is iterative in nature. This method, however, requires that conjunctive or disjunctive models be specified in advance. Moreover, because the QRM is a nonparametric method, parametric methods should provide more powerful results if the underlying model is assumed, specifically when N is large.

The primary purpose of this paper is to propose new empirically-based Q-matrix methods to validate the correctness of attribute specifications in the Q-matrix. To this end, the *iterative* Jensen-Shannon divergence (iJSD) index and the *iterative* G-DINA model discrimination index (iGDI) are introduced to obviate the need to assume any specific reduced CDMs in the item calibration. The main advantage of the iJSD is that it is applicable for dichotomous and nondichotomous models such as the continuous G-DINA model (Minchen & de la Torre, 2016) and MC-DINA model (Yigit, Sorrel, & de la Torre, 2016). Therefore, the iJSD is more general than the iGDI in that it can be applied under different types of models. The main advantage of the iGDI, an extension of the original GDI, is the inclusion of an iterative algorithm that could provide better results. As with the iGDI, the iJSD was also implemented iteratively. In the iterative process, if any changes occur in attribute specifications, a new calibration is carried out with the suggested (i.e., updated) Q-matrix so that any potential effect of misspecified entries can be eliminated. The feasibility of the iJSD and iGDI was investigated using simulated and fraction subtraction data.

3.3 Background

3.3.1 The G-DINA Model

The DINA model has been a commonly used CDM within the last two decades. This model partitions examinees into two groups: those who have all the required attributes and those who do not. The assumption is that missing any one of the required attributes is the same as missing all of them. However, this restriction does not usually reflect the reality that examinees with more attributes can be more capable than those with less attributes. Given this limitation of the DINA model, de la Torre (2011) proposed the generalized DINA model (G-DINA). The G-DINA model classifies examinees into 2^{K_j} latent groups, where K_j is the number of the required attributes for item j (i.e., $K_j = \sum_{k=1}^{K} q_{jk}$). Therefore, examinees who have mastered different attributes can have different probabilities of correctly answering an item.

Assume that item j requires the first $1, \cdots, K_j$ attributes. The reduced attribute vector can be denoted by $\boldsymbol{\alpha}_{lj}^*$, which represents the columns of the required attributes (i.e., $l = 1, \cdots, 2^{K_j}$). Thus, $P(X_j = 1|\boldsymbol{\alpha}_{lj}^*) = P(\boldsymbol{\alpha}_{lj}^*)$ is the probability of correctly answering an item j by examinees with attribute pattern $\boldsymbol{\alpha}_{lj}^*$. The item response function of the G-DINA model for the identity link is given by

$$P(\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j} \sum_{k=1}^{K_j-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{j12\ldots K_j} \prod_{k=1}^{K_j} \alpha_{lk}, \tag{3.1}$$

where δ_{j0} is the intercept for item j; δ_{jk} is the main effect of α_k; $\delta_{jkk'}$ is the interaction effect of α_k and $\alpha_{k'}$; and $\delta_{j12\ldots K_j}$ is the interaction effect of $\alpha_1, \ldots, \alpha_{K_j}$.

The G-DINA model is a saturated model that subsumes several commonly-used reduced CDMs, which are the DINA model, the DINO model, the A-CDM, the LLM, and the R-RUM. These reduced models can be obtained from the G-DINA model by

using different constraints and link functions (de la Torre, 2011). The item response function of the DINA model can be obtained from the G-DINA model by setting all the parameters in Equation 3.1 to zero, except for δ_{j0} and $\delta_{j12\ldots K_j}$. That is,

$$P(\boldsymbol{\alpha}^*_{lj}) = \delta_{j0} + \delta_{j12\ldots K_j} \prod_{k=1}^{K_j} \alpha_{lk}. \tag{3.2}$$

3.3.2 The G-DINA Model Discrimination Index

The G-DINA model discrimination index (GDI), denoted by $\hat{\varsigma}^2_j$ $(1 \leq l \leq 2^K)$, was proposed by de la Torre and Chiu (2016) to empirically validate the Q-matrix specifications for the G-DINA model. A theorem was discussed to justify the use of the index for the proposed validation method. The suggested q-vector is based on the proportion of variance accounted for (PVAF) by a q-vector relative to the maximum $\hat{\varsigma}^2_j$, which is obtained when all the attributes are specified (de la Torre & Chiu, 2016). In particular, the q-vector with the fewest number of attributes required corresponding to $\hat{\varsigma}^2_j$ that approximates the maximum GDI is suggested. The approximation was done with a predetermined cutoff value for PVAF set at $\varepsilon = 0.95$ (de la Torre & Chiu, 2016).

Given an attribute distribution, the ς^2_j measures the weighted variance of the probabilities of correctly answering item j. Let the first K_j attributes be required. The GDI of an item with a specification $\mathbf{q}_{K':K''}$ is defined as

$$\begin{aligned} \varsigma^2 = \varsigma^2_{K':K''} &= \sum_{\alpha_{K'}=0}^{1} \cdots \sum_{\alpha_{K''}=0}^{1} w(\alpha_{K':K''})[p(\alpha_{K':K''}) - \overline{p}(\alpha_{K':K''})]^2 \\ &= \sum_{\alpha_{K'}=0}^{1} \cdots \sum_{\alpha_{K''}=0}^{1} w(\alpha_{K':K''})p^2(\alpha_{K':K''}) - \overline{p}^2(\alpha_{K':K''}), \end{aligned} \tag{3.3}$$

where $\overline{p}(\alpha_{K':K''}) = \sum_{\alpha_{K'}=0}^{1} \cdots \sum_{\alpha_{K''}=0}^{1} w(\alpha_{K':K''})p(\alpha_{K':K''})$ is the weighted probability

of success across all the $2^{K''-K'+1}$ possible patterns of $p(\alpha_{K':K''})$; and $w(\alpha_{K':K''})$ is the posterior probability of examinees being in class $(\alpha_1, \ldots \alpha_{K''})$, which is equal to $\sum_{\alpha_1=0}^{1} \cdots \sum_{\alpha_{K'}-1=0}^{1} w(\alpha_{1:K''})$ (de la Torre & Chiu, 2016).

As briefly mentioned previously, the performance of the iGDI can be improved because the iterative procedure was applied to the original GDI. For example, when $N = 1,000$ and $J = 31$ with the medium quality item, including 5% misspecifications, the true-positive rate of the iGDI was 0.81 given the iterative algorithm; however, it would be 0.71 unless the iterative algorithm in the GDI was included. More specifically under the preceding condition, an original q-vector $(0,0,0,1,0)'$ was randomly misspecified as $(1,0,0,0,1)'$ in the provisional Q-matrix. The following vectors were the suggested q-vectors at each succeeding iteration: $(1,0,1,1,1)'$, $(1,0,1,1,0)'$, and $(0,0,0,1,0)'$.

3.3.3 The Jensen-Shannon Divergence Index

Over the past decades, the Kullback-Leibler discrimination (KLD) index has been widely used in educational assessment (e.g., Chang & Ying, 1996; Henson & Douglas, 2005; Madigan & Almond, 1996; Veldkamp & van der Linden, 2002). It is a distance measure between two distributions over the same random variable. The KLD, which can be used with both discrete and continuous variables, is an alternative to the Fisher information index, which can only be used with continuous variables (Cheng, 2009). Thus, the KLD is more suitable for CDMs due to the discrete nature of attribute patterns. However, the KLD has some limitations in that it only compares two distributions, is not symmetric, and does not have a maximum value (i.e., it ranges from 0 to ∞) (Henson, Roussos, Douglas, & He, 2008). As an alternative to the KLD, the Shannon Entropy (SHE; Cover & Thomas, 2006) has been applied in the context of CDM (Xu,

Chang, & Douglas, 2003) to overcome some of the limitations of the KLD. For example, the SHE is symmetric (Lin, 1991), has a maximum value when the probabilities of distributions are equal (Xu et al., 2003), and can compare more than two distributions. In Xu et al.'s (2003) study, the SHE was described as an item selection index in cognitive diagnosis computerized adaptive testing (CD-CAT), where the authors found that the SHE provided higher correct classification rates than the KLD.

The JSD index is known as a measure of similarity between probability distributions (Gómez-Lopera, Martínez-Aroza, Robles-Pérez, & Román-Roldán, 2000; Lin, 1991), which measures the average distances among multiple probability distributions. Unlike the KLD, the JSD has better properties. Specifically, it is symmetric, bounded, and always well-defined with finite values (Castner, 2014). Furthermore, the JSD can offer more flexibility than the SHE to measure the spread of multiple distributions because a different weight to each probability distribution can be assigned (Lin, 1991). The JSD computation for multiple probability distributions is given by

$$JSD_{w_1,w_2,\ldots,w_n}(P_1,P_2,\ldots,P_n) = H(\sum_{i=1}^{n} w_i P_i) - \sum_{i=1}^{n} w_i H(P_i), \tag{3.4}$$

where $w_1, w_2, \ldots, w_n$ are weights that sum to 1; $P_1, P_2, \ldots, P_n$ are the probability functions; and $H(\cdot)$ is the SHE of the probability distribution expressed by

$$H(X) = -\int P(x) ln[P(x)] dx. \tag{3.5}$$

Higher values of the JSD imply a greater spread in the predicted class probability distributions, and it is zero if and only if the distributions are identical (Melville & Mooney, 2004). In this present study, the JSD was adapted for Q-matrix validation purposes based on binary attributes under the G-DINA model.

3.3.4 The iJSD for Q-Matrix Validation

The JSD index was first introduced in the context of CDM by Minchen and de la Torre (2016) for the continuous G-DINA (C-G-DINA) model. They adapted the JSD index as an item selection algorithm for continuous responses in CD-CAT. According to this study, the JSD provided higher attribute-wise and vector-wise classification rates than random item selection algorithm (Minchen & de la Torre, 2016). Yigit et al. (2016) further adapted the JSD for polytomous response data due to the complexity in its Q-matrix. Specifically, they proposed a CD-CAT item selection rule based on the JSD index for the MC-DINA model. They found that the JSD provided high attribute classification accuracy even with a short test or low quality items.

This current study modified the original JSD index to be used for Q-matrix validation purposes, where assuming the underlying process in the estimation of item parameters and posterior distributions is not required. For the purpose of empirically-based Q-matrix validation, the equation of the JSD for item j corresponding to a q-vector $\boldsymbol{q}_l$, $1 \leq l \leq 2^K$, can be expressed as follows

$$\begin{aligned} JSD_{jl}(P_1, P_2, \ldots, P_{2^K}) &= JSD_{jl}(P_1^{(l)}, P_2^{(l)}, \ldots, P_{2^{K_l}}^{(l)}) \\ &= H(\sum_{g=1}^{2^{K_l}} w_g^{(l)} P_g^{(l)}) - \sum_{g=1}^{2^{K_l}} w_g^{(l)} H(P_g^{(l)}), \end{aligned} \tag{3.6}$$

where $P_1, \ldots, P_{2^K}$ are the probabilities of success associated with the 2^K latent classes; K_l is the total number of attributes required in $\boldsymbol{q}_l$; $w_1^{(l)}, \ldots, w_g^{(l)}, \ldots, w_{2^{K_l}}^{(l)}$ and $P_1^{(l)}, \ldots, P_g^{(l)}, \ldots, P_{2^{K_l}}^{(l)}$ are posterior weights and success probabilities of each 2^{K_l} latent

group, respectively. $H(\cdot)$ is the SHE defined as

$$\begin{aligned} H(P_g^{(l)}) &= E[-ln(P_g^{(l)})] \\ &= -[P_g^{(l)} ln(P_g^{(l)}) + (1 - P_g^{(l)})(1 - ln(P_g^{(l)}))]. \end{aligned}$$

Higher values for the JSD imply a greater spread in the latent group success probabilities. The first term in Equation 3.6 is fixed for each studied item, and represents the SHE of the weighted sum of all success probabilities. The second term of Equation 3.6 is the sum of the weighted SHE of each probability. This shows the sum of success in latent groups. A higher value of JSD is related to a lower value of the second term of the JSD.

Furthermore, q-vectors with more attribute specifications usually have higher JSDs. Similar to the rationale in the GDI, in practice, real data introduce some noise, which can affect the quality of estimation, and ultimately the accuracy of posterior weights and success probabilities. This can result in choosing the full q-vector (i.e., $\boldsymbol{q} = \mathbf{1}$). Therefore, a q-vector with the lowest JSD that is within the confidence interval (CI) of the $\boldsymbol{q} = \mathbf{1}$ was chosen. The CI of the $\boldsymbol{q} = \mathbf{1}$ was obtained based on the variance of the JSD. Referring to each term of Equation 3.6, the mean of the SHE is the expected value of $ln(P_g^{(l)})$, the variance of the SHE would be the expectation of the square of $ln(P_g^{(l)})$ minus the square of the expectation of $ln(P_g^{(l)})$. The variance of the JSD was further standardized due to the fact that the size of each q-vector varies (i.e., 2^{K_l}). Thus, a q-vector with the fewest number of attribute specifications corresponding to the lowest JSD that is within the confidence interval of the $\boldsymbol{q} = \mathbf{1}$ can be chosen as the correct q-vector for item j.

Both the iJSD and iGDI were carried out in two steps. First, item parameters and posterior distributions of attribute patterns were estimated using the provisional Q-matrix. The estimates in item parameters were based on an empirical Bayesian implementation of the expected-maximization algorithm. Second, the $iJSD_{jl}$ and $iGDI_{jl}$

were computed to identify the correct q-vector for item j. If a Q-matrix entry is modified at the first iteration, these steps are repeated for both methods and a new calibration is carried out with the updated Q-matrix as the new provisional Q-matrix. A stopping rule for the iterative cycle occurs when it stops suggesting additional changes in attribute specifications. The code for this study was written in Ox (Doornik, 2009).

3.4 Design and Analysis

This study proposed two methods, the iJSD and iGDI, for Q-matrix validation under a situation where a specific reduced CDM is unknown when calibrating the data. In this present paper, the simulated data were only generated from the DINA model because it has a more straightforward interpretation and requires a smaller sample size for accurate parameter estimates (Rojas, de la Torre, & Olea, 2012). Therefore, using the DINA model, 100 datasets were simulated under a number of conditions. Specifically, sample sizes (N = 1,000 and 2,000), complexity of q-vectors considered in the Q-matrix (see Table 3.1 for $J = 30$, and 31), item qualities ($s_j = g_j = 0.1$, 0.2, and 0.3), and amount of misspecifications (5% and 10%) were controlled. In each condition, 100 misspecified Q-matrices, which consist of 5% and 10% misspecified q-entries, were randomly generated from the true Q-matrices. Moreover, attribute structures were generated from uniform and higher-order distributions (HO; de la Torre & Douglas, 2004). In the uniform distribution, all the possible attribute patterns are equally likely, whereas, in the HO distribution, mastery or nonmastery of an attribute k is related to a unidimensional latent variable θ_i for examinee i. The probability of mastering α_k as a function of θ_i can be formulated as in

$$P(\alpha_k|\theta_i) = \frac{exp(\lambda_{0k} + \lambda_{1k}\theta_i)}{1 + exp(\lambda_{0k} + \lambda_{1k}\theta_i)}, \tag{3.7}$$

where $\boldsymbol{\lambda}_0 = \{-1.0, -0.5, 0.0, 0.5, 1.0\}$ and $\lambda_1 = 1.0$ are the attribute difficulty and discrimination parameters, respectively. The ability of examinee i, θ_i, was drawn from $N(0,1)$.

Table 3.1: True Q-matrix for the Simulated Data

Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0	11	1	1	0	0	0	21	1	1	1	0	0
2	0	1	0	0	0	12	1	0	1	0	0	22	1	1	0	1	0
3	0	0	1	0	0	13	1	0	0	1	0	23	1	1	0	0	1
4	0	0	0	1	0	14	1	0	0	0	1	24	1	0	1	1	0
5	0	0	0	0	1	15	0	1	1	0	0	25	1	0	1	0	1
6	1	0	0	0	0	16	0	1	0	1	0	26	1	0	0	1	1
7	0	1	0	0	0	17	0	1	0	0	1	27	0	1	1	1	0
8	0	0	1	0	0	18	0	0	1	1	0	28	0	1	1	0	1
9	0	0	0	1	0	19	0	0	1	0	1	29	0	1	0	1	1
10	0	0	0	0	1	20	0	0	0	1	1	30	0	0	1	1	1

Note that we are not interested in differences between the performance across two different test lengths (i.e., $J = 30$ and 31), instead, across different levels of complexity in the Q-matrix (i.e., 1-, 2-, and 3-attribute q-vectors, and 1-, 2-, ..., and 5-attribute q-vectors, where the latter one is much more complex). All the possible attribute combinations are specified in the 31-item Q-matrix for $K = 5$, excluding the $\boldsymbol{q} = \mathbf{0}$. Therefore, this can allow us to investigate whether or not including all the possible combinations of attribute specifications can provide a higher recovery of attribute misspecifications. The number of attributes K was fixed to 5.

3.4.1 Results

Results based on the false-positive and true-positive rates of the iJSD and iGDI were reported. The false-positive rate (i.e., Type-I error) is the proportion of correctly specified q-vectors that are modified; and the true-positive rate (i.e., power) is the proportion of misspecified q-vectors that are correctly identified (de la Torre & Chiu,

2016). For the iGDI, in addition to the cut-off value for PVAF set at $\varepsilon = 0.95$ in the de la Torre and Chiu's (2016) study, this current study further investigated results for $\varepsilon = 0.90$ and 0.99.

Results were reported in Tables 3.2 and 3.3. In each table, findings were also split into two parts based on attribute structures generated from uniform and HO distributions. Note that optimum results of the iGDI were obtained when $\varepsilon = 0.95$ for the high quality item and $\varepsilon = 0.90$ for the medium and low quality items.

Table 3.2 reports the false-positive rates of the iJSD and iGDI. When attributes were uncorrelated (i.e., uniform), the false-positive rates of the two methods were around zero with at least medium quality items. When the item quality was low, the iGDI showed lower false-positive rates than the iJSD, except when $N = 1,000$ and $J = 31$. A larger sample size resulted in lower false-positive rates for both methods. In general, the false-positive rates were lower when $J = 30$ than $J = 31$. When attributes were correlated (i.e., HO), the false-positive rates increased for the high and medium quality items. Even though the iJSD had higher inflation of the false-positive rates, it was still around the nominal level. However, when the item quality was low, the iJSD had lower false-positive rates under the HO distribution than that under the uniform distribution. Similarly, the iGDI had lower false-positive rates under the HO distribution than that under the uniform distribution, in particular, when $N = 1,000$ with the low quality item. Again, increasing the sample size lowered the false-positive rates for both methods, and the false-positive rates were generally lower when $J = 30$ than $J = 31$.

Based on the attribute structure generated from the uniform distribution, the iGDI provided equally well as or better true-positive rates than the iJSD when the item quality was high, except when $N = 1,000$ and $J = 31$ with 10% misspecifications, where the iJSD (0.98) had a higher true-positive rate than the iGDI (0.95). Given the medium quality item, the iJSD had higher true-positive rates than the iGDI under two conditions where $N = 1,000$ and $J = 30$. Specifically, when $N = 2,000$ and $J = 30$, both methods

Table 3.2: False-Positive Rate of the iJSD and iGDI

Quality	%	iJSD				iGDI			
		N= 1,000		*N* = 2,000		*N*= 1,000		*N* = 2,000	
		J = 30	*J* = 31	*J* = 30	*J* = 31	*J* = 30	*J* = 31	*J* = 30	*J* = 31
		Uniform							
H	5	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00
	10	0.02	0.00	0.00	0.00	0.01	0.02	0.00	0.00
M*	5	0.00	0.02	0.00	0.02	0.00	0.01	0.00	0.00
	10	0.00	0.03	0.00	0.03	0.00	0.01	0.00	0.00
L*	5	0.18	0.33	0.09	0.26	0.08	0.44	0.00	0.00
	10	0.24	0.39	0.14	0.47	0.19	0.52	0.00	0.00
		Higher-Order							
H	5	0.06	0.04	0.04	0.03	0.00	0.00	0.00	0.00
	10	0.06	0.04	0.03	0.04	0.00	0.00	0.00	0.01
M*	5	0.05	0.05	0.02	0.04	0.00	0.00	0.00	0.00
	10	0.05	0.09	0.02	0.08	0.00	0.04	0.02	0.04
L*	5	0.09	0.07	0.02	0.04	0.01	0.01	0.00	0.00
	10	0.11	0.11	0.02	0.10	0.02	0.05	0.00	0.06

Note. * indicates that the iGDI provided higher recovery when $\varepsilon = 0.90$ for the medium and low quality items.

provided perfect true-positive rates with at least medium quality items. However, the iGDI had a higher true-positive rate with 31-item tests. For the high and medium quality items, the true-positive rates were generally acceptable (> 0.8) for both methods, except when $N = 1,000$ and $J = 31$ with both 5% and 10% misspecifications for the iJSD, and with 10% misspecifications for the iGDI. Furthermore, when the item quality was low, the true-positive rates were not acceptable (< 0.8) for both methods. Similar to the findings of the false-positive rates in the preceding paragraph, the true-positive rates were higher when $J = 30$ than $J = 31$ for both methods.

Table 3.3: True-Positive Rate of the iJSD and iGDI

Quality	%	iJSD				iGDI			
		N= 1,000		*N* = 2,000		*N*= 1,000		*N* = 2,000	
		J = 30	*J* = 31	*J* = 30	*J* = 31	*J* = 30	*J* = 31	*J* = 30	*J* = 31
		Uniform							
H	5	0.98	0.98	1.00	0.99	0.99	1.00	1.00	1.00
	10	0.98	0.98	1.00	0.99	0.99	0.95	1.00	1.00
M*	5	0.98	0.77	1.00	0.88	0.92	0.81	1.00	0.97
	10	0.98	0.70	1.00	0.83	0.93	0.79	1.00	0.96
L*	5	0.36	0.16	0.56	0.28	0.41	0.16	0.75	0.40
	10	0.35	0.13	0.53	0.21	0.33	0.15	0.69	0.38
		Higher-Order							
H	5	0.93	0.96	0.98	0.97	1.00	1.00	1.00	1.00
	10	0.94	0.95	0.97	0.95	1.00	0.99	1.00	0.99
M*	5	0.95	0.95	0.98	0.97	1.00	0.99	1.00	1.00
	10	0.95	0.88	0.98	0.90	0.99	0.94	1.00	0.93
L*	5	0.89	0.84	0.98	0.95	1.00	0.93	1.00	0.97
	10	0.88	0.77	0.97	0.86	0.97	0.87	0.99	0.89

Note. * indicates that the iGDI provided higher recovery when $\varepsilon = 0.90$ for the medium and low quality items.

Results for the true-positive rates were more stable when attributes were generated from HO distribution. Overall, the true-positive rates were acceptable (> 0.8) for both methods, except for a condition of the low quality item when $N = 1,000$ and $J = 31$ with 10% misspecifications for the iJSD. Even though the iGDI outperformed the iJSD under the HO distribution, both methods had the true-positive rates above 0.90, in particular, when $N = 2,000$. There was only one exception observed for both methods – the true-positive rates were below 0.90 when $J = 31$ with the low quality item including 10% misspecifictions, which was still above the acceptable level. Similar to the previous findings, the true-positive rates were almost always higher when $J = 30$ than $J = 31$ for both methods.

3.5 Implementation with Real Data

The iJSD and iGDI were further investigated using the fraction-subtraction data (Tatsuoka, 1984). This real data analysis can help compare the performance of the iJSD and iGDI in practice. The fraction-subtraction test used in this study consists of 12 items administered to 536 middle school students. Only four attributes were included in the analysis: α_1 – performing a basic fraction subtraction operation; α_2 – simplifying/reducing; α_3 – separating a whole number from fraction; and α_4 – borrowing one from a whole number to fraction. The 12 items with the corresponding attribute specifications are shown in Table 3.4. Results from the two methods were examined based on suggested q-vectors under a saturated model (i.e., the G-DINA) with an assumption that the specific reduced model is unknown when calibrating the data.

3.5.1 Results

Findings based on the iJSD and iGDI are displayed in Table 3.5. Results indicated that the iJSD was more stringent than the iGDI because the iJSD had fewer 1s than the iGDI. The iJSD suggested to change 10 attribute specifications in items 1, 2, 4, 5, 6, 7, 8, 9, 11, and 12 and the iGDI suggested to change six attribute specifications in items 1, 5, 6, 11, and 12 when $\varepsilon = 0.95$.

Item 1 originally required only α_1, which is not necessarily true because the solution requires more than "performing a basic fraction subtraction operation." That is,

$$\begin{aligned}\frac{3}{4}-\frac{3}{8} &= \frac{2\times 3}{2\times 4}-\frac{3}{8}\\ &= \frac{6-3}{8}=\frac{3}{8},\end{aligned}$$

meaning that an attribute "finding a common denominator" was necessary to solve item

Table 3.4: Q-Matrix for Fraction-Subtraction Items

		Attributes			
Item		α_1	α_2	α_3	α_4
1.	$\frac{3}{4}-\frac{3}{8}$	1	0	0	0
2.	$3\frac{1}{2}-2\frac{3}{2}$	1	1	1	1
3.	$\frac{6}{7}-\frac{4}{7}$	1	0	0	0
4.	$3\frac{7}{8}-2$	1	0	1	0
5.	$4\frac{4}{12}-2\frac{7}{12}$	1	1	1	1
6.	$4\frac{1}{3}-2\frac{4}{3}$	1	1	1	1
7.	$\frac{11}{8}-\frac{1}{8}$	1	1	0	0
8.	$3\frac{4}{5}-3\frac{2}{5}$	1	0	1	0
9.	$4\frac{5}{7}-1\frac{4}{7}$	1	0	1	0
10.	$7\frac{3}{5}-\frac{4}{5}$	1	0	1	1
11.	$4\frac{1}{10}-2\frac{8}{10}$	1	1	1	1
12.	$4\frac{1}{3}-1\frac{5}{3}$	1	1	1	1

Note. α_1 – performing a basic fraction subtraction operation; α_2 – simplifying/reducing; α_3 – separating a whole number from fraction; and α_4 – borrowing one from a whole number to fraction.

1 correctly. Despite being different from this attribute, α_2 was suggested by the iJSD, and α_3 and α_4 were suggested by the iGDI on the top of α_1. A explanation could be due to the fact that the attribute "finding a common denominator" was not included in the Q-matrix used in this study.

For items 4, 8, and 9, the iJSD recommended to exclude α_3, however, the iGDI retained α_3. Furthermore, the iJSD excluded α_2 from item 7, whereas, the iGDI concurred with the original Q-matrix, requiring α_1 and α_2. The iJSD and iGDI also indicated α_2 unnecessary in items 5, 6, 11, and 12, and that in item 2 for the iJSD only. These suggestions could be, among others, due to the employment of another strategy to solve the problems. For example, items 6, 11, and 12 can be answered correctly without mastering α_2 (i.e., simplifying/reducing), such that, – borrowing one from a

whole number to fraction, separating a whole number from fraction, and performing a basic fraction – happens to give the correct answer. The following example shows the strategy to solve item 11

$$
\begin{aligned}
4\frac{1}{10} - 2\frac{8}{10} &= 3\frac{(1\times 10)+1}{10} - 2\frac{8}{10} \\
&= (3-2)\frac{11-8}{10} = 1\frac{3}{10},
\end{aligned}
$$

meaning that item 11 can be answered correctly with another strategy, which does not require to use the same attributes specified in the original Q-matrix.

Table 3.5: Suggested Q-Matrix by the iJSD and iGDI for the G-DINA model

	iJSD				iGDI			
Item	α_1	α_2	α_3	α_4	α_1	α_2	α_3	α_4
1.	1	1*	0	0	1	0	1*	1*
2.	1	0*	1	1	1	1	1	1
3.	1	0	0	0	1	0	0	0
4.	1	0	0*	0	1	0	1	0
5.	1	0*	1	1	1	0*	1	1
6.	1	0*	1	1	1	0*	1	1
7.	1	0*	0	0	1	1	0	0
8.	1	0	0*	0	1	0	1	0
9.	1*	0	0*	0	1	0	1	0
10.	1	0	1	1	1	0	1	1
11.	1	0*	1	1	1	0*	1	1
12.	1	0*	1	1	1	0*	1	1

Note. α_1 – performing a basic fraction subtraction operation; α_2 – simplifying/reducing; α_3 – separating a whole number from fraction; and α_4 – borrowing one from a whole number to fraction; iGDI results were obtained when $\varepsilon_j = 0.95$; * indicates suggested attribute specifications.

3.6 Summary and Discussion

This paper proposed two new methods, the iJSD and iGDI, for empirically validating the Q-matrix for the DINA model. Both methods can be used for dichotomous models, however, the iJSD is more general than the iGDI that can also be applied under nondichotomous models such as the C-G-DINA and MC-DINA. Furthermore, the iGDI, an extension of the original GDI, provided better results because the iterative procedure was applied to the original GDI.

Results showed that the iJSD and iGDI can identify misspecified q-entries at a high rate, in particular, when the item quality was at least medium. Under favorable conditions, the false-negative rate was around the nominal level. Generally, results were stable for the iJSD, and similar to those of the iGDI, specifically, when the item quality was high and medium.

When $J = 30$, both procedures provided better results than $J = 31$. At least based on the findings in this study, including more single-attribute q-vectors provided higher recovery of misspecified entries than fewer single-attribute items. In other words, more complexity of q-vectors considered in the Q-matrix for $J = 31$ in comparison to $J = 30$ did not provide better results.

Attribute structures also affected the results of the iJSD and iGDI. When attributes were uncorrelated (i.e., uniform), both procedures showed lower false-positive rates for the high and medium quality items, which was inflated with the low quality items. In terms of the true-positive rates, the iJSD and iGDI were not too different with at least medium quality items. But, both procedures provided low true-positive rates with the low quality items.

Given correlated attributes (i.e., HO), the true-positive and false-positive rates were more stable. The iGDI showed lower false-positive rates, which was around the nominal level for the iJSD. The iJSD and iGDI presented quite high true-positive rates above

the acceptable level, but the iGDI was generally better.

The running time of the code was shorter for the iJSD than the iGDI. For example, it took the code 36.88 and 102.17 minutes using a 3.50-GHz I7 computer to run the iJSD and iGDI, respectively, for 100 iterations under the condition where $N = 2,000$, $J = 31$, and medium quality items with 10% misspecifications under a uniform distribution. The number of iterations did not go beyond four.

A reason for different attribute suggestions based on the procedures could be because the fraction subtraction data have a small sample size and test item, and the number of attributes is smaller than that of the simulation study. Thus, it would be interesting to investigate how the iJSD and iGDI could behave under a small number of sample sizes, test items, and attributes. More discussions about multiple strategies in cognitive diagnosis using the fraction subtraction data can be found in de la Torre and Douglas (2008), Hou, de la Torre, and Nandakumar (2014), and Mislevy (1996).

In the future, the iJSD and iGDI should be investigated in depth to see how they would work when other reduced CDMs are involved. Furthermore, relaxing the assumption that the number of attributes is fixed (i.e., $K = 5$) could provide a broader perspective for the performance of the procedures. Finally, both indices should be improved particularly when the low quality items are involved in a test.

The main idea of this study is not to replace existing validation procedures but rather to serve as a supplementary tool from a statistical point of view. Moreover, employing both existing validation methods and domain experts may be a more appropriate process for Q-matrix validation.

3.7 References

Castner, J. A. (2014). Measures of cognitive distance and diversity. *Available at SSRN: https://ssrn.com/abstract=2477484*.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213–229.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*, 619–632.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*, 598–618.

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

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. Hoboken, New Jersey: John Wiley & Sons, Inc.

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

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

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*, 163–183.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253–273.

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

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*, 595–624.

Doornik, J. A. (2009). An object-oriented matrix programming language Ox 6. [Computer software]. London, UK: Timberlake Consultants Ltd.

Gómez-Lopera, J. F., Martínez-Aroza, J., Robles-Pérez, A. M., & Román-Roldán, R. (2000). An analysis of edge detection by using the jensen-shannon divergence. *Journal of Mathematical Imaging and Vision*, *13*, 35–56.

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

Hartz, S., Roussos, L., Henson, R., & Templin, J. (2005). *The fusion model for skill diagnosis: Blending theory with practicality. Unpublished manuscript.*

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

Henson, R. A., Roussos, L., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement*, *32*, 275–288.

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*, 191–210.

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*, 98–125.

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.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, *37*, 145–151.

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

Madigan, D., & Almond, R. G. (1996). On test selection strategies for belief networks. In *Learning from data* (pp. 89–98). New York: Springer.

Melville, P., & Mooney, R. J. (2004). Diverse ensembles for active learning. In *Proceedings of the twenty-first international conference on machine learning* (pp. 584–591). Morgan Kaufmann.

Minchen, N., & de la Torre, J. (2016). *The continuous G-DINA model and the Jensen-Shannon Divergence*. Paper presented at the international meeting of Psychometric Society, Asheville, NC.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*, 379–416.

Rojas, G., de la Torre, J., & Olea, J. (2012). Choosing between general and specific cognitive diagnosis models when the sample size is small. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

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

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

Tatsuoka, K. K. (1984). *Analysis of errors in fraction addition and subtraction problems*. Computer-based Education Research Laboratory, University of Illinois.

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

Terzi, R., & de la Torre, J. (2015). *An iterative method of empirically-based Q-matrix validation*. Paper presented at the annual aeeting of the National Council on Measurement in Education, Chicago, IL.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575–588.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.

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

Yigit, H., Sorrel, M., & de la Torre, J. (2016). *Computerized adaptive testing for cognitively based multiple-choice data*. Paper presented at the VII European Congress of Methodology, Mallorca, Spain.

Chapter 4

Study III: The Wald Test for Empirical Q-Matrix Validation

4.1 Abstract

Cognitive diagnosis models (CDMs) have the advantage of providing finer-grained information based on examinee responses to a cognitively diagnostic assessment. The Q-matrix, which specifies the required attributes for each item, is an important component of these models. However, specifying the Q-matrix is an inherently subjective process. Over- or under-specifying one or more entries in the Q-matrix may negatively affect item parameter estimates, and lead to examinee misclassifications. Thus, the misspecification of Q-matrix entries is of serious validity concern. To ensure the validity of inferences from CDMs, Q-matrices developed by experts need to be validated. This study proposes the Wald-Q – a Wald test-based Q-matrix validation method. Simulation studies are carried out to examine the false-positive and true-positive rates of the Wald-Q in comparison to the IMSSA and iGDI under various conditions. Results show that the Wald-Q can identify misspecified q-entries at a high rate, especially when the test is long, and the false-positive rate is around the nominal level under favorable conditions. The Wald-Q is also applied to fraction-subtraction data. The study concludes with the strengths and limitations of the Wald-Q, and suggestions for future studies.

4.2 Introduction

Recently, interest in formative assessment based on cognitive diagnosis models (CDMs) has been growing. Traditional test theories used for summative assessments, such as classical test theory (CTT) and item response theory (IRT), have some limitations. In particular, a single overall ability score based on CTT or IRT does not provide useful diagnostic information about specific skills (e.g., Leighton & Gierl, 2007). This limitation has led researchers (Akbay, Terzi, Kaplan, & Karaaslan, 2017; Tatsuoka, 1984; Tjoe & de la Torre, 2014) in the testing field to develop the latent construct as finer-grained and interrelated, but separable latent skills within a domain of interest. Compared to traditional item response models, CDMs provide more specific information relevant to classroom instruction and learning. In other words, CDMs can be used to determine examinees' mastery profiles that can be used for targeted instruction. At their core, CDMs specify the relationship, or interaction, between skills or attributes and tasks.

The need to discover more specific skills has led researchers to develop different reduced and general CDMs, which have been described in the literature. Examples of reduced models include the deterministic inputs, noisy "and" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model, the deterministic input, noisy "or" gate (DINO; Templin & Henson, 2006) model, the compensatory and reduced reparameterized unified model (C-RUM and R-RUM; Hartz, Roussos, Henson, & Templin, 2005), the *additive* CDM (*A*-CDM; de la Torre, 2011), and the linear logistic model (LLM; de la Torre & Douglas, 2004). Examples of general models include the general diagnosis model (GDM; von Davier, 2008), the log-linear CDM (Henson, Templin, & Willse, 2009), and the generalized DINA model (G-DINA; de la Torre, 2011). For the most part, reduced CDMs can be expressed as special cases of general CDMs. Regardless of different formulations, the Q-matrix (Tatsuoka, 1983) is a common component in all

these CDMs.

The Q-matrix is needed for CDMs to relate the specific subset of attributes to each item. It is generally a binary $J \times K$ matrix where the jth item can be correctly answered if the required attributes for item j have been mastered by the examinee. The row j and column k entry of the Q-matrix, q_{jk}, is 1 if the kth attribute is required to correctly answer item j, and is 0 if it is not required. The process of constructing the Q-matrix partly involves domain experts. As expected, the process can be questionable because of the inherent subjectivity of the expert judgments and has caused serious validity discussions among researchers (e.g., Chiu, 2013; de la Torre, 2008; Liu, Xu, & Ying, 2012; Rupp & Templin, 2008). Including or omitting multiple attribute specifications in the Q-matrix may have a serious impact on the accuracy of the attribute classifications because of the inaccurate estimation of the model parameters (de la Torre & Chiu, 2016). Additionally, any model-fit analysis becomes unreliable if the correctness of the Q-matrix is not checked. However, the Q-matrix is usually assumed to be correct after it has been specified by domain experts because of the lack of well-established methods to verify attribute specifications in the Q-matrix (Chiu, 2013; DeCarlo, 2011; de la Torre, 2008), particularly when general CDMs are involved (de la Torre & Chiu, 2016; Liu et al., 2012).

To date, only few novel statistical methods have been developed for validating attribute specifications in the Q-matrix. For instance, DeCarlo (2011) proposed a model-based approach where misspecified q-vectors are treated as random variables and estimated with the rest of the model parameters. However, aside from being computationally time-consuming, this method sometimes cannot identify the misspecified q-vectors, and requires any misspecified q-vectors to be identified in advance. From a different perspective, Liu et al. (2012) proposed a data-driven approach to identify the Q-matrix and estimate related model parameters. This approach is based on examinees' responses without involving any input from experts. However, among other

limitations, the identifiability of the Q-matrix may be weaker under some conditions, such as if there is an unknown guessing parameter (Liu et al., 2012). Additionally, since this method has only been used with specific CDMs, further investigation is needed to see if it can be used when the underlying process is unknown.

Chiu (2013) developed another Q-matrix refinement method (QRM) based on a nonparametric classification procedure (Chiu & Douglas, 2013). The QRM minimizes the residual sum of squares (RRS) between the observed responses and the ideal responses among all the possible q-vectors of a given Q-matrix. The algorithm recovers the Q-matrix by replacing the q-vector of the highest RSS (i.e., the worst) with the q-vector of the lowest RSS (i.e., the best). The algorithm iterates until convergence is met. One issue of the QRM is that it requires specifying in advance whether the underlying process is conjunctive or disjunctive method. Additionally, due to its nature as a nonparametric method, parametric methods should provide more powerful results, if the underlying model is known, particularly when N is large.

Another method, which was proposed by de la Torre (2008), is an empirically based δ-method implemented through a sequential search algorithm for the DINA model. In this method, the correct q-vector among all the possible $2^K - 1$ q-vectors, excluding the $\boldsymbol{q} = \mathbf{0}$, is defined based on the δ_{jl} value, the discrimination index of item j, when q_l, the lth q-vector, is used. The index is computed as differences in the probabilities of correct responses between examinees who have the required attributes (i.e., $\eta_j = 1$) and those who do not (i.e., $\eta_j = 0$). Among all the possible q-vectors, the correct specification of a q-vector should maximize differences between the success probabilities of the two groups (i.e., $\eta_j = 1$ and $\eta_j = 0$). However, this method has some limitations. First, uncertainty remains in defining ε values, which are applied to prevent over or under corrections. In practice, it is unclear how a formal single value for a threshold (i.e., ε) can be determined. The value can be more liberal or more stringent depending on

many factors, such as sample sizes, test lengths, item qualities, and amount of misspecifications, which were all fixed in de la Torre's (2008) study. Additionally, the amount of misspecifications was limited to 5 of 150 attributes in the Q-matrix ($J = 30, K = 5$), which is approximately 3.3% misspecifications in total.

Given these issues in de la Torre's (2008) paper, Terzi and de la Torre (2015) introduced an empirically based iterative δ-method for Q-matrix validation as an extension of the empirically based δ-method (de la Torre, 2008) to address some of the concerns in de la Torre's (2008) work. As the name suggests, the method involved an iterative algorithm, which offered additional improvements in identifying attribute specifications. Moreover, optimal values of ε across different conditions were defined based on the estimated item qualities. The effectiveness of the algorithm was further improved by focusing on single-attribute specifications (Terzi & de la Torre, 2015). Therefore, the K single-attribute vectors were enough to determine correct attribute specifications. However, this method was only examined using a uniform attribute structure and the DINA model.

The empirically based δ-method (de la Torre, 2008) was recently expanded to a wider class of CDMs by de la Torre and Chiu (2016) by developing a method based on the G-DINA model (de la Torre, 2011). Even though the discrimination index (i.e., ς_j^2) has greater applicability under the generalized model, the findings have limited generality due to the fixed sample size and test length examined in the study. As with the δ-method, de la Torre and Chiu's (2016) work does not prescribe a formal way for defining optimal ε values. Furthermore, the method does not include an iterative algorithm, as in, it stops validating attribute specifications at the first step.

The primary objective of this study is to propose a new empirically based method for validating the correctness of attribute specifications in a provisional Q-matrix. In particular, the method, Wald-Q, adapts the Wald test (Morrison, 1967) for multivariate hypothesis testing that performs all $2^K - 2$ tests for one item at a time. The Wald-Q

should eliminate some of the limitations of the existing methods. For instance, although the method is an empirically based procedure similar to that used by de la Torre (2008), Terzi and de la Torre (2015), and de la Torre and Chiu (2016), the use of a single optimal ε value is not required. Furthermore, the Wald-Q includes an iterative process that continues validating attribute specifications after the first step as long as any changes occur in attribute specifications. Using the updated Q-matrix, a new calibration is carried out to eliminate any potential effect of misspecified entries at the succeeding steps. If no changes occur in attribute specifications, the iterative process is terminated. Moreover, the Wald-Q can be designed for specific and general CDMs based on the restriction matrix. Even though assuming a reduced or saturated model results in a different restriction matrix, the data were always calibrated using the G-DINA model. The performance of the Wald-Q was further compared to the *iterative* modified sequential search algorithm (IMSSA; Terzi & de la Torre, 2015) and *iterative* G-DINA model discrimination index (iGDI; Terzi & de la Torre, in preparation) in the context of the DINA and G-DINA models, respectively. The iGDI is an extension of the original GDI with an iterative algorithm in that better results were obtained (Terzi & de la Torre, in preparation).

The rest of the paper consists of the following sections. In the second section, background is provided about the G-DINA and DINA models, the G-DINA model discrimination index, the Wald test, and the Wald test for Q-matrix validation. The next section investigates the viability of the method by examining the false-positive and true-positive rates of the Wald-Q in comparison to the IMSSA and iGDI. An empirical example is included to examine the procedure with real data. The paper concludes with a summary and discussion, and suggestions for further research.

4.3 Background

4.3.1 G-DINA and DINA Models

The G-DINA model is one of the most commonly used generalized CDMs. The G-DINA model creates 2^{K_j} latent groups, where K_j is the total number of the required attributes for item j, as $K_j = \sum_{k=1}^{K} q_{jk}$ (de la Torre, 2011). Assuming the first $1, \cdots, K_j$ attributes to be required for item j, the reduced attribute vector $\boldsymbol{\alpha}_{lj}^*$ can be used to represent the columns of the required attributes, where $l = 1, \cdots, 2^{K_j}$. The probability of correctly answering item j by examinees with attribute pattern $\boldsymbol{\alpha}_{lj}^*$ will be denoted by $P(X_j = 1|\boldsymbol{\alpha}_{lj}^*) = P(\boldsymbol{\alpha}_{lj}^*)$. The item response function (IRF) of the G-DINA model for the identity link is given by

$$P(\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j} \sum_{k=1}^{K_j-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{j12\ldots K_j} \prod_{k=1}^{K_j} \alpha_{lk}, \qquad (4.1)$$

where δ_{j0} is the intercept for item j, δ_{jk} is the main effect due to α_k, $\delta_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$, and $\delta_{j12\ldots K_j}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K_j}$.

As a saturated model, the G-DINA model subsumes several commonly-used reduced CDMs, such as the DINA model, the DINO model, the A-CDM, the LLM, and the R-RUM. Applying appropriate parameterization, these reduced models can be derived from the G-DINA model under different constraints and link functions (de la Torre, 2011). In this present paper, among these reduced models, the DINA model was used because it has more straightforward interpretations, requires smaller sample sizes for accurate parameter estimation (Rojas, de la Torre, & Olea, 2012), and is flexible for extension to more general CDMs. By setting all the parameters in Equation 4.1 to zero, except for δ_{j0} and $\delta_{j12\ldots K_j}$, the IRF of the DINA model can be formulated as

$$P(\alpha_{lj}^{*}) = \delta_{j0} + \delta_{j12\ldots K_j} \prod_{k=1}^{K_j} \alpha_{lk}. \tag{4.2}$$

4.3.2 The G-DINA Model Discrimination Index

The empirically-based sequential δ-method (de la Torre, 2008) was proposed to empirically validate a Q-matrix in conjunction with the DINA model. Because of the limitation in the parameterization of the DINA model, a more general discrimination index was needed for Q-matrix validation purposes. The idea of the δ-method for empirically validating the correctness of attribute specifications was extended to the G-DINA model, where a statistic based on 2^{K_j} groups is computed.

De la Torre and Chiu (2016) introduced the G-DINA model discrimination index (GDI), denoted by ς_j^2, which is a Q-matrix validation index for general CDMs. In the paper, they discussed a theorem to justify the use of the index for Q-matrix validation with the G-DINA model. The suggested q-vector is determined based on the proportion of variance accounted for (PVAF) by a q-vector relative to the maximum $\hat{\varsigma}_j^2$ under which all attributes are specified (de la Torre & Chiu, 2016). In particular, the q-vector with the fewest attribute specifications corresponding to ς_j^2 that approximates the maximum GDI is suggested. The approximation was done with a predetermined cutoff value for PVAF set at $\varepsilon = 0.95$ (de la Torre & Chiu, 2016).

Given a particular attribute distribution, the ς_j^2 measures the weighted variance of the probabilities of correctly answering item j. Let the first K_j attributes be required.

The GDI of an item with the specification $\mathbf{q}_{K':K''}$ is defined as

$$\begin{aligned}\varsigma^2 = \varsigma^2_{K':K''} &= \sum_{\alpha_{K'}=0}^{1} \cdots \sum_{\alpha_{K''}=0}^{1} w(\alpha_{K':K''})[p(\alpha_{K':K''}) - \overline{p}(\alpha_{K':K''})]^2 \\ &= \sum_{\alpha_{K'}=0}^{1} \cdots \sum_{\alpha_{K''}=0}^{1} w(\alpha_{K':K''})p^2(\alpha_{K':K''}) - \overline{p}^2(\alpha_{K':K''}),\end{aligned} \tag{4.3}$$

where $\overline{p}(\alpha_{K':K''}) = \sum_{\alpha_{K'}=0}^{1} \cdots \sum_{\alpha_{K''}=0}^{1} w(\alpha_{K':K''})p(\alpha_{K':K''})$ is the weighted probability of success across all the $2^{K''-K'+1}$ possible patterns of $p(\alpha_{K':K''})$; and $w(\alpha_{K':K''})$ is the posterior probability of examinees being in class $(\alpha_1, \ldots \alpha_{K''})$, which is equal to $\sum_{\alpha_1=0}^{1} \cdots \sum_{\alpha_{K'}-1=0}^{1} w(\alpha_{1:K''})$ (de la Torre & Chiu, 2016).

4.3.3 The Wald Test

The Wald test (Morrison, 1967) has been a popular statistical test for decades. In comparing the Wald, Lagrange multiplier (LM), and likelihood ratio (LR) tests, Buse (1982) underscored that the Wald test has the advantage that it only requires estimating the larger (i.e., unrestricted) model, and does not require derivatives. In contrast, the LR test requires estimating both (i.e. unrestricted and restricted) models, whereas, the LM test requires obtaining the derivatives to carry out the test. Recently, Sorrel, Abad, Olea, de la Torre, and Barrada (2017) compared the performance of four inferential item-fit statistics (i.e., Wald, LR, LM tests, and $S-X^2$) in the context of CDM comparison, and observed that the Wald and LR tests performed better than the LM and $S-X^2$.

Several studies have applied the Wald test in the context of CDMs (de la Torre, 2011; de la Torre & Lee, 2013; Hou, de la Torre, & Nandakumar, 2014; Ma, Iaconangelo, & de la Torre, 2016). The Wald test for CDM applications, first introduced by de la Torre (2011), was used to examine whether the G-DINA model can be replaced by one of the reduced models (i.e., DINA, DINO, or A-CDM). The null hypothesis to test the fit of a reduced model with $p < 2^{K_j}$ parameters can be written as $\boldsymbol{R}_{jp} \times \boldsymbol{P}_j = 0$,

where $\boldsymbol{P}_j = \{P(\boldsymbol{\alpha}^*_{lj})\}$, and $\boldsymbol{R}_{jp}$ is the $(2^{K_j} - p) \times 2^{K_j}$ matrix of restrictions. The Wald statistic W_j to test the null hypothesis for item j is computed as

$$W_j = [\boldsymbol{R}_{jp} \times \boldsymbol{P}_j]'[\boldsymbol{R}_{jp} \times Var(\boldsymbol{P}_j) \times \boldsymbol{R}'_{jp}]^{-1}[\boldsymbol{R}_{jp} \times \boldsymbol{P}_j], \tag{4.4}$$

where Var($\boldsymbol{P}_j$) is the variance-covariance matrix of the item parameters for the saturated model computed from the inverse of the information matrix. Under the null hypothesis for the DINA model, the Wald statistic is assumed to be an asymptotically χ^2 distributed with $(2^{K_j} - p)$ degrees of freedom. For example, the $\boldsymbol{R}_{jp}$ matrix for the DINA model when $K_j = 3$ is

$$\boldsymbol{R}_{jp6\times 8} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \end{pmatrix},$$

which constrains all the parameters in Equation 4.1 to zero except for δ_{j0} and $\delta_{j12\ldots K_j}$ (de la Torre, 2011).

The Wald test was further applied by de la Torre and Lee (2013) to investigate the most appropriate CDM at the item level by comparing the fit of a saturated model against the fits of reduced models. They examined the performance of the Wald test with both simulated and real data analyses. The simulation study showed that the Wald test had excellent power to identify the true underlying model even for small sample sizes, while controlling the Type-I error for large sample sizes with a small number of attributes. Furthermore, the Wald test was used to examine differential item functioning (DIF) in the context of CDMs (Hou et al., 2014). The viability of the Wald test to detect

both uniform and nonuniform DIF in the DINA model was explored via a simulation study. The study showed Type I error rates close to the nominal level, and medium to high power rates for reference item parameter values less than 0.3. The Wald test application in the de la Torre and Lee's (2013) study was extended by Ma et al. (2016), who evaluated the Wald test across several popular additive models. It was shown that the Wald test can identify correct reduced models and improve attribute classifications, in particular, when the sample size is small and items are of low quality. Given the previous applications of the Wald test in the context of CDMs, this study proposes the Wald test for Q-matrix validation under the DINA and G-DINA models.

4.3.4 The Wald Test for Q-Matrix Validation

This paper introduces the Wald-Q, a new application of the Wald test, specifically, for Q-matrix validation purposes in conjunction with both reduced and general models. Among others, its primary appeal is that the Wald test only requires estimating the larger model (Buse, 1982). The Wald-based procedure for Q-matrix validation is a multivariate hypothesis test that performs all $2^K - p$ tests for each item at once. The test simultaneously compares all the possible q-vectors to the saturated specification at the item level. This involves testing $\mathbf{0} \prec \boldsymbol{q}^* \prec \mathbf{1}$ against $\boldsymbol{q} = \mathbf{1}$. In its current application, the Wald-Q can be considered as an all-subset search algorithm carried out iteratively. For this study, a q-vector with a saturated specification (i.e., $[1,1,\ldots,1]'$) is referred to as the full q-vector (i.e., $\boldsymbol{q} = \mathbf{1}$), and the others are called reduced q-vectors (i.e., $\mathbf{0} \prec \boldsymbol{q}^* \prec \mathbf{1}$).

The Wald-Q can be used when the underlying specific model is either known or unknown. The null hypothesis for Q-matrix validation can be written as $\boldsymbol{R}_{jp} \times \boldsymbol{P}_j = 0$, where $\boldsymbol{P}_j = P(\boldsymbol{\alpha}^*_{lj})$ is the vector of item parameters of the saturated model (i.e., G-DINA model with $\boldsymbol{q} = \mathbf{1}$), and $\boldsymbol{R}_{jp}$ is the $(2^K - p) \times 2^K$ matrix of restrictions, where

p is the number of parameters of the reduced model (i.e., DINA or G-DINA model with $\boldsymbol{q}^* \prec \mathbf{1}$). The rows and columns in the restriction matrix represent the contrasts and latent classes, respectively (see Tables 4.2 and 4.4). The number of rows of $\boldsymbol{R}_{jp}$, $(2^K - p)$, also indicates the degrees of freedom. The Wald statistic W_j to test the null hypothesis for Q-matrix validation is computed using Equation 4.4. Under the null hypothesis, the W_j is assumed to be asymptotically $\chi^2_{2^K-2}$ distributed when the smaller model is the DINA model, and asymptotically $\chi^2_{2^K-2^{K_j}}$ distributed when the reduced model is another G-DINA model, where the former assumes a particular underlying process, whereas the latter does not. Note that the larger model is always the G-DINA model that involves K attributes, whereas the smaller model can be the DINA or another G-DINA model that involves $K_j < K$ attributes. Additional constraints are necessary to specify a particular reduced CDM such that constraints are imposed to move from K to K_j.

For validation purposes, we need to estimate $\boldsymbol{P}_j$ and $Var(\boldsymbol{P}_j)$ for the full q-vector (e.g., $[1,1,1]'$ for $K = 3$). In the all-subset search method, the full q-vector is fixed; the remaining reduced q-vectors can be obtained by applying different constraints in the restriction matrix. For example, six constraints (i.e., restrictions) are needed for the DINA model, whereas, four constrains are needed for the G-DINA model when $K = 3$ and $K_j = 2$.

DINA model. Assume that $K = 3$. Each row of Table 4.1 represents a candidate q-vector. Entries in a row denote the latent groups to which attribute patterns belong. This grouping is essentially the same as η_{lj}, where $\eta_{lj} = 1$ if a latent group is expected to answer an item correctly, and 0 otherwise. For example, if α_1 and α_2 are required for an item, as in the q-vector $(1,1,0)'$, the eight latent classes are classified into two unique groups: $([0,0,0]', [1,0,0]', [0,1,0]', [0,0,1]', [1,0,1]', [0,1,1]')$ in group 0 and $([1,1,0]', [1,1,1]')$ in group 1.

Table 4.1: Number of Parameters (i.e., Latent Groups) for Different Q-Vectors when $K = 3$ (DINA Model Assumed)

	Attribute Patterns								
q-vector	(0,0,0)	(1,0,0)	(0,1,0)	(0,0,1)	(1,1,0)	(1,0,1)	(0,1,1)	(1,1,1)	#Par
(1,0,0)	0	1	0	0	1	1	0	1	2
(0,1,0)	0	0	1	0	1	0	1	1	2
(0,0,1)	0	0	0	1	0	1	1	1	2
(1,1,0)	0	0	0	0	1	0	0	1	2
(1,0,1)	0	0	0	0	0	1	0	1	2
(0,1,1)	0	0	0	0	0	0	1	1	2
(1,1,1)	0	0	0	0	0	0	0	1	2

Note. DINA: deterministic inputs, noisy "and" gate model. Entries in row denote the latent groups to which attribute vectors belong. #Par is the number of parameters associated with the corresponding q-vector.

According to the grouping formed by the q-vector $(1,1,0)'$, $\boldsymbol{R}_{jp}$ matrix for the DINA model is presented in Table 4.2. The first five restrictions in $\boldsymbol{R}_{jp}$, the success probabilities for the six attribute vectors in group 0, are constrained to be equal; the last restriction constrains the success probabilities of the two attribute vectors in group 1 to be equal. The restriction matrices for the remaining q-vectors can be derived in the same manner. Applying each restriction matrix to test each q-vector provides a hypothesis test. A reduced q-vector associated with retained null hypothesis can replace the full q-vector. In other words, if H_0 is retained, the q-vector with the fewest attribute specifications corresponding to a highest nonsignificant p-value is deemed correct for the item. If the null hypotheses for all the reduced q-vectors are rejected, we can conclude that the true q-vector is the full q-vector.

It should be noted that the restriction matrix in Table 4.2 simultaneously tests a particular reduced q-vector and a specific reduced CDM. However, in practice, we may only be interested in the former. This can be done by using restriction matrices corresponding to the G-DINA model. Conducting the restriction for the latter is the same as conducting a model comparison at the item level (de la Torre & Lee, 2013; Ma

Table 4.2: Restriction Matrix for the Q-Vector $(1,1,0)'$ ($K_j = 2$) (DINA Model Assumed)

(0,0,0)	(1,0,0)	(0,1,0)	(0,0,1)	(1,1,0)	(1,0,1)	(0,1,1)	(1,1,1)	$P(\alpha^*_{lj})$
1	-1	0	0	0	0	0	0	$P(0,0,0) = P(1,0,0)$
0	1	-1	0	0	0	0	0	$P(1,0,0) = P(0,1,0)$
0	0	1	-1	0	0	0	0	$P(0,1,0) = P(0,0,1)$
0	0	0	1	0	-1	0	0	$P(0,0,1) = P(1,0,1)$
0	0	0	0	0	1	-1	0	$P(1,0,1) = P(0,1,1)$
0	0	0	0	1	0	0	-1	$P(1,1,0) = P(1,1,1)$

et al., 2016).

G-DINA model. The restriction matrices for the G-DINA model can be obtained in a manner similar to the DINA model. However, in this case, the form of a reduced CDM is not specified. To test whether the same q-vector, $(1,1,0)'$, is the correct q-vector under the GINA model, four constraints are required because the item has four parameters (i.e., $2^K - p$), where $p = 2^{K_j}$, as shown in the last column of Table 4.3. Given the q-vector $(1,1,0)'$, four unique groups are defined: $([0,0,0]',[0,0,1]')$ is group 0, $([1,0,0]',[1,0,1]')$ is group 1, $([0,1,0]',[0,1,1]')$ is group 2, $([1,1,0]',[1,1,1]')$ is group 3.

Table 4.4 shows the four restrictions corresponding to the q-vector $(1,1,0)'$ for the G-DINA model. It also shows that the attribute patterns within the same group are assumed to have equal probability of correctly answering an item. Similar to the DINA model, identifying the suggested q-vector is based on the fewest attribute specifications among the retained reduced q-vectors. The Wald-Q for both the DINA and G-DINA models are implemented iteratively in that if any changes in attribute specifications are suggested, a new calibration is carried out with the suggested (i.e., updated) Q-matrix. In this iteration, the Wald-Q is again implemented. If further changes occur, the process is repeated, otherwise, the process is terminated.

Table 4.3: Number of Parameters (i.e., Latent Groups) for Different Q-Vectors when $K = 3$ (No Model Assumed)

	Attribute Patterns								
q-vector	(0,0,0)	(1,0,0)	(0,1,0)	(0,0,1)	(1,1,0)	(1,0,1)	(0,1,1)	(1,1,1)	#Par
(1,0,0)	0	1	0	0	1	1	0	1	2
(0,1,0)	0	0	1	0	1	0	1	1	2
(0,0,1)	0	0	0	1	0	1	1	1	2
(1,1,0)	0	1	2	0	3	1	2	3	4
(1,0,1)	0	1	0	2	1	3	2	3	4
(0,1,1)	0	0	1	2	1	2	3	3	4
(1,1,1)	0	1	2	3	4	5	6	7	8

Note. G-DINA: generalized deterministic inputs, noisy "and" gate model. Columns with the same number belong to the same group. #Par is the number of parameters associated with the corresponding q-vector.

Table 4.4: Restriction Matrix for the Q-Vector $(1,1,0)'$ ($K_j = 2$) (No Model Assumed)

(0,0,0)	(1,0,0)	(0,1,0)	(0,0,1)	(1,1,0)	(1,0,1)	(0,1,1)	(1,1,1)	$P(\boldsymbol{\alpha}^*_{lj})$
1	0	0	-1	0	0	0	0	$P(0,0,0) = P(0,0,1)$
0	1	0	0	0	-1	0	0	$P(1,0,0) = P(1,0,1)$
0	0	1	0	0	0	-1	0	$P(0,1,0) = P(0,1,1)$
0	0	0	0	1	0	0	-1	$P(1,1,0) = P(1,1,1)$

The implementation of the Wald-Q for each item can be summarized in the following five steps:

1. Calibrate the entire test using the provisional Q-matrix and the G-DINA model.

2. For item j, obtain the $\hat{\boldsymbol{P}}_j$ and $Var(\hat{\boldsymbol{P}}_j)$ based on $\boldsymbol{q}_j = \mathbf{1}$ and the G-DINA model using the $P(\alpha_l|\boldsymbol{X})$ and $\boldsymbol{X}_j$.

3. Based on $\hat{\boldsymbol{P}}_j$ and $Var(\hat{\boldsymbol{P}}_j)$ from step 2, carry out $(2^K - p)$ Wald tests to determine the correct q-vector for item j.

4. After completing the preceding steps for the first item, repeat the same steps for each of the J items.

5. If a Q-matrix entry is modified, repeat steps 1 through 3, and use the updated Q-matrix as the provisional Q-matrix.

The estimates of the item parameters and posterior distribution of the attribute patterns were obtained by an empirical Bayesian implementation of the expected-maximization algorithm (de la Torre, 2009) using the provisional Q-matrix. The code for the simulation study was written in Ox (Doornik, 2009).

4.4 Design and Analysis

This simulation study examined the performance of the Wald-Q by validating the correctness of attribute specifications in the Q-matrix, and the true-positive and false-positive rates were reported under various conditions. In the context of Q-matrix validation, the true-positive rate is the proportion of misspecified q-vectors that are correctly identified; the false-positive rate is the proportion of correctly specified q-vectors that are modified (de la Torre & Chiu, 2016). The true-positive and false-positive rates can be considered as analogous to power and Type-I error rates, respectively.

The viability of the Wald-Q was analyzed using two simulation studies to compare the effectiveness of: (1) the Wald-Q against the IMSSA when the underlying model can be assumed to be the DINA model; and (2) the Wald-Q to the iGDI under a saturated model when the underlying model cannot be assumed. In the simulation studies, 100 datasets were simulated using the DINA model with the following factors: sample sizes (N = 1,000 and 2,000), test lengths ($J = 15$ and 30), item qualities ($s_j = g_j = 0.1$, 0.2, and 0.3), attribute structures (uniform and higher-order distributions), and amount of misspecifications. In each condition, 100 misspecified Q-matrices, which contain 5% and 10% randomly misspecified q-entries, were generated from the true Q-matrix shown in Table 4.5. The true Q-matrix has only 1- and 2-attribute q-vectors included in $J = 30$, which was obtained by doubling the first $J = 15$ items. Therefore, results across the test lengths can be comparable. Using the higher-order (HO) structure for attribute generations, mastery or nonmastery of an attribute k is assumed to be related

to a unidimensional latent variable θ_i for examinee i. The probability of mastering α_k as a function of θ_i can be formulated as

$$P(\alpha_k|\theta_i) = \frac{exp(\lambda_{0k} + \lambda_{1k}\theta_i)}{1 + exp(\lambda_{0k} + \lambda_{1k}\theta_i)}, \tag{4.5}$$

where $\boldsymbol{\lambda}_0 = \{-1.0, -0.5, 0.0, 0.5, 1.0\}$ and $\lambda_1 = 1.0$ are the attribute difficulty and discrimination parameters, respectively; and θ_i, the ability of examinee i, was drawn from N(0,1).

Table 4.5: Q-matrix for the Simulated Data

Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0	11	0	1	0	1	0	21	1	1	0	0	0
2	0	1	0	0	0	12	0	1	0	0	1	22	1	0	1	0	0
3	0	0	1	0	0	13	0	0	1	1	0	23	1	0	0	1	0
4	0	0	0	1	0	14	0	0	1	0	1	24	1	0	0	0	1
5	0	0	0	0	1	15	0	0	0	1	1	25	0	1	1	0	0
6	1	1	0	0	0	16	1	0	0	0	0	26	0	1	0	1	0
7	1	0	1	0	0	17	0	1	0	0	0	27	0	1	0	0	1
8	1	0	0	1	0	18	0	0	1	0	0	28	0	0	1	1	0
9	1	0	0	0	1	19	0	0	0	1	0	29	0	0	1	0	1
10	0	1	1	0	0	20	0	0	0	0	1	30	0	0	0	1	1

Note. Each attribute is measured 10 times when J = 30 and five times when J = 15 with an equal number of 1- and 2-attribute q-vector.

4.4.1 Results

Results were compared in two respects. First, the Wald-Q was compared to the IMSSA when the DINA model is assumed, and second to the iGDI when a reduced model cannot be assumed. For the IMSSA, $\varepsilon^{(1)}$ values were set at 2.2, 1.9, and 1.7 for high, medium, and low quality items, respectively (Terzi & de la Torre, 2015). For the iGDI, in addition to setting the cut-off value for PVAF at $\varepsilon = 0.95$, as in de la Torre and Chiu (2016), this study investigated results for $\varepsilon = 0.90$ and 0.99 to further examine

which ε value can provide optimal results for the iGDI. Results were divided into two parts based on attribute patterns generated from the uniform and higher-order distributions. Note that since $(2^K - p)$ Wald tests were carried out to determine the correct q-vector for each item, the nominal α was modified using the Bonferroni correction (i.e., $\alpha/(2^K - p)$).

Table 4.6 shows the false-positive rates of the Wald-Q in comparison to the IMSSA. When attributes were uncorrelated (i.e., uniform), false-positive rates of both methods were around zero. The Wald-Q sometimes showed lower false-positive rates than the IMSSA, except when $N = 1,000$ and $J = 15$, and $N = 2,000$ and $J = 30$ with 10% misspecifications. Larger sample sizes and test items provided lower false-positive rates for both methods. When attributes were correlated (i.e., HO), the false-positive rates for both methods were higher than when attributes were uncorrelated. In particular, the IMSSA had higher Type-I error inflation than the Wald-Q. Again, increasing the sample size and test length lowered the false-positive rates for both methods, except when $J = 15$ and item quality was low for the IMSSA.

Table 4.7 presents the true-positive rates of the Wald-Q against the IMSSA. When attributes were generated from the uniform distribution, the Wald-Q provided equally well as or higher true-positive rates across all the conditions. In particular, the Wald-Q had perfect true-positive rates when item quality was high, with an exception when $N = 1,000$ and $J = 15$ with 10% misspecifications (0.99). The IMSSA also had perfect true-positive rates when item quality was high, with the exception of two conditions where $N = 1,000$, $N = 2,000$, and $J = 15$ with 10% misspecifications. Given the medium quality item, the Wald-Q had above 0.95 true-positive rates when $J = 15$; however, it was perfect when $J = 30$. The IMSSA showed above 0.84 true-positive rates, and were at least 0.99 when $J = 30$. As expected, both methods provided lower true-positive rates when the item quality was low. The true-positive rates of the Wald-Q ranged from 0.62 to 0.91 depending on the conditions. When $J = 15$ with 10% misspecifications, it was

Table 4.6: False-Positive Rates of the Wald-Q and IMSSA (DINA Model Assumed)

Quality	%	Wald-Q				IMSSA			
		N= 1,000		*N* = 2,000		*N*= 1,000		*N* = 2,000	
		J = 15	*J* = 30	*J* = 15	*J* = 30	*J* = 15	*J* = 30	*J* = 15	*J* = 30
		Uniform							
H	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	10	0.01	0.00	0.00	0.00	0.03	0.00	0.03	0.00
M	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L	5	0.02	0.00	0.00	0.00	0.02	0.01	0.00	0.00
	10	0.07	0.01	0.01	0.01	0.02	0.01	0.01	0.00
		Higher-Order							
H	5	0.04	0.01	0.02	0.00	0.26	0.23	0.20	0.18
	10	0.07	0.02	0.03	0.01	0.33	0.24	0.32	0.19
M	5	0.04	0.00	0.01	0.00	0.33	0.30	0.30	0.23
	10	0.04	0.01	0.04	0.01	0.45	0.30	0.43	0.25
L	5	0.03	0.01	0.00	0.00	0.44	0.36	0.46	0.31
	10	0.12	0.01	0.07	0.01	0.49	0.38	0.52	0.34

Note. IMSSA: iterative modified sequential search algorithm.

under the acceptable level (< 0.8). However, increasing the test length to 30 improved the rates to above the acceptable level (> 0.8), and ranged from 0.86 to 0.91. For the IMSSA, the true-positive rates were not acceptable (< 0.8). When attributes were generated from the HO distribution, the Wald-Q showed more stable results, which were above the acceptable level (> 0.8) throughout all the conditions. In particular, the true-positive rates were all perfect when $J = 30$. When $J = 15$, all the rates were above 0.93 except for $N = 1,000$ under the low quality item with 10% misspecifications. For the IMSSA, the true-positive rates were not acceptable (< 0.8) except for conditions where $N = 2,000$, $J = 15$ and 30 with 5% misspecifications, and $N = 2,000$ and $J = 30$ with 10% misspecifications.

Table 4.7: True-Positive Rate of the Wald-Q and IMSSA (DINA Model Assumed)

Quality	%	Wald-Q				IMSSA			
		N= 1,000		*N* = 2,000		*N*= 1,000		*N* = 2,000	
		J = 15	*J* = 30	*J* = 15	*J* = 30	*J* = 15	*J* = 30	*J* = 15	*J* = 30
		Uniform							
H	5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	10	0.99	1.00	1.00	1.00	0.95	1.00	0.97	1.00
M	5	0.97	1.00	0.98	1.00	0.91	1.00	0.93	1.00
	10	0.95	1.00	0.98	1.00	0.84	0.99	0.86	1.00
L	5	0.72	0.88	0.73	0.91	0.45	0.72	0.46	0.75
	10	0.62	0.86	0.68	0.91	0.36	0.72	0.36	0.75
		Higher-Order							
H	5	0.97	1.00	0.98	1.00	0.74	0.78	0.81	0.82
	10	0.96	1.00	0.97	1.00	0.60	0.76	0.64	0.81
M	5	0.96	1.00	0.98	1.00	0.64	0.70	0.67	0.79
	10	0.96	1.00	0.96	1.00	0.48	0.67	0.50	0.75
L	5	0.94	1.00	0.97	1.00	0.51	0.66	0.54	0.71
	10	0.87	1.00	0.93	1.00	0.36	0.62	0.40	0.67

Note. IMSSA: iterative modified sequential search algorithm.

Table 4.8 displays the false-positive rates of the Wald-Q and iGDI. When attributes were generated from the uniform distribution, the false-positive rates were very low (i.e., 0) for the Wald-Q and iGDI with at least medium quality items. When $N = 1,000$ and $J = 15$ with the low quality item, the false-positive rates were inflated for the Wald-Q, whereas, it was around the nominal level for the iGDI. When attributes were generated from the HO distribution, the false-positive rates were higher than when the attributes were generated from the uniform distribution; however, they were around the nominal level. In comparison to the Wald-Q, the iGDI showed similar or lower false-positive rates except when $N = 1,000$, $J = 30$, and the low quality item.

The true-positive rates of the Wald-Q and iGDI are reported in Table 4.9. Based on attributes generated from the uniform distribution, the iGDI performed equally well as

Table 4.8: False-Positive Rate of the Wald-Q and iGDI (No Model Assumed)

Quality	%	Wald-Q				iGDI			
		N= 1,000		*N* = 2,000		*N*= 1,000		*N* = 2,000	
		J = 15	*J* = 30	*J* = 15	*J* = 30	*J* = 15	*J* = 30	*J* = 15	*J* = 30
		Uniform							
H	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M*	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L*	5	0.26	0.01	0.01	0.00	0.01	0.03	0.00	0.00
	10	0.39	0.02	0.03	0.00	0.03	0.06	0.01	0.00
		Higher-Order							
H	5	0.04	0.01	0.01	0.00	0.01	0.00	0.01	0.00
	10	0.08	0.01	0.05	0.01	0.01	0.00	0.01	0.00
M*	5	0.02	0.02	0.01	0.00	0.01	0.01	0.02	0.00
	10	0.10	0.03	0.09	0.01	0.05	0.02	0.06	0.00
L*	5	0.03	0.01	0.00	0.00	0.04	0.14	0.02	0.00
	10	0.12	0.01	0.07	0.01	0.10	0.15	0.07	0.01

Note. * indicates that the iGDI provided higher recovery when $\varepsilon = 0.90$ for the medium and low quality items.

or better than the Wald-Q when the item quality was medium and high, where the true-positive rates were above 0.95 for both methods. When the item quality was low, the Wald-Q outperformed the iGDI. In particular, it was above 0.93 when $N = 2,000$ and $J = 30$. For the other conditions, the true-positive rates were not acceptable (< 0.8) for both methods. Results were more stable when attributes were generated from the HO distribution. Overall, the Wald-Q performed equally well as or better than the iGDI. The true-positive rates for the Wald-Q were at least close to perfect (above 0.99) when $J = 30$ across all the conditions, and it was above 0.90 when $J = 15$. The true-positive rates for the iGDI were also close to perfect (above 0.99) when $N = 2,000$ and $J = 30$. The rates were also acceptable (> 0.8) for the iGDI across the rest of the conditions, except when $N = 1,000$ and $J = 15$ with 10% misspecifications (0.70).

Table 4.9: True-Positive Rate of the Wald-Q and iGDI (No Model Assumed)

Quality	%	Wald-Q				iGDI			
		N= 1,000		*N* = 2,000		*N*= 1,000		*N* = 2,000	
		J = 15	*J* = 30	*J* = 15	*J* = 30	*J* = 15	*J* = 30	*J* = 15	*J* = 30
		Uniform							
H	5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
M*	5	0.97	1.00	0.95	1.00	0.97	1.00	1.00	1.00
	10	0.96	1.00	0.96	1.00	0.95	1.00	1.00	1.00
L*	5	0.53	0.72	0.68	0.96	0.30	0.57	0.44	0.92
	10	0.43	0.70	0.61	0.93	0.23	0.48	0.38	0.87
		Higher-Order							
H	5	0.97	1.00	0.98	1.00	0.97	1.00	0.97	1.00
	10	0.96	1.00	0.97	1.00	0.97	1.00	0.97	1.00
M*	5	0.97	0.99	0.96	1.00	0.95	0.99	0.95	1.00
	10	0.90	0.99	0.92	1.00	0.89	0.98	0.90	1.00
L*	5	0.93	0.99	0.95	1.00	0.85	0.82	0.92	1.00
	10	0.90	0.99	0.92	0.99	0.70	0.82	0.82	0.99

Note. * indicates that the iGDI provided higher recovery when $\varepsilon = 0.90$ for the medium and low quality items.

4.5 Implementation with Real Data

The viability of the Wald-Q against the IMSSA and iGDI was further investigated using real data. The fraction-subtraction test (Tatsuoka, 1984) with 12 items taken by 536 middle school students were examined. The following four attributes were used in the Q-matrix: α_1 – performing a basic fraction subtraction operation; α_2 – simplifying/reducing; α_3 – separating a whole number from fraction; and α_4 – borrowing one from a whole number to fraction.

Table 4.10 shows the 12 items with the corresponding attribute specifications. Results from the Wald-Q, IMSSA, and iGDI were compared based on suggested q-vectors. First comparison of the Wald-Q to the IMSSA assumed that the underlying reduced model is DINA, whereas, the next comparison of the Wald-Q to the iGDI assumed that

Table 4.10: Q-Matrix for Fraction-Subtraction Items

Item		Attributes			
		α_1	α_2	α_3	α_4
1.	$\frac{3}{4}-\frac{3}{8}$	1	0	0	0
2.	$3\frac{1}{2}-2\frac{3}{2}$	1	1	1	1
3.	$\frac{6}{7}-\frac{4}{7}$	1	0	0	0
4.	$3\frac{7}{8}-2$	1	0	1	0
5.	$4\frac{4}{12}-2\frac{7}{12}$	1	1	1	1
6.	$4\frac{1}{3}-2\frac{4}{3}$	1	1	1	1
7.	$\frac{11}{8}-\frac{1}{8}$	1	1	0	0
8.	$3\frac{4}{5}-3\frac{2}{5}$	1	0	1	0
9.	$4\frac{5}{7}-1\frac{4}{7}$	1	0	1	0
10.	$7\frac{3}{5}-\frac{4}{5}$	1	0	1	1
11.	$4\frac{1}{10}-2\frac{8}{10}$	1	1	1	1
12.	$4\frac{1}{3}-1\frac{5}{3}$	1	1	1	1

Note. (α_1) performing a basic fraction subtraction operation, (α_2) simplifying/reducing, (α_3) separating a whole number from fraction, and (α_4) borrowing one from a whole number to fraction.

the underlying reduced model is unknown.

4.5.1 Results

The comparison of the Wald-Q to the IMSSA is shown in Table 4.11. The Wald-Q suggested changes in nine attribute specifications (i.e., items 1, 2, 4, 5, 6, 7, 8, 9, and 12). In contrast, the IMSSA suggested changes in 10 attribute specifications (items 1, 3, 4, 7, 8, 9, and 10). These changes could be due to issues beyond the scope of this paper, such as the incompleteness of the Q-matrix, different strategies of solving questions, the small sample sizes, and short tests among others.

For example, the incompleteness of the Q-matrix in the data is due to the fact that only 58 of 256 ($K = 8$; Chiu, 2013) and 10 of 32 ($K = 5$; Chiu & Köhn, 2015) possible attribute patterns can be identified by the items. Therefore, multiple classes

Table 4.11: Suggested Q-Matrix by the Wald-Q and IMSSA (DINA Model Assumed)

	Wald-Q				IMSSA			
Item	α_1	α_2	α_3	α_4	α_1	α_2	α_3	α_4
1.	1	0	1	0	1	1	1	1
2.	1	0	1	1	1	1	1	1
3.	1	0	0	0	1	1	1	0
4.	1	0	0	0	1	1	1	0
5.	1	0	1	1	1	1	1	1
6.	1	0	1	1	1	1	1	1
7.	1	0	0	0	1	1	1	0
8.	1	0	0	0	1	1	1	0
9.	1	0	0	0	1	1	1	0
10.	1	0	1	1	1	1	1	1
11.	1	1	1	1	1	1	1	1
12.	1	0	1	1	1	1	1	1

Note. (α_1) performing a basic fraction subtraction operation, (α_2) simplifying/reducing, (α_3) separating a whole number from fraction, and (α_4) borrowing one from a whole number to fraction.

can be merged (Chiu, 2013). Another possibility is that examinees can apply different strategies to answer items correctly. To give an example for item 2, where attribute suggestions by the IMSSA concurred with the provisional Q-matrix that require all the four attributes. However, the Wald-Q suggested excluding α_2. It is interesting that the other three attributes – borrowing one from a whole number to fraction, separating a whole number from fraction, and performing a basic fraction – happen to give the correct answer. The following example shows strategies step by step how to solve item 2:

$$3\frac{1}{2} - 2\frac{3}{2} = 2\frac{(1\times 2)+1}{2} - 2\frac{3}{2}$$
$$= (2-2) + (\frac{3}{2} - \frac{3}{2}) = 0 + \frac{3-3}{2} = 0,$$

meaning that mastering these three attributes would be enough to answer item 2 correctly rather than mastering all the four attributes specified in the Q-matrix.

Table 4.12: Suggested Q-Matrix by the Wald-Q and iGDI (No Model Assumed)

	Wald-Q				iGDI			
Item	α_1	α_2	α_3	α_4	α_1	α_2	α_3	α_4
1.	1	1	0	0	1	0	1	1
2.	1	0	1	1	1	1	1	1
3.	1	0	0	0	1	0	0	0
4.	1	0	0	0	1	0	1	0
5.	1	0	1	1	1	0	1	1
6.	1	0	1	1	1	0	1	1
7.	1	0	0	0	1	1	0	0
8.	1	0	0	0	1	0	1	0
9.	1	0	0	0	1	0	1	0
10.	1	0	1	1	1	0	1	1
11.	1	0	1	1	1	0	1	1
12.	1	0	1	1	1	0	1	1

Note. (α_1) performing a basic fraction subtraction operation, (α_2) simplifying/reducing, (α_3) separating a whole number from fraction, and (α_4) borrowing one from a whole number to fraction. iGDI results were obtained based on $\varepsilon = 0.95$.

The Wald-Q was further compared to the iGDI for the G-DINA model shown in Table 4.12. Results indicated that the Wald-Q was more liberal than the iGDI. The iGDI suggested changing only six attribute specifications in items 1, 5, 6, 11, and 12; however, the Wald-Q suggested changing 10 attribute specifications in items 1, 2, 4, 5,

6, 7, 8, 9, 11, and 12. The interpretations of the previous results can be applied to these results obtained for the iGDI and Wald-Q under the G-DINA model.

It is important to state that the Wald-Q provided different attribute changes in two items depending on whether or not the underlying DINA model can be assumed. For item 1, the Wald-Q suggested α_3 and α_2 under the DINA and G-DINA models, respectively. For item 11, the Wald-Q under the DINA model agreed with the original Q-matrix, however, the Wald-Q under the G-DINA model excluded α_2.

4.6 Summary and Discussion

This study adapted the Wald test as a method of empirically validating the Q-matrix that can be used with reduced and general CDMs. Results showed that the Wald-Q can identify misspecified q-entries at a high rate, especially when the test is long.

Under favorable conditions, the false-positive rates were around the nominal level. The Wald-Q also behaved similarly for the reduced and general models. Longer tests had more positive impact than larger sample sizes on the Wald-Q under both models, which is similar to findings in Sorrel et al. (2017)'s study. Results were stable under both models, especially when the item quality was high and medium.

Attribute structures affected the use of the Wald-Q with reduced and general models differently. For instance, when attributes were correlated (i.e., HO), the reduced and general models had similar true-positive and false-positive rates. When attributes were uncorrelated (i.e., uniform), the reduced model showed higher true-positive rates, and similar or lower false-positive rates than when attributes were correlated.

The performance of the Wald-Q was better than that of the IMSSA across the board, particularly when the attributes were correlated. The Wald-Q and iGDI were not too different when item quality was medium or high, but the former was generally better than the latter with low quality items.

The time to implement the Q-matrix validation procedures using a 3.50-GHz I7 computer was the shortest for the IMSSA, followed by the iGDI and Wald-Q. Specifically, it took the code 2.68, 14.02, and 21.10 minutes to run the IMSSA, iGDI, and Wald-Q, respectively, for 100 iterations under the condition where $N = 2,000$, $J = 30$, and medium quality items with 10% misspecifications under a uniform distribution. The average number of iterations was two, and did not go beyond four.

In this study, assuming a reduced or saturated model differed based on the restriction matrix after calibrating the data using the G-DINA model. However, it is possible that better results could be obtained if a reduced model can be assumed in advance in calibrating the data. More specifically, in this study the data were calibrated using the G-DINA model for the Wald-Q and iGDI. However, the Wald-Q for the Q-matrix validation was implemented using different restriction matrices based on the assumption that if a reduced model can be assumed or not. For the IMSSA, the data calibration and Q-matrix validation were carried out using the DINA model.

A general idea of this study is not to replace any existing validation methods or domain experts but rather to provide a statistical perspective as an additional supplementary tool. It should be emphasized that including domain experts in the process of Q-matrix validation is as important as including available statistical methods for validating the correctness of attribute specifications.

Future research is necessary to better understand the Wald-Q. Additional work should be done to determine how the Wald-Q behaves when other reduced CDMs are involved. The Wald-Q should also be examined including a different number of attributes. Furthermore, different methods of computing $Var(\boldsymbol{P}_j)$ should be explored to see if the performance of the Wald-Q can be improved, particularly with the low quality item.

4.7 References

Akbay, L., Terzi, R., Kaplan, M., & Karaaslan, K. G. (2017). Expert-based attribute identification and validation: An application of cognitively diagnostic assessment. *Journal on Mathematics Education*, *9*(1). Retrieved from `http://ejournal.unsri.ac.id/index.php/jme/article/view/4341`

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician*, *36*, 153–157.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*, 598–618.

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

Chiu, C.-Y., & Köhn, H.-F. (2015). Consistency of cluster analysis for cognitive diagnosis the DINO model and the DINA model revisited. *Applied Psychological Measurement*, *39*, 465–479.

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

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

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253–273.

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

de la Torre, J., & Lee, Y. S. (2013). Evaluating the wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*, 355–373.

Doornik, J. A. (2009). An object-oriented matrix programming language Ox 6. [Computer software]. London, UK: Timberlake Consultants Ltd.

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

Hartz, S., Roussos, L., Henson, R., & Templin, J. (2005). *The fusion model for skill diagnosis: Blending theory with practicality. Unpublished manuscript.*

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*, 191–210.

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the wald test to

investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*, 98–125.

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.

Leighton, J., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, UK: Cambridge University Press.

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

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*, 200–217.

Morrison, D. F. (1967). *Multivariate statistical methods*. New York, NY: McGraw-Hill.

Rojas, G., de la Torre, J., & Olea, J. (2012). Choosing between general and specific cognitive diagnosis models when the sample size is small. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

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

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement*, 1–18.

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

Tatsuoka, K. K. (1984). *Analysis of errors in fraction addition and subtraction problems*. Computer-based Education Research Laboratory, University of Illinois.

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

Terzi, R., & de la Torre, J. (2015). *An iterative method of empirically-based Q-matrix validation*. Paper presented at the annual aeeting of the National Council on Measurement in Education, Chicago, IL.

Terzi, R., & de la Torre, J. (in preparation). *The Jensen-Shannon Divergence index and iterative GDI for Q-matrix validation under a general CDM*. Manuscript in preparation.

Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal*, *26*, 237–255.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.

Chapter 5
Summary

Traditional unidimensional item response theory (IRT) models describe examinees' proficiency by providing single overall ability scores. A main limitation of a single score is its inability to provide detailed diagnostic information about specific skills students should master to become proficient in a domain of interest (e.g., Leighton & Gierl, 2007). In contrast, cognitively diagnostic assessments (CDAs) have been designed to provide more detailed information by uncovering examinees' current knowledge, skill sets, and capabilities within a particular content area so that specific attributes that may need academic support while learning is occurring can be identified (de la Torre, 2009). The latent skills possessed by examinees can be discovered based on examinee responses to test items primarily in conjunction with cognitive diagnosis models (CDMs).

Most, if not all, CDMs require a Q-matrix to specify attributes measured by each item. If attributes have been correctly specified, CDMs can accurately identify examinees' mastery or nonmastery of attributes. However, conventional Q-matrix development process has some degree of subjectivity due to the involvement of human judgments, and has raised validity concerns due to the possibility of inaccurate attribute classifications. Although some statistical procedures exist in the literature, additional work is still needed to address remaining concerns in Q-matrix validation. With a general aim to validate the accuracy of attribute specifications in cognitive diagnosis modeling framework, this dissertation proposed new Q-matrix validation procedures, and addressed concerns in some of the current validation methods.

In the first study, a new search algorithm, *iterative modified sequential search algorithm* (IMSSA), based on the sequential EM-based δ-method (de la Torre, 2008) was proposed to empirically validate the correctness of attribute specifications. The IMSSA is an extension of the *sequential search algorithm* (SSA; de la Torre, 2008) in that the former addressed some limitations of the latter in various ways. Using two simulation studies, the IMSSA was compared to three methods without an iterative algorithm and to a method with an iterative algorithm. Among the noniterative algorithms, the *modified sequential search algorithm* (MSSA) showed better results, which also provided higher recovery of attribute specifications than the *Q-matrix refinement method* (QRM) on average across the conditions. Moreover, the IMSSA had much better results than the noniterative algorithms. On average, the IMSSA and QRM had perfect recovery with large sample sizes and long tests, and very high recovery rates with short tests based on data generated from the high quality items; and the IMSSA outperformed the QRM when data were generated from the medium and low quality items.

In the second study, the Jensen-Shannon divergence (iJSD) index and the *iterative* generalized deterministic inputs, noisy "and" gate (G-DINA) model discrimination index (iGDI) were proposed. Both indices are used as empirically-based Q-matrix validation methods to verify the correctness of attribute specifications in the Q-matrix. As with the iGDI, the iJSD was also implemented iteratively. Results showed that when the item quality was at least medium, the iJSD and iGDI can identify misspecified q-entries at a high rate. Attribute structures had a different impact on the results of the iJSD and iGDI. When attributes were generated from a higher-order (HO) distribution, the true-positive and false-positive rates were more stable than those generated from a uniform distribution. Finally, given the complexity of q-vectors, a higher recovery of attribute specifications was obtained based on the Q-matrix with more single-attribute q-vectors than fewer single-attribute q-vectors. In other words, more complexity of q-vectors considered in the Q-matrix did not provide better results.

In the third study, the Wald test (Morrison, 1967) was adapted to carry out multivariate hypothesis testing to validate Q-matrix entries, which is called Wald-Q. The Wald-Q can be applied to reduced and general CDMs based on the restriction matrix. The effectiveness of the Wald-Q was compared to the IMSSA proposed in the first study and to iGDI proposed in the second study in conjunction with the DINA and G-DINA models, respectively. The Wald-Q outperformed the IMSSA across the conditions, particularly when the attributes were correlated. The Wald-Q and iGDI were not too different when item quality was at least medium, but the former was generally better than the latter with low quality items. Results displayed that the Wald-Q can identify misspecified q-entries at a high rate, especially when the test was long. Moreover, longer tests had a more positive impact than larger sample sizes on the Wald-Q under both models. This interpretation was supported by Sorrel, Abad, Olea, de la Torre, and Barrada (2017)'s findings that the impact of increasing the test length on the Wald test was larger than increasing the sample size.

The Wald-Q under reduced and general models performed differently based on the attribute structures. In particular, when attributes were generated from a HO distribution, the reduced and general models showed similar true-positive and false-positive rates. When attributes were generated from a uniform distribution, the reduced model had higher true-positive rates, and similar or lower false-positive rates than the general model.

Attribute specifications in the Q-matrix should be correctly identified to obtain maximum information from a CDM estimation (de la Torre, 2008). Hence, given the results of the three studies as a whole, this dissertation showed considerable improvement in verifying the correctness of attribute specifications. The dissertation's first study was useful in addressing some of the limitations with a current method on which the proposed new method was based. The second study was important in proposing more general indices with an iterative algorithm that can be extended to a wider class

of CDMs so that assuming an underlying process would not be required. The third study was essential in obtaining more accurate validity of attribute specifications in the Q-matrix by providing comparisons of the Wald-Q to the other two methods, the IMSSA and iGDI, proposed in the previous two studies.

A successful implementation of the proposed methods can lead to the advancement of the use of CDAs in educational settings by accurately estimating attribute classifications. Results leading to improvements in Q-matrix validation can also help other components of cognitive diagnosis modeling, such as the estimation of model parameters, model-data fit analyses, the accuracy of attribute classifications, and ultimately, validity of CDA inferences. Nonetheless, there are still questions that need further investigation. For example, the simulated response data were generated based on the DINA model in the three studies. However, it would be interesting to analyze the efficiency of the new procedures using response data that would be generated based on other reduced models. Moreover, the new procedures should be analyzed using real data with a complete Q-matrix. Unfortunately, it was stated by Chiu, Douglas, and Li (2009) and Chiu (2013) that the Q-matrix for the fraction-subtraction data is not complete. So, all possible attribute patterns cannot be specified, which is actually a similar problem of model misfit caused by an incomplete set of the skills in the Q-matrix (de la Torre & Chiu, 2016). Finally, the proposed procedures need to be improved in cases with short tests, small sample sizes, and low quality items.

5.1 References

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*, 598–618.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*, 633–665.

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.

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*, 163–183.

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253–273.

Leighton, J., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, UK: Cambridge University Press.

Morrison, D. F. (1967). *Multivariate statistical methods*. New York, NY: McGraw-Hill.

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement*, 1–18.