A HYPOTHESIS TESTING PROCEDURE DESIGNED FOR Q-MATRIX VALIDATION OF DIAGNOSTIC CLASSIFICATION MODELS

by

Ruchi Jain Sachdeva

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

A Hypothesis Testing Procedure Designed for Q-Matrix Validation of Diagnostic Classification Models

Ruchi Jain Sachdeva

Cognitive diagnosis models have become very popular largely because these models provide educators with an explanation for a student not performing well based on skills that have not yet been mastered, making it possible for educators to provide targeted remediation and tailor instruction to address individual strengths and weaknesses. However, in order for these procedures to be effective, the Q-matrix which establishes the relationships between latent variables representing knowledge structures (columns) and individual items on an assessment (rows) must be carefully considered. The goal of this work is to develop a new test statistic for the detection of model misspecifications of the Q-matrix, which include both underfitting the Q-matrix and overfitting the Q-matrix. In addition to the development of this new test statistic, this dissertation evaluated the performance of this new test statistic and developed an estimator of the asymptotic variance based on the Fisher Information Matrix of the slip and guess parameters.

The test statistic was evaluated by two simulation studies and also applied to the fraction subtraction dataset. The first simulation study investigated the true Type-I error rates for the test under four levels of sample size, three levels of correlation among attributes and three levels of item discrimination. Results showed that as the sample size increases the Type I error reduces to 5%. Surprisingly, the results for the relationship between Type I error and Item discrimination show that the most discriminating items (Item Discrimination of 4) have the largest Type I error rates. The power study showed

that the statistic is very powerful in the detection of under-specification or over-specification of the Q-matrix with large sample sizes and/or when items are highly discriminating between students that have mastered or have not mastered a skill. Interestingly, the results when the Q matrix has multiple misspecifications the detection of under-specification is better than for over-specification when two misclassifications are being tested simultaneously. The analysis of the fraction subtraction dataset found 15% of the q-entries had enough evidence to reject the Null hypothesis. This clearly indicates that the test finds misfit in the original expert designed Q-matrix.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGEMENTS

This work could not have been completed without the extensive support and guidance of my advisor, Professor Matthew Johnson. From the start of this project to its completion, Professor Johnson provided crucial insights and suggestions that kept the project on-track, and I am indebted to him for his patient mentoring. I'm truly grateful to the members of my committee—Professors Young-Sun Lee, James Corter, Charles Lang, and Ronald Neath—who contributed greatly to this work through their helpful questions, feedback, and suggestions. Collectively they are all tremendous researchers and exceptional teachers and I'm grateful to have had their input at crucial steps along the way.

Nobody has been more important to me in the pursuit of this endeavor than the members of my family. I would like to thank my mother and father, whose unconditional love, guidance and unwavering support are with me in whatever I pursue and shaped who I am today. Mom and Dad, I'm forever grateful for your selfless love and sacrifice. Both of you have always cherished with me every great moment and supported me whenever I've needed it. My little brother, Neerav I am eternally grateful for your steadfast love, support and encouragement. On the days that I almost didn't believe I could get to the finish line your encouragement and belief in me made all the difference.

Most importantly, I wish to give my heartfelt thanks to my husband, Anil, whose unconditional love, patience, and continual support of my academic endeavors over the past several years enabled me to complete this dissertation. This journey has been long and difficult but I am thankful that I had you by my side. I may have had the all-nighters but you stayed up late just to give me company and supported me in a way only a best friend and companion can do. I love you and I could not have completed this journey without you.

Finally, I dedicate this work to my two wonderful children, Divya and Arjan, who have made me stronger, more fulfilled and provide me with unending inspiration. Divya, you have grown up watching mommy study and in your special way you have always helped make the tough days easier and the finish-line sweeter. Arjan, my little one, your immense capacity to show affection has been my strength and joy. Divya and Arjan, if you read this when you are older, we love you and we will make sure you get to follow your dreams just as we could.

DEDICATION

To my husband, Anil.

To my children, Divya and Arjan.

And to my parents and brother, Virender, Usha and Neerav.

CHAPTER 1: INTRODUCTION

Cognitive diagnosis models (Leighton & Gierl, 2007) also known as diagnostic classification models (Rupp, Templin, & Henson, 2010) have received increasing attention in the recent years within educational and psychological measurement because they prove useful when diagnostic decisions about proficiency are required (Rupp & Templin, 2008a) The popularity of these models is largely due to their focus on obtaining information about student mastery or possession of a set of predefined skills or attributes based upon their responses to a set of test items not readily gained from alternative latent variable models, including those based on Item Response Theory (IRT). The terms skill and attribute are used interchangeably to describe the knowledge procedural or declarative that an examinee needs to successfully complete a task in a specific domain (Leighton & Gierl, 2007). For assessments to be used as a formative tool, they need to be designed to classify students' performance at a finer-grained level. This characteristic of CDMs provides educators with an explanation for a student not performing well based on skills that have not yet been mastered, making it possible for educators to provide targeted remediation and tailor instruction to address individual strengths and weaknesses. However, in order for these procedures to be effective, the input for the models must be carefully considered. The set of attributes or skills an instrument measures as well as the patterns of these skills which are reflected in the items on a test determine the quality of the diagnostic information that can be obtained from these models. One critical step when implementing CDMs is the specification of which skills (latent) are required to successfully answer each item (observed behavior) on the diagnostic assessment. This matrix of specification is called a Q-matrix (Tatsuoka, 1983) and relates latent skills to observed behavior. Through a pattern of 0's and 1's, the Q-matrix establishes the relationships between latent variables representing knowledge structures (columns) and individual items on an assessment (rows) (Rupp & Templin, 2008b). Henson & Douglas (2005) have defined this Q-matrix to have elements q_{jk} for J items and K attributes indicating whether mastery of attribute k is required by item j such that $q_{jk} = 1$ if item j requires attribute k, and 0 otherwise. However, recent studies (Rupp & Templin, 2008a; DeCarlo, 2011; Kunina-Habenicht, Rupp, & Wilhelm, 2012) have shown that considering the Q-matrix as known or fixed and solely based on theory determined by domain experts, item developers, teachers and psychometricians may be misleading. A proposed Q-matrix by content experts may not be identical to the 'true' Q-matrix, even if the experts carefully take into account instructional purposes and students' protocols. The correct specification of a Q-matrix requires the correct identification of the cognitive processes underlying the assessment item responses. However, multiple strategies different from what are specified in the Q-matrix may be used by test takers, making the derivation of the Q-matrix very subjective in nature. In some cases, determining the necessity of a given skill for each item can be relatively straightforward. For instance, the skills required to answer an item on a math test can be easily established by visual examination of the symbols for addition, subtraction, etc. However, on more subjective tests such as reading comprehension tests, the underlying skills required by each item may be more difficult to determine.

Since the information contained within a Q-matrix is the primary driver of the usefulness of a CDM, correct specification of the Q-matrix is essential in providing accurate diagnostic results for each examinee. When a Q-matrix has been incorrectly specified, the misspecification can be classified as either *under-specification* or *over-specification*. If a Q-matrix is under-specified, some attributes that are in fact measured by an item are not recorded as such, i.e. an under-specified Q-matrix will have at least one entry that should be a '1' but is mistakenly recorded as a '0.' In previous research, this type of misspecification has been referred to as an incomplete Q-matrix. Similarly, in an over-specified Q-matrix attributes identified as measured by an item are not in fact related to the item. The misspecification of a Q-matrix would lead to undesirable consequences for example poor model fit, inaccurate model parameter estimation (Henson, Templin, & Willse, 2009) (Rupp & Templin, 2008a), and incorrect interpretations of the set of user-specified attributes. Therefore, it becomes essential to develop statistical techniques to both validate the Q-matrix specifications as developed by the content expert and possibly improve upon the accuracy of that Q-matrix using empirical data.

The necessity of the present research arises because few studies have attempted to develop methods of validating the Q-matrix for CDMs. Most previous research has focused on examining the effect of Q-matrix misspecification, for example, (Rupp & Templin, 2008a) investigated the effect of Q-matrix misspecification on item parameter estimation for the DINA model. DeCarlo (2011) examined the impact of Q-matrix misspecifications on latent class sizes under the DINA model. Kunina-Habenicht, Rupp, & Wilhelm (2012) examined the effects of model misspecification due to Q-matrix misspecifications on item parameter estimation and respondent classification within a broader CDM framework.

Few studies have proposed methods to validate a Q-matrix for CDMs. One such study by de la Torre (2008) has proposed an empirically based sequential search algorithm to correct an incorrectly specified Q-matrix under two conditions: 1) the response data is modeled by a DINA model; 2) the number of incorrectly specified q-vectors is small compared to the number of items in the assessment. In addition, Liu, Xu, & Ying (2012) stated that under the DINA or DINO model, if a Q-matrix is correctly specified, the Euclidean distance between expected proportions of positive responses to all items and a model-based combination of items and the corresponding observed proportions converges to zero in probability. They suggest that a procedure can be formed to validate an existing Q-matrix by inspecting the closeness of the Euclidean distance between the above two vectors to zero.

DeCarlo (2012) demonstrated that suspected misspecified Q-matrix entries can be validated for the DINA and reparameterized DINA if we assume that a level of uncertainty exists regarding whether correctly completing a task depends on possessing a particular skill. This uncertainty was modeled using the Bayesian framework and more specifically it involved treating those questionable Q-matrix entries as Bernoulli random variables and estimating them simultaneously with the other parameters in the model. Chung (2014) extends his work by using the Bayesian framework to estimate the whole Q-matrix rather than just some questionable entries. Xiang (2013) uses a mathematical framework to estimate the true Q-matrix based on item response data.

Notably, the key elements of CDMs are not unique to these models but have their roots in other major psychometric and statistical frameworks such as structural equation modeling (e.g., Kline, 2002). Similar to CDMs, Structural Equation Modeling (SEM) is a statistical framework for modelling the relationship between observed and latent variables as well as among latent variables themselves. Like CDMs, SEM is comprised of both a measurement (Q-matrix in CDMs) and a structural component. The measurement component relates observed responses or indicators to latent variables and is evaluated using Confirmatory Factor Analysis in SEM. The structural model specifies the relations among latent variables and is estimated using Path analysis in SEM. Similar to CFA, CDMs contain multiple latent predictor variables, each indexing one of the postulated skills for the diagnostic assessment. The number of latent variables depends on the number of skills that researchers hope to numerically separate in a reliable manner with the assessment. The main difference between the two model families is that in SEM the observed variables can be categorical or continuous but the latent variable is assumed to be continuous while in the CDM family both the observed and the latent variables are categorical. The SEM framework predates CDMs and has faced similar concerns around the validation of the measurement portion before undertaking the structural. As such, in SEM the measurement model is often tested and validated using modification indices. These modification indices quantify the expected drop in X^2 if a previously fixed parameter is set free in order to be estimated from the data and improve the fit of the model.

The goal of this work is to create a method of Q-matrix validation in a manner similar to the way modification indices work to improve the fit of the model. It involves inferential procedures to test statistical hypotheses on individual q matrix entries. These procedures would be evaluated based on the improvement in accuracy of the Q-matrix originally developed using the domain expert's knowledge. Next, we will test whether these indices will aid in retrieving the "true" Q-matrix under various Q-matrix misspecifications. The primary hypothesis of this investigation is that using these inferential procedures will yield a CDM with a Q-matrix with fewer misspecifications and in turn lead to better fit to the data, more stable parameter estimates and more accurate classification rates.

CHAPTER 2: LITERATURE REVIEW

Chapter 2 consists of three major sections. The first section introduces general information about the Cognitive Diagnostic Models, followed by a more specific description of the DINA model. The second section discusses the existing Q-matrix misspecification and validation literature, and the final section details the SEM modification indices.

Overview of Cognitive Diagnosis Models

In contrast to IRT models which model ability a continuous measure, CDMs are discrete latent variable models, with a simple between-item or a complex within-item multidimensional loading structure between the items and the latent skills. These models target to evaluate an examinee’s competencies in the domain of interest represented by a set of micro-level skills. In contrast to a handful of well-known statistical techniques for criterion-referenced assessment that were based on Classical Test Theory, Item Response Theory, and Factor Analysis, CDMs eliminate multiple sources of error arising from the latent-variable scores, assumption of the latent variable scores themselves, and the determination of cut-scores (Rupp, Templin, & Henson, 2010). Thus, CDMs serve as better models to classify examinees based on a set of multiple micro-level skills. While CDMs are specifically designed for cognitively diagnostic assessments, they can also be used in traditional tests, provided a small set of finely-grained skills can be shown to underlie test performance.

Cognitive Diagnostic Models can be classified based on three characteristics: (1) the scale type of the observed (response) variables, (2) the scale type of the latent (attribute) variables, and (3) whether a compensatory or non-compensatory combination of the latent attributes variables is needed (Rupp & Templin, 2008b). Although there are many different models, all serve the same purpose: to classify respondents into a predetermined number of latent classes which are attribute patterns. More specifically, the classification of CDMs splits product Bernoulli models with binary latent student parameters into two groups: non-compensatory and compensatory models. Non-compensatory models, also called conjunctive models, are those that require a student to possess all of the skills that an item requires in order to answer it correctly. Alternatively, compensatory models allow for a lacking skill to be compensated by the presence of other skills, i.e. a student who is missing a skill deemed necessary to answer a particular item still has a non-zero probability of correctly answering the item. An extreme case of compensatory models are disjunctive models which only require one or more traits to be present for a student to correctly answer an item (Rupp & Templin, 2008b).

Prior to discussing different types of CDMs, several key concepts and terminology need to be clarified. As mentioned earlier, an important characteristic of CDMs is their capability for classifying test takers by their ability of interest represented by a group of latent classes, often called attribute mastery profiles, which are defined by the mastery or non-mastery of a set of interrelated but separable latent attributes measured by the test. The attribute mastery profile of each examinee i (i =1,…, I) is characterized by an attribute vector of length K denoted by $\alpha_i = (\alpha_{i1},...,\alpha_{ik},...,\alpha_{iK})'$ for an assessment involving K attributes (Samejima, 1995). Specifically, α_{ik} =1 if the k*th* attribute has been mastered by the *i*th examinee and α_{ik} = 0 if the k*th* attribute has not been mastered. Each attribute vector defines a unique latent class. Thus, K attributes define $L = 2^K$ latent classes for the analysis. The primary objective of a CDM is to classify examinees into these latent classes.

The degree of the definitional specificity of attributes is associated with the task's cognitive complexity underlying the test items. Both of these are tied to the population for whom the diagnostic assessment is constructed. Most of CDMs define attributes as latent binary

variables, although the existence of polytomous category attribute models is not excluded (Templin, 2004; Von Davier & Yamamoto, 2004; Chen & de la Torre, 2013)
Given a test with items j = 1,…J, one can map single or multiple necessary attributes required by each item for a positive response. Under this design, experts can generate a matrix of relations between items and attribute required to solve these items, with the elements of the matrix indicating whether the k^{th} attribute is required for successful response on jth item. Such a J × K matrix of zeros and ones is referred to as a Q-matrix, with elements $\{q_{jk}\}$, where q_{jk} =1 denotes that the k^{th} attribute is necessary for success on item j and 0, the attribute is not necessary for the item (Tatsuoka, 1983).

DINA Model

The DINA (deterministic input noisy and) model (Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977) model, one of the most parsimonious CDMs that require only two interpretable item parameters, is the foundation of other models applied in cognitive diagnostic tests (Tatsuoka C. , 2002). The DINA model is a non-compensatory, conjunctive CDM, and assumes that an examinee must know all the required attributes in order to answer an item correctly (Henson, Templin, & Willse, 2009). An examinee mastering only some of the required attributes for an item will have the same success probability as another examinee possessing none of the attributes. For each item, the examinee item respondents are scored into two latent classes: one class indicates answering the item correctly (scored 1), containing examinees who possess all attributes required for answering that item correctly; the other class indicates incorrectly answering the item (scored 0), containing examinees who lack at least one of the required attributes for answering that item correctly. This feature is true for any number of attributes specified in the Q-matrix (de la Torre J. , 2009). The complexity of the DINA model is not influenced by the number of attributes measured by a test because its parameters are estimated for each item but not for each attribute, unlike other non-compensatory conjunctive cognitive diagnostic models (e.g., the RUM) (Rupp & Templin, 2008a). The DINA model has two item parameters, slip (s_j) and guess (g_j). The term “slip” refers to the probability of an examinee possessing all the required attributes but failing to answer the item correctly. The term “guess” refers to the probability of a correct response in the absence of one or more required attributes. However, the two item parameters also encompass other nuisances. Those nuisances confound the reasons why examinees who have not mastered some required attributes can answer an item correctly, and the reasons why examinees who have mastered all the required attributes can miss the correct response. Two examples of the common nuisances are the misspecifications in the Q-matrix, and the usage of alternative strategies, as Junker and Sijtsma (2001) described when they first advocated the DINA model. Below are the two item parameters:

$$g_j = P(X_{ij} = 1 | \eta_{ij} = 0) \tag{1.1}$$

$$s_j = P(X_{ij} = 0 | \eta_{ij} = 1) \tag{1.2}$$

and the item response function in the DINA model is defined as

$$P_j(\alpha_i) = P(\mathrm{X}_{ij} = 1 | \alpha_i) = g^{1-\eta_{ij}}(1-s_j)^{\eta_{ij}} \tag{1.3}$$

where the η_{ij} matrix refers to a matrix of binary indicators showing whether the examinee attribute profile pattern *i* has mastered all of the required skills for item *j*. The formula is defined as:

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \tag{1.4}$$

where α_{ik} refers to the binary mastery status of the k^{th} skill of the i^{th} skill pattern (1 denotes mastery of skill *k*, and 0 denotes non-mastery). And, as discussed in the previous section, q_{jk} here is the Q-matrix entries specifying whether the *j*th item requires the *k*th skill. The value of this deterministic latent response, η_{ij}, is zero if an examinee is missing at least one of the required attributes. ∏ indicates that the expression following it is multiplied across all attributes from attribute 1 (α = 1) to attribute K. If an attribute is not measured by an item, then q_{ik}= 0, which implies that the value of α_{ik} does not matter. If an attribute is measured, then q_{ik}= 1, which implies that it matters whether α_{ik}= 0 or α_{ik} =1.

Table 1: *Response Probabilities in the DINA Model*

	$X_{ik}=1$ Correct response	$X_{ik}=0$ Incorrect response
$\eta_{ij}=0 \Leftrightarrow \alpha_l^{'}q_j < q_j^{'}q_j$ (Non-mastery of at least one measured attribute)	g_j	$1-g_j$
$\eta_{ij}=1 \Leftrightarrow \alpha_l^{'}q_j = q_j^{'}q_j$ (Mastery of all measured attributes)	$1-s_j$	s_j

Specifically, an examinee belonging to a latent class in which all measured attributes are mastered $(\eta_{ij}=1)$ will answer item i correctly with probability $1-s_j$, that is, they will answer correctly if they do not slip. An examinee belonging to a latent class in which at least one of the measured attributes is not mastered $(\eta_{ij}=0)$ will answer correctly with probability g_j, i.e. they will only answer the item correctly by guessing. Under the DINA model, the probability of responding correctly to an item can only increase by mastering all measured attributes; the item response probability does not increase incrementally for each additional attribute mastered. The model predicts that the performance of examinees having mastered only some of the measured attributes is equivalent to those not having mastered any of the measured attributes.

Analyzing a DINA model requires test content specialists to first construct a Q-matrix to specify which item measures the appropriate attributes, similar to implementing many other CDMs. However, many CDM analyses assume that the specification of a Q-matrix is correct (or true), without verifying its suitability statistically. An incorrectly specified Q-matrix would mislead the results of the analysis. If the results show a model misfit because of an inappropriate Q-matrix, the misfit issue is hard to detect and solve (de la Torre, 2008). Hence, de la Torre (2008) proposed a sequential EM-based δ-method for validating the Q-matrices when implementing the DINA model. In his method, δ_j is defined as "the difference in the probabilities of correct responses between examinees in groups $\eta_j = 1$ and $\eta_j = 0$" (i.e., examinees with latent responses 1 and 0) (as cited in de laTorre, 2008, p. 344). δj serves as a discrimination index of item quality that accounts for both the slip and guessing parameters. Below is the computation formula for item *j*:

$$\delta_j = 1 - s_j - g_j \tag{1.5}$$

The higher the guessing and/or slip parameters are, the lower the value of δ_j. This signifies that the less-discriminating items have high guessing and slip parameters, and have a smaller discrimination index value of δ_j. In contrast, an item that perfectly discriminates between examinees in groups $\eta_j = 1$ and $\eta_j = 0$ has a discrimination index of $\delta_j = 1$ because there is no guessing and slip. Therefore, the higher the value of δ_j is, the more discriminating the item.

Q-matrix misspecification literature

The Q-matrix plays a critical role in cognitive assessment development. Below we review the current work of researchers investigating the effects of misspecifications of the Q-matrix. Henson and Templin (2009) point out that, while constructing the Q-matrix is the most crucial and difficult step in DCM, it is often taken for granted. It is assumed that experts correctly identify exactly the skills needed; no more, no less. However, this assumption may not always be true, and the consequences of violating it can be seen in model parameter estimates, classification rates for examinees, and overall model fit. The appropriateness of the Q-matrix is often overlooked and as a result poorly fitting models due to Q-matrix misspecification cannot be identified as poorly fitting or corrected (de la Torre, 2008). Because of these concerns, studies are now being conducted to determine the consequences of Q-matrix misspecification across a range of conditions. The impact of Q-matrix misspecification will greatly influence how a cognitive model in its entirety fits the data. Additionally, the effects of this misspecification will be different for a conjunctive model than it is for a disjunctive model. For non-compensatory models, a Q-matrix is properly defined if the attributes specified as 1s in the Q-matrix are all needed for giving the maximum probability of correctly answering each item and only those attributes are required. In a compensatory model, at least one of the attributes must be a 1 to give the maximum probability of correct response (de la Torre, 2008).

When a Q-matrix is not properly specified it can be due to these three types of misspecifications: underspecified, over specified, or combination of both. In an underspecified q-vector (i.e., Q-matrix row vector), entries of ‘1’ are recoded as ‘0’ so that fewer model parameters are estimated for the item under consideration. An over specified q-vector entry of ‘0’ are recoded as ‘1’ so that parameters that represent pure noise are unduly estimated. The misspecification of a Q-matrix would lead to undesirable consequences, e.g., poor model fit, inaccurate model parameter estimation (e.g. Henson and Templin, 2009; Rupp and Templin 2008a), and incorrect interpretations of the set of user-specified attributes. Therefore, the development of validation methods to access the specification accuracy of an existing Q-matrix by learning it from empirical data is important for the successful implementation of DCMs.

In the current use of cognitive diagnostic assessment, the Q-matrix is considered known (fixed) as solely determined by content specialists through theory. Recent studies have shown that considering the Q-matrix as known may be misleading. For example, DeCarlo (2011) highlights the ongoing debate about the true Q-matrix of Tatsuoka's fraction subtraction data (Tatsuoka, 1983) following the findings in studies such as, de la Torre and Douglas, 2004; de la Torre, 2008; and Henson et al., 2009. In fact, Templin & Henson (2006) recognize the need to investigate empirical techniques for determining the entries of the Q-matrix. In their own words, "Techniques that allow the empirical data to mold the entries of the Q-matrix would provide helpful feedback for the construction of reliable instruments developed for use with cognitive diagnosis models." An intuitive method would be to compare model fit indices among models with possible Q-matrices. However, this method involves intense computation. For an assessment with I items and K attributes, there are 2^{K*I} possible Q-matrices and the model fit indices for 2^{K*I} models need to be compared. As the K and I items get large, the number of possible Q-matrices increases exponentially, and so is the computation involved.

Rupp and Templin (2008a) examined the effect of the Q-matrix misspecification on parameter estimates and skill classification accuracy using the DINA model. The authors generated a known and true Q-matrix and then created several modified Q-matrices by deleting or adding extra items. Comparisons of models using the true Q-matrix were then compared to models using the modified Q-matrices. Their results showed that when an extra skill is required, the slip parameter will be inflated while the guess parameter remains somewhat unaffected. However, when a required skill is omitted from the Q-matrix the guess parameter will be overestimated while the slip parameter remains somewhat unaffected (Rupp & Templin, 2008a). Items on an assessment, like individuals, have skill patterns or combinations. Each item may require from one to all of the skills in any combination. In order for individuals to be accurately classified, it is important that their skill pattern be reflected in at least one item on an assessment. In their simulation study, Rupp and Templin (2008a) also found that when an individual's skill pattern was not represented by any item on an assessment that individual was completely misclassified. Im & Corter (2011) investigated the statistical consequences of attribute misspecifications in the rule space method. Two types of attribute misspecifications were examined in the study: the exclusion of an essential attribute and the inclusion of a superfluous attribute. The results showed that the exclusion of an essential attribute tends to lead to underestimation of examinees' mastery probabilities for the remaining attributes, while the inclusion of a superfluous attributes generally leads to overestimation of attribute mastery for the other attributes.

Although some research now exists for understanding the consequences of Q-matrix misspecification little research has focused on developing methods for detecting and correcting a misspecified Q-matrix.

Q-matrix validation literature

For example, de la Torre (2008) proposed an empirically based sequential search method to validate a Q-matrix. The search algorithm is based on the comparison of correct response probabilities between two specific groups of people. With reasonable computation time, the method is able to correct a misspecified Q-matrix under two conditions: 1) the response data is modeled by a DINA model; 2) the number of misspecified q-vectors is small compared to the number of items in the assessment. The method defined by de la Torre searches for a Q-matrix that maximizes the difference in the probabilities of a correct response to an item between examinees who possess all the skills required for a correct response to that item and examines who do not. de la Torre named his method, the delta method. The delta δ_j is the difference in the probabilities of the correct response between examinees who have mastered skill j and those who have not mastered skill j. Delta is therefore, a discrimination index for item j in that items that are highly discriminating have high δ_j values while those that are not highly discriminating have low δ_j values. This δ_j changes as the q-vector of an item changes. Using the delta method in conjunction with the DINA model, de la Torre conducted a simulation study and showed that the EM -based delta method can correctly replace the q-vectors that were misspecified in the Q-matrix with the correct ones while simultaneously retaining the q-vectors that had been correctly specified, most of the time. de la Torre concluded that although the delta method provides statistical information about the Q-matrix, it should not be used in isolation but in conjunction with substantive knowledge and domain expertise. In other words, the delta method for validating the Q-matrix does not replace theory. The delta procedure seems best characterized as a method for modestly adjusting an existing Q matrix, rather than a method for deriving a Q matrix as such. In his method the skills are fixed, and only the set of items posited to require the skill is adjusted. This method was only tested on the DINA model and it is unknown if this method will work equivalently when the observed item responses conform to models other than the DINA model.

Following this study, de la Torre & Chiu (2016) extended the sequential EM based delta method to the generalized DINA (G-DINA) model. The G-DINA model uses a more flexible parameterization. Based on different parameterizations, the G-DINA model can be converted to a class of reduced CDMs such as DINA and DINO models, additive CDMs, linear logistic models, and reparameterized unified models. In this study, a generalization of the discrimination index δj was proposed, represented by index ς_j^2 . The researchers pointed out that “a correct q-vector will yield homogeneous latent groups in terms of the probability of success [within-group probabilities] and therefore will result in groups with the highest variability of probabilities of success given a parsimonious subset of attributes” (p.10). Based on the sequential search algorithm and five CDMs, a simulation study displayed that all misspecified Q-vectors were accurately identified and 39 replaced while the correct Q-vectors were retained. The findings indicated the viability of the general index ς_j^2 for validating the Q-matrix. However, for both indices δj and ς_j^2, the researchers also pointed out that these statistical methods cannot be used in isolation, whereas they should be implemented with other methods such as information about the items, or expert knowledge, to create a more integrative framework for selecting and validating a Q-matrix.

DeCarlo (2012) proposed a Bayesian model-based method for Q-matrix validation for the R-DINA model, a reparameterization of the DINA model. DeCarlo's method requires both a provisional Q-matrix based on expert input and advance knowledge and/or identification of the possible misspecified entries in the Q-matrix. His method specifies prior distributions for the suspected misspecified Q-matrix entries reflecting the uncertainty regarding whether correctly completing a task depends on possessing a particular skill (DeCarlo, 2012). More specifically, DeCarlo (2012) treated the suspected misspecified Q-matrix entries as Bernoulli random variables while holding the rest of the Q-matrix entries as fixed. Later these random variables are estimated simultaneously with the other parameters in the model. The method was tested in the simulation studies and showed that the posterior distributions for the random Q-matrix elements provided useful information about which elements should be or should not be included. The method recovered uncertain elements of the Q-matrix quite well in a number of simulation conditions with the rest of the elements correctly specified. But as DeCarlo cautioned, further studies are needed to assess the robustness and generalizability of the method. However, this research set the foundation for possibly using the Bayesian framework to estimate the complete Q-matrix without an initial provisional expert based Q-matrix. That is the Q-matrix is derived solely based on students' responses without any expert input. Along those lines Chung (2014) extended DeCarlo's (2012) Bayesian approach to estimate the whole Q-matrix of the

DINA model and the reduced reparameterized unified model (rRUM). A saturated multinomial model was used to estimate the correlated attribute patterns for the DINA and rRUM models. Closed forms of the posteriors for guess and slip parameters were derived for the DINA model. The random walk Metropolis-Hastings algorithm was applied for parameter estimation in the rRUM. Additional exploratory research on estimating the Q-matrix without any expert input was done by Xiang (2013). Xiang (2013) set up a mathematical framework to estimate the true Q-matrix based on item response data. The research evaluates the method through simulation studies and applies it to estimate Q matrix from real item response data.

Some researchers have attempted to develop nonparametric methods of Q-matrix validation that do not rely on the estimation of model parameters. Problems encountered in the development of these methods include ambiguities in determining the attributes needed to respond correctly to test items, and, as was the case for the model-based methods, heavy computational burdens. For example, the hill-climbing algorithm proposed by Barnes (2010) attempted to reconstruct the Q-matrix of a test directly from examinees' observed item responses. Unfortunately, the algorithm often terminates with the estimated q-entries having values between 0 and 1, limiting the usefulness of the estimated Q-matrix for cognitive diagnosis.

Based on the DINA model, Tu, Cai, & Dai (2012) developed another method (γ method) to validate the Q-matrix. The indexes for validating the Q-matrix were the slip and guessing parameters and the score differences between the groups mastering and without mastering the attributes. Specifically, when (1) the guessing value of an item was too big, greater than the critical guessing values, and (2) the item score of the group mastering attribute k was not significantly different from that of the group without mastering attribute k, that is, the effect size was less than .20, then it suggested that attribute k probably was a unnecessary attribute for the item, and the element “1” of the Q-matrix was changed to “0”. On the other hand, when (1) the slip value of an item was too big, greater than the critical values, and (2) the item score of the group mastering attribute k was significantly different from the non-mastering group, with a effect size of .20 or greater, then attribute k probably was a necessary attribute for the item, and the element “0” of the Q-matrix was changed to “1”. In the third situation, when both the guessing and slip were too big, two or more attribute entries for an item needed to be changed. As the study of Q-matrix misspecification by Baker (1993), Tu et al. (2012) designed six Q-matrices with different percent of miss-specified elements, from zero, 5%, 40 to 25%. The miss-specified entries of the Q-matrices based on the simulated data were randomly selected and incorrectly re-specified. The researchers supposed that there were three levels of the observed response error ratio, which were 5%, 10%, and 15% of the expected examinee response patterns. They also set up five critical values for the slip and guessing parameters, from .10 to .30. Through comparing the original Q-matrices with the modified Q-matrices, Tu et al. found that (1) there was not any modification for the error-free Q-matrix; (2) when the critical values for the slip and guessing were .20, .25, and .30, the γ method effectively improved the Q-matrix specifications; and (3) the γ method was sensitive to the wrong Q-matrices when the critical values of the slip and guessing were small, while it was not sensitive when the critical values were high: the lower critical values of the slip and guessing led to more incorrect modifications of the Q-matrix, while the higher critical values of the slip and guessing resulted in less or no modifications of the wrong Q-matrices. Also, Tu et al. (2012) compared the γ method with de la Torre’s (2008) EM-based δ-method using the same Q-matrix. They generated the very similar results. Considering the δ-method was based on the complex and sequential EM computation, the γ method was relatively simpler. In addition, the γ method raised the correct ratios of cognitive diagnosis, which were measured by the marginal match ratio and pattern match ratio. The study indicated that the γ method was proved to be an effective method for validating a Q-matrix. However, the researchers pointed out that the γ method should be used with the experts' views together; the slip and guessing greater than the critical values did not mean the Q-matrix must be wrong.

Chiu (2013), defines a method that minimizes the residual sum of square (RSS) between the real responses and the ideal responses that follow from a given Q-matrix. The algorithm adjusts the Q-matrix by choosing the item with the worst RSS over to the data, and replaces it with the one has the lowest RSS, and iterates until convergence. Liu, Xu, & Ying (2012**)** stated that under the DINA or DINO model, if a Q-matrix is correctly specified, the Euclidean distance between expected proportions of positive responses to all items and a model-based combination of items and the corresponding observed proportions converges to zero in probability. After which they suggest a procedure can be setup to validate an existing Q-matrix by checking the closeness of the Euclidean distance between the above two vectors to zero. Kunina-Habenicht, Rupp, & Wilhelm (2012) examined the effects of model misspecification due to Q-matrix misspecifications on item parameter estimation and respondent classification within a broader DCM framework.

SEM Modification Indices

The three basic methods of model modification are the likelihood ratio or chi-square difference, Lagrange multiplier (LM) and Wald tests (Ullman, 2006). All are asymptotically equivalent under the null hypothesis but approach model modification differently. A major difference between the three methods (LR, Wald, and Score) is that they require different models to be estimated. The LR test requires both the restricted and unrestricted models to be estimated. The Wald test requires only the unrestricted model to be estimated and the Score test requires only the restricted model to be estimated. While the LM test asks which parameters, if any, should be added to a model, the Wald test asks which, if any, could there be any parameters that are currently being estimated that could instead be fixed to zero? Or, equivalently, which parameters are not necessary in the model? The Wald test is analogous to backward deletion of variables in stepwise regression, in which one seeks a nonsignificant change in R^2 when variables are left out. If the goal is the development of a parsimonious model the Wald test would be used to evaluate deletion of unnecessary parameters.

The LM test also compares nested models but requires estimation of only one model, the reduced model. This makes the LM test a computationally efficient test for model under-specification, that is, for testing whether the addition of certain parameters would significantly improve model fit. This modification index widely used in structural equation modeling is in fact a one degree of freedom score statistic (Sorbom, 1989). The LM test asks would the model be improved if one or more of the parameters in the model that are currently fixed were estimated. Or, equivalently, what parameters should be added to the model to improve the fit of the model?

In the present paper a new test statistic is developed based on the Lagrange Multiplier for use in the DCM framework. In the same way that the LM test compares nested models but requires the estimation of only one model this new test statistic will require that only one model be estimated (the Null model). In addition to the development of this new test statistic, this dissertation will also evaluate the performance of this new test statistic for the detection of diagnostic model misspecifications of the Q-matrix, which include both underfitting the Q-matrix (i.e., specifying 0s where there should be 1s) and overfitting the Q-matrix (i.e., specifying 1s where there should be 0s).

CHAPTER 3: METHODOLOGY

EM algorithm for the DINA model

The response probabilities in the DINA model are expressed through the following item response function:

$$P_j(\alpha_i) = P(X_{ij} = 1 \mid \alpha_i) = P\left(X_{ij} = 1 \mid \eta_{ij}\right) g_j^{1-\eta_{ij}} (1-s_j)^{\eta_{ij}} \ \log\frac{1-s_j}{s_j} = \log\frac{g_j}{1-g_j} + \psi + \frac{z_j}{3} \qquad \textit{(1.6)}$$

$$P_j\left(\alpha_l\right) = \left\{\begin{matrix} g_j & if\,\alpha_l^{'}q_j < q_j^{'}q_j \rightarrow \eta_{lj} = 0 \\ 1-s_j & if\,\alpha_l^{'}q_j = q_j^{'}q_j \rightarrow \eta_{lj} = 1 \end{matrix}\right\} \qquad \textit{(1.7)}$$

where the η_{ij} matrix refers to a matrix of binary indicators showing whether the examinee attribute profile pattern *i* has mastered all of the required skills for item *j*. That is, the DINA model involves only two probabilities g_j and $1-s_j$ for responding item *j*. The probability of positively responding to item *j* increases from g_j to $1-s_j$ only if examinees possess all required skills and thus η_{ij} is non-zero.

Parameter estimation of the DINA model requires both item parameters and skill profiles to be estimated together since in practice skill attribute pattern are unobservable and item parameters are unknown. One way to accomplish this would be to use joint maximum likelihood, as seen in traditional item response models' joint maximization of item parameters (structural parameter) and the skill profiles (incidental parameter) may yield inconsistent estimations of the item parameters (Baker, 1993). Additionally, in cases where the number of attributes is fairly large, the number of attribute patterns can be enormous resulting in a saturated model that may be computationally inefficient. The optimal option would be to use the marginal maximum likelihood (MML) estimation which would be implemented by the expectationmaximization (EM) algorithm. A major benefit of using the EM algorithm is that it provides desirable convergence behavior and has simple estimation steps.

In order to aid in the discussion of the proposed test statistic I restate the major steps of the EM Algorithm originally presented by de la Torre (2009). As shown by de la Torre (2009) for estimating the DINA model, the marginalized log-likelihood of the response vector of examinee *i* is maximized with respect to the item parameters $\beta_{j\eta}\begin{Bmatrix}\beta_{j0}=g_j\\ \beta_{j1}=s_j\end{Bmatrix}$. The goal of the maximum marginal likelihood estimation (MMLE) is to find the parameter values that maximize the marginalized likelihood function $L\left(X\right)$, or more conveniently, its logarithm form,

$l\left(X\right)=\log\prod_{i=1}^{I}L(X_i)$.

$$\begin{aligned} L\left(X\right)&=\prod_{i=1}^{I}L\left(X_i\right)=\prod_{i=1}^{I}\sum_{l=1}^{L}L\left(X_i\,|\,\alpha_l\right)p\left(\alpha_l\right)\\ l\left(X\right)&=\log\prod_{i=1}^{I}L(X_i)\\ &=\sum_{i=1}^{I}\log L(X_i)\end{aligned}\tag{1.8}$$

Where $L\left(X_i\right)$ is the marginalized likelihood of the response vector of the ith examinee; $L\left(X_i\,|\,\alpha_l\right)=\prod_{j=1}^{J}P\left(\alpha_l\right)^{X_{ij}}\left[1-P\left(\alpha_l\right)^{1-X_{ij}}\right]$ is the conditional likelihood of response vector of examinee *i* conditioned on the attribute vector, $p\left(\alpha_l\right)$ is the prior probability of the skills vector α_l and $L=2^K$ is the number of possible combinations of K attributes. We begin by the individual posterior distribution which can be deduced using the Bayes' theorem:

$$P\left(\alpha\ \ |\,X_i\right)=\frac{P\left(X_i\,|\,\alpha_l\right)P\left(\alpha_l\right)}{\sum_{l=1}^{L}P\left(X_i\,|\,\alpha_l\right)P\left(\alpha_l\right)},l=1,\ldots,L\tag{1.9}$$

Where $P\left(\alpha_l \mid X_i\right)$ is the posterior probability that examinee *i* has the attribute pattern α_l. From the posterior probabilities, the following expected counts are derived:

1. $I_l = \sum_{i=1}^{I} P\left(\alpha_l \mid X_i\right)$ is the expected number of examinees with attribute pattern α_l
2. $R_{jl} = \sum_{i=1}^{I} P\left(\alpha_l \mid X_i\right) X_{ij}$ is the expected number of examinees with attribute pattern α_l answering item *j* correctly.

In the next step, the first derivative of the log-likelihood with respect to the item parameters is set to zero. This derivative involves only the two counts obtained in the earlier step and thus allows for updating the item parameters. Below are the significant steps involved in this process:

$$\frac{\partial l(X)}{\partial \beta_{jn}} = \sum_{i=1}^{I} \frac{1}{L(X_i)} \sum_{l=1}^{L} p(\alpha_l) \frac{\partial L(X_i \mid \alpha_l)}{\partial \beta_{jn}} \qquad \textit{(1.10)}$$

$$\begin{aligned}\frac{\partial L\left(X_i \mid \alpha_l\right)}{\partial \beta_{j\eta}} &= \prod_{j' \neq j} P_{j'}\left(\alpha_l\right)^{X_{ij'}} \left[1 - P_{j'}\left(\alpha_l\right)\right]^{1-X_{ij'}} \frac{\partial P_j\left(\alpha_l\right)^{X_{ij}} \left[1 - P_j\left(\alpha_l\right)\right]^{1-X_{ij}}}{\partial \beta_{j\eta}} \\ &= \left[\prod_{j=1}^{J} P_j\left(\alpha_l\right)^{X_{ij}} \left[1 - P_j\left(\alpha_l\right)\right]^{1-X_{ij}}\right] \frac{\partial P_j\left(\alpha_l\right)}{\partial \beta_{j\eta}} \left[\frac{X_{ij} - P_j\left(\alpha_l\right)}{P_j\left(\alpha_l\right)\left[1 - P_j\left(\alpha_l\right)\right]}\right] \qquad \textit{(1.11)} \\ &= L\left(X_i \mid \alpha_l\right) \frac{\partial P_j\left(\alpha_l\right)}{\partial \beta_{j\eta}} \left[\frac{X_{ij} - P_j\left(\alpha_l\right)}{P_j\left(\alpha_l\right)\left[1 - P_j\left(\alpha_l\right)\right]}\right]\end{aligned}$$

Next after some interchanging and substituting into the $\frac{\partial l(X)}{\partial \beta_{jn}}$ equation we get the following:

$$\frac{\partial l(X)}{\partial \beta_{jn}} = \sum_{l=i}^{L} \frac{\partial P_j\left(\alpha_l\right)}{\partial \beta_{j\eta}} \left[\frac{1}{P_j\left(\alpha_l\right)\left[1 - P_j\left(\alpha_l\right)\right]}\right] \left[\sum_{i=1}^{I} p\left(\alpha_l \mid X_i\right) X_{ij} - P_j\left(\alpha_l\right) \sum_{i=1}^{I} P\left(\alpha_l \mid X_i\right)\right] \textit{(1.12)}$$

$$\frac{\partial l(X)}{\partial \beta_{jn}} = \sum_{l=i}^{L} \frac{\partial P_j\left(\alpha_l\right)}{\partial \beta_{j\eta}} \left[\frac{1}{P_j\left(\alpha_l\right)\left[1 - P_j\left(\alpha_l\right)\right]}\right] \left[R_{jl} - P_j\left(\alpha_l\right) I_l\right] \qquad \textit{(1.13)}$$

$$\frac{\partial l(X)}{\partial \beta_{jn}} = \sum_{\alpha_l : \alpha_l' q_j < q_j' q_j} \frac{\partial P_j\left(\alpha_l\right)}{\partial \beta_{j\eta}} \left[\frac{1}{P_j\left(\alpha_l\right)\left[1 - P\left(\alpha_l\right)\right]}\right] \left[R_{jl} - P_j\left(\alpha_l\right) I_l\right] + \sum_{\alpha_l : \alpha_l' q_j = q_j' q_j} \frac{\partial P_j\left(\alpha_l\right)}{\partial \beta_{j\eta}} \left[\frac{1}{P_j\left(\alpha_l\right)\left[1 - P\left(\alpha_l\right)\right]}\right] \left[R_{jl} - P_j\left(\alpha_l\right) I_l\right] \textit{(1.14)}$$

Earlier we saw that the simple form of the DINA model partitioned the attribute space into two parts. The first portion represents the group that is missing at least one required attribute to answer item *j* ($\eta_{lj} = 0$) correctly or specifically $\alpha_l^{'} q_j < q_j^{'} q_j$ and therefore their probability of success is $\partial P_j(\alpha_l) = g_j$. The second portion represents the group that has all the required attributes ($\alpha_l^{'} q_j = q_j^{'} q_j$) to get item *j* correct therefore their probability of success is $\partial P_j(\alpha_l) = 1 - s_j$.

Therefore, using the above information we get the following:

$$\frac{\partial l(X)}{\partial \beta_{jn}} = \frac{\partial g_j}{\partial \beta_{j\eta}} \left[\frac{1}{g\,[1-g_j]} \right] \sum_{\alpha_l : \alpha_l^{'} q_j < q_j^{'} q_j} \left[R_{jl} - P_j(\alpha_l) I_l \right] + \frac{\partial 1 - s_j}{\partial \beta_{j\eta}} \left[\frac{1}{(1-s_j)s_j} \right] \sum_{\alpha_l : \alpha_l^{'} q_j = q_j^{'} q_j} \left[R_{jl} - P_j(\alpha_l) I_l \right] \quad \textit{(1.15)}$$

- I_{jl}^{0} is the expected number of examinees lacking at least one of the required attributes for item *j*.
- I_{jl}^{1} is the expected number of examinees with all the required attributes for item *j*.
- R_{jl}^{0} is the expected number of examinees among I_{jl}^{0} that correctly answer item *j*.
- R_{jl}^{1} is the expected number of examinees among I_{jl}^{1} that correctly answer item *j*.
- $I_{jl}^{1} + I_{jl}^{0} = I_l$ for all item *j*.

$$\frac{\partial l(X)}{\partial \beta_{jn}} = \frac{\partial g_j}{\partial \beta_{j\eta}} \left[\frac{1}{g_j\left[1-g_j\right]} \right] \left[R_{jl}^{0} - g_j I_{jl}^{0} \right] + \frac{\partial\left(1-s_j\right)}{\partial \beta_{j\eta}} \left[\frac{1}{\left(1-s_j\right)s} \right] \left[R_{jl}^{1} - \left(1-s_j\right) I_{jl}^{1} \right] \quad \textit{(1.16)}$$

For the maximization of $\partial l(X)$ with respect to $\beta_{j\eta} \begin{Bmatrix} \beta_{j0} = g_j \\ \beta_{j1} = s_j \end{Bmatrix}$ the above equation simplifies to

solving for g_j in $\left[\frac{1}{g_j\left(1-g_j\right)} \right] \left[R_{jl}^{0} - g_j I_{jl}^{0} \right] = 0$ and solving for s_j in

$-\left[\frac{1}{\left[1-s_j\right]s}\right]\left[R_{jl}^1-\left[1-s_j\right]I_{jl}^1\right]=0$. Therefore, the MLE estimators are $\hat{g}_j=\frac{R_{jl}^0}{I_{jl}^0}$ and $\hat{s}_j=\frac{\left[I_{jl}^1-R_{jl}^1\right]}{I_{jl}^1}$. Next, we outline the exact procedure followed in the EM algorithm.

In the beginning the EM algorithm iterates using initial item parameters and skill distribution parameters which have to be chosen. Then, the EM algorithm alternates between the E-step and the M-step. In E-step the expected counts for each item and each group [I_l & R_{jl}] are derived from the posterior. This is a prerequisite for the calculation of the required statistics in the M-step. Next, the M-step involves updating the parameter estimates $\left\{\begin{matrix}\beta_{j0}=g_j\\ \beta_{j1}=s_j\end{matrix}\right\}$ for the DINA using maximization methods. The item parameter estimates for item *j* are updated according to the previously calculated counts [I_l & R_{jl}] since the $\hat{g}_j=\frac{R_{jl}^0}{I_{jl}^0}$ and $\hat{s}_j=\frac{\left[I_{jl}^1-R_{jl}^1\right]}{I_{jl}^1}$ are a function of only these two counts which were obtained in the E-step. Then the skill class distributions $P\left(\alpha_l\right)$ are updated. Finally, the E-step and M-step alternated until a previously set convergence criterion is attained. The convergence criterion used for estimating the DINA model was that the maximum difference between the previous and the current parameter estimate was smaller than 0.0001 (de la Torre, 2008). Other commonly used criteria include the absolute log-likelihood convergence criterion and the relative log-likelihood convergence criterion (Muthen & Muthen, 1998-2010).

Detailed look at de la Torre's Method of Q-matrix Validation

For item *i,* its correct q-vector, q_i, could be one of the 2^K-1 possible attribute patterns. The q-vector could not be a vector with all 0 elements. Let be a binary vector of length *K,*

$l=1,2,\ldots., 2K-1$, and let δ_{il} be the difference in probabilities of correct responses for item *i* between respondents who have mastered the required attributes when item *i'*s q-vectors is specified as α_l and respondents who are lack of at least one of the required attributes, that is $\delta_{il} = P(Y_{ij} = 1 | \eta_{lj} = 1) - P(Y_{ij} = 1 | \eta_{lj} = 0)$. The correct q-vector for item *i* is then defined (de la Torre, 2008) as the binary vector that maximizes δ_{il}.

$$q_i = \underset{\alpha_l}{\arg\max}(\delta_{il}) \tag{1.17}$$

Table 3 shows how to identify the correct q-vector using this definition for a hypothetical item in a four-attribute domain as shown in Table 2. This hypothetical item requires the first and the second attribute, and guessing and slipping parameter under the DINA model of 0.2 and 0.2, respectively.

Table 2: A Hypothetical Item under the DINA Model in a Four-Attribute Domain

Attribute					
1	2	3	4	Guessing	Slipping
1	1	0	0	.2	.2

Table 3 lists the *δ* for all the 16 attribute patterns in a four-attribute domain. The second and third column list the membership to the two groups ξ=1 and ξ=0, and probability of providing a correct answer to the hypothesized item for examinees with the corresponding attribute patterns. Here, a flat distribution is assumed for the attribute pattern distribution, i.e. every respondent is equally likely to be classified into one of the 16 possible attribute patterns. The last three columns provide the probabilities of correct answers in the two groups $\eta_{ij} = 1$ and $\eta_{ij} = 0$, and their difference *δ*. These probabilities are found as the mean probabilities of correct response for the two groups given the item q-vector is the corresponding q-vector. Take pattern 6 for instance, examinees with 12 patterns (pattern 1- 5, 7-11, 13, 14) fall in group $\eta_{ij} = 0$ when the item q-vector is the attribute pattern (1, 1, 0, 0). The mean probability of correct response in $\eta_{ij} = 0$ is thus found by adding the true correct response probabilities for these 12 patterns and dividing the sum by 12 (2.4/12 = 0.2). Examinees with pattern 6, 12, 15, 16 fall into group $\eta_{ij} = 1$ and the mean probability of correct response is given by (0.8+0.8+0.8+0.8)/4=0.8. Based on Table 3, the largest δ is obtained at the true item q-vector.

Table 3: Probabilities of Correct Response for a Hypothetical Item under the DINA Model in a Four-Attribute Domain

Patterns	Under True Q-vector		Attributes				Probability of Correct Response		d
	ξ	$P(X=1\mid\xi)$	1	2	3	4	$\xi=1$	$\xi=0$	
1	0	0.2	0	0	0	0	-	-	-
2	0	0.2	1	0	0	0	0.5	0.2	0.3
3	0	0.2	0	1	0	0	0.5	0.2	0.3
4	0	0.2	0	0	1	0	0.35	0.35	0
5	0	0.2	0	0	0	1	0.35	0.35	0
6	1	0.8	1	1	0	0	0.8	0.2	0.6
7	0	0.2	1	0	1	0	0.5	0.3	0.2
8	0	0.2	1	0	0	1	0.5	0.3	0.2
9	0	0.2	0	1	0	1	0.5	0.3	0.2
10	0	0.2	0	1	1	0	0.5	0.3	0.2
11	0	0.2	0	0	1	1	0.35	0.35	0
12	1	0.8	1	1	1	0	0.8	0.29	0.51
13	0	0.2	0	1	1	1	0.5	0.33	0.17
14	0	0.2	1	0	1	1	0.5	0.33	0.17
15	1	0.8	1	1	0	1	0.8	0.29	0.51
16	1	0.8	1	1	1	1	0.8	0.32	0.48

Searching for the correct q-vector by definition is straightforward but computationally intensive. As *K* increases, the number of attribute patterns increases exponentially. A more efficient algorithm, the sequential search algorithm (de la Torre, 2008), was proposed for searching for correct q-vectors. This algorithm starts by comparing the δ for all single-attribute patterns. The attribute resulting in the largest δ is selected as one of the attributes in the q-vector. Then, all two-attribute patterns with the first selected attribute are compared by their δs. The second attribute is chosen based on two criteria: (1) its corresponding two-attribute pattern has the largest δ, say $\delta_{(2)}$; and (2) $\delta_{(2)} > \delta_{(1)}$. If $\delta_{(2)} < \delta_{(1)}$, it is unnecessary to include a second attribute.

The process stops here and the correct q-vector is a single-attribute pattern. The process continues in the same way to choose the rest of the attributes. Table 4 demonstrates how this sequential method works for the item in Table 2.

Table 4: Selecting Attribute based on δ

	Attributes δ			
Number of Attributes	1	2	3	4
One	.3	.3	0	0
Two (including $\alpha 1$)	-	.6	.2	.2
Three(including $\alpha 1, \alpha 2$)	-	-	.51	.51

As shown in Table 4, δs are compared for all single-attribute patterns at the first step. Both the pattern with the first attribute only and that with the second attribute only have the highest difference of 0.3, and either one can be included in the correct q-vector. Suppose we choose the first attribute for this step. In the second step, δs are compared for those two-attribute patterns which include the first attribute. At this step, the second attribute is picked because it has the highest difference 0.6, and it is larger than the highest difference in the first step (0.3). A third attribute is not needed because the largest difference for q-vectors with three attributes is 0.51, which is less than 0.6. Thus, the correct q-vector is (1,1,0,0). The correct q-vector found using the sequential method agrees with that from the definition.

Proposed Method of Q-matrix validation

The delta sequential search method (de la Torre, 2008) was shown to successfully correct Q-matrix misspecifications when response data was modeled by the DINA model. This work extends this research by developing an inference test that utilizes the delta statistic to test for Q matrix misspecifications in the DINA model. This technique will empirically test, inform and therefore augment but not replace the theory used to derive a Q-matrix. Currently researchers like Templin and Henson (2006) recognize the need to investigate empirical techniques for validating the entries of the Q-matrix mainly because a proposed Q-matrix by content experts may not be identical to the ‘true’ Q-matrix, even if the experts carefully take into account instructional purposes and students’ protocols.

Next let us consider how the Q-matrix of the DINA model might be improved. This test provides a means to assess the strength of the relationship between an item and the skill it is posited to measure and therefore useful in the detection of Q matrix misspecification. This provides a new way to improve the DINA model fit in a manner similar to the role of modification indices in Structural equation models (SEM). Traditionally in SEM literature modification indices are used as a guide in determining which parameters to add to the model so as to significantly improve model fit based on the empirical data. Every model requires making certain assumptions, modification indices can be used to help evaluate how reasonable these assumptions are by giving the researcher a sense of what happens when those assumptions are relaxed. In a comparable manner, the goal of this research is to develop and evaluate a method for significance testing for the Q-matrix entries normally assumed as given.

We begin by describing a hypothesis test that will allow us to utilize de la Torre’s delta for hypothesis testing:

$$\begin{array}{lcl} H_0 : q_{jk} = 0 & & H_0 : q_{jk} = 1 \\ H_A : q_{jk} = 1 & \text{or} & H_A : q_{jk} = 0 \end{array}$$

The Q-matrix represents the relationship between the skills and each item on a test. It is a binary *J* x *K* matrix where *J* represents the total number of items and *K* represents the total number of skills assessed by a test. Therefore, a value of $q_{jk} = 0$ indicates knowledge of skill *k* is not necessary to answer item *j* and $q_{jk} = 1$ implies knowledge of the k^{th} skill is necessary to answer item *j*.

Next, we denote $A_j\left(q_j\right)$ as the set of skill profiles α that have all the skills necessary to answer item *j* with q-vector q_j and $\overline{A_j}\left(q_j\right)$as the set of skill profiles that lack at least one of the skills to answer item *j*. These skill profiles α are called pseudo-empirical mainly because they are not observed. Subsequently, we define in equation 1.18 the delta $\left[\delta_j\left(q_j\right)\right]$method which compares the difference in proportions between those who have mastered all the skills ($\alpha_l{}' q_j = q_j' q_j$ or $\alpha \in A_j\left(q_j\right)$) to those who lack at least one skill ($\alpha_l{}' q_j < q_j' q_j$ or $\alpha \in \overline{A}_j\left(q_j\right)$) to answer item j correctly.

$$\delta_j\left(q_j\right) = \hat{P}\left\{X_j = 1 \mid \alpha_l{}' q_j = q_j' q_j\right\} - \hat{P}\left\{X_j = 1 \mid \alpha_l{}' q_j < q_j' q_j\right\} = \hat{P}\left\{X_j = 1 \mid \alpha \in A_j\left(q_j\right)\right\} - \hat{P}\left\{X_j = 1 \mid \alpha \in \overline{A}_j\left(q_j\right)\right\} \qquad (1.18)$$

The probability to get item *j* correct given they possess all the required attributes is $\hat{P}\left\{X_j = 1 \mid \alpha \in A_j\left(q_j\right)\right\}$ or $\hat{P}\left\{X_j = 1 \mid \alpha_l{}' q_j = q_j' q_j\right\}$ and is defined as the estimated proportion of individuals that got item j correct given they have all the necessary skills. Whereas, the probability to get item *j* correct given they do not possess all the necessary skills $\hat{P}\left\{X_j = 1 \mid \alpha_l{}' q_j < q_j' q_j\right\} = \hat{P}\left\{X_j = 1 \mid \alpha \in \overline{A}_j\left(q_j\right)\right\}$ is equivalent to estimated proportion of individuals that got item j correct given they do not possess at least one skill necessary to answer item *j*. These probabilities can also be defined with respect to estimated proportions which look familiar because de la Torre defines them in the EM algorithm for estimating the DINA model parameters. During the E step these posterior probabilities were calculated using the expected counts for each item and each group$\left(I_{jl}^0, I_{jl}^1\right)$& $\left(R_{jl}^0, R_{jl}^1\right)$.

$$\hat{P}\left\{X_j = 1 \mid \alpha_l{}' q_j = q_j' q_j\right\} = \hat{P}\left\{X_j = 1 \mid \alpha \in A_j\left(q_j\right)\right\} = \frac{\sum_{i=1}^{I} \sum_{\alpha_l{}' q_j = q_j' q_j} X_{ij} P\left(\alpha_l \mid X_i\right)}{\sum_{i=1}^{I} \sum_{\alpha_l{}' q_j = q_j' q_j} P\left(\alpha_l \mid X_i\right)} = \frac{R_{jl}^1}{I_{jl}^1} \qquad \textit{(1.19)}$$

$$\hat{P}\left\{X_j = 1 \mid \alpha_l^{'} q_j < q_j^{'} q_j\right\} = \hat{P}\left\{X_j = 1 \mid \alpha \in \overline{A}_j\left(q_j\right)\right\} = \frac{\sum_{i=1}^{I} \sum_{\alpha_l^{'} q_j < q_j^{'} q_j} X_{ij} P\left(\alpha_l \mid X_i\right)}{\sum_{i=1}^{I} \sum_{\alpha_l^{'} q_j < q_j^{'} q_j} P\left(\alpha_l \mid X_i\right)} = \frac{R_{jl}^{0}}{I_{jl}^{0}} \quad \textit{(1.20)}$$

In a similar manner, the delta for item *j* can also be defined with respect to these expected counts as shown in the equation below:

$$\delta_j\left(q_j\right) = \frac{\sum_{i=1}^{I} \sum_{\alpha_l^{'} q_j = q_j^{'} q_j} X_{ij} P\left(\alpha_l \mid X_i\right)}{\sum_{i=1}^{I} \sum_{\alpha_l^{'} q_j = q_j^{'} q_j} P\left(\alpha_l \mid X_i\right)} - \frac{\sum_{i=1}^{I} \sum_{\alpha_l^{'} q_j < q_j^{'} q_j} X_{ij} P\left(\alpha_l \mid X_i\right)}{\sum_{i=1}^{I} \sum_{\alpha_l^{'} q_j < q_j^{'} q_j} P\left(\alpha_l \mid X_i\right)} = \frac{R_{jl}^{1}}{I_{jl}^{1}} - \frac{R_{jl}^{0}}{I_{jl}^{0}} \quad \textit{(1.21)}$$

In order to derive the test statistic, we describe in detail the method for hypothesis testing for a single Q-entry. One possible hypothesis for altering a single Q-entry is $\begin{matrix} H_0 : q_{jk} = 0 \\ H_A : q_{jk} = 1 \end{matrix}$. The null (*H₀*) here says that the kth attribute is not necessary to master in order to answer item *j* correctly while in the alternate model the kth attribute is necessary. The null and alternate models are nested meaning that one model is a subset of another. This is analogous to a modification index which is a one degree of freedom chi-square test of the addition of a new parameter or the deletion of a parameter. Similarly, this hypothesis test is comparing nested models that either add or delete a single Q-entry while holding everything else constant including parameter estimates. Next, as shown in Table 5 we identify A_{jk}^{0} as the set of skill profiles α that have mastered all the skills necessary to answer item *j* correctly under the null hypothesis (*H_0*). The null hypothesis says that the kth attribute is not necessary and or required for mastery in order to answer item *j* correctly therefore the skill profiles in this set includes those that do not have mastery of the kth skill ($\alpha_k = 0$). In contrast A_{jk}^{1} is the set of skill profiles that have mastered all the skills necessary to answer item *j* correctly under the alternate hypothesis (*H_A*) which posits that the kth attribute is necessary to master in order to answer item *j* correctly. This set of skill profiles will require everyone to have mastered the kth skill in order to be included. Therefore the $\overline{A}^{1}_{jk}$ (the opposite of the set A^{1}_{jk}) would represent all the skill profiles that lack at least one of the skills to answer item *j* under H_A which would also include those that do not have mastery of the kth skill ($\alpha_k = 0$) and/or another skill necessary to answer item *j*.

Table 5: Skill profiles Sets for Null and Alternate Hypotheses

A^{0}_{jk}	Master all the skills for item j under the H_0. Includes those that do not have mastery of the kth skill (α_k=0) because it is not a necessary skill to get item j correct.
A^{1}_{jk}	Master all the skills for item j under the H_A. Excludes anyone that does not master the kth skill (α_k=0).
$\overline{A}^{1}_{jk}$	Does not master all the skills necessary to answer item j under the H_A. Includes skill profiles where someone does not master the kth skill (α_k=0).
$B_{jk} = A^{0}_{jk} \cap \overline{A}^{1}_{jk}$	Master all the skills to answer item j under the H_0 and those that have all the other skills except the kth skill (α_k=0) to answer item j correctly.

The last and final set we define is the intersection of the following two sets: $\beta_{jk} = A^{0}_{jk} \cap \overline{A}^{1}_{jk}$. This would include only those elements which are in both sets; this set represents only those skill profiles which have mastered all the skills *except* the kth skill ($\alpha_k = 0$) necessary to answer item *j*.

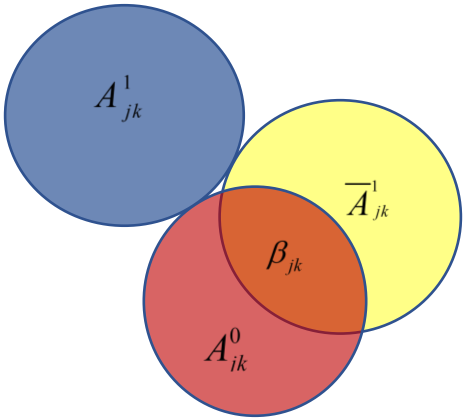


Figure 1: *Euler Diagram depicting the skill sets of the Null & Alternate Hypotheses*

Example using a Q-matrix entry

To illustrate this more clearly an example of a single Q-matrix Entry is briefly described. The Null q-entry represents the expert designed Q-entry that states that only the first skill is necessary to answer Item 8 correctly. Below are the details for the q-entry for Item 8 both at the Null and the alternate hypotheses. This tests a single Q-entry (q_{82}) while the rest of Q-matrix is assumed to be correct.

H_0: $q_{82} = 0$

	Skill 1	Skill 2	Skill 3
Item 8	1	0	0

H_A: $q_{82} = 1$

	Skill 1	Skill 2	Skill 3
Item 8	1	1	0

The group of concern is β_{82}^{0} because if the students in this group are able to answer item 8 correctly then the null hypothesis is supported as shown in Table 6. Whereas if the group of students in β_{82}^{0} are not able to answer item 8 correctly often enough then there will be enough evidence to reject the null hypothesis.

Table 6: Skill sets for a single Q-entry

Pattern		Master all the skills for item 8 under the H_0. Includes those that do not have mastery of the 2nd skill because it is not a necessary skill to get item 8 correct.	Master all the skills for item 8 under H_A. Excludes anyone that does not master the 2nd skill.	Does not master atleast one of the skills necessary to answer item 8 under the H_A. Includes skill profiles where someone does not master the 2nd skill.	Master all the skills to answer item 8 under the H0 and those that have all the other skills except the 2nd skill to answer item 8 correctly.
		A_{82}^{0}	A_{82}^{1}	$1-A_{82}^{1}$	β_{82}^{0}
1	000	0	0	1	0
2	010	0	0	1	0
3	001	0	0	1	0
4	110	1	1	0	0
5	011	0	0	1	0
6	111	1	1	0	0
7	100	1	0	1	1
8	101	1	0	1	1

Subsequently, the test statistic is derived by calculating the difference in the delta vector under both the null and alternate hypotheses. The difference of the deltas as shown in equation

1.22 below is calculated using the expected counts which are derived using the MLE parameter estimates under the null hypothesis.

$$\delta_{jk}^{Alt} - \delta_{jk}^{Null} = \left[R_{jl}^{1(Alt)} \Big/ I_{jl}^{1(Alt)} - R_{jl}^{0(Alt)} \Big/ I_{jl}^{0(Alt)} \right] - \left[R_{jl}^{1(Null)} \Big/ I_{jl}^{1(Null)} - R_{jl}^{0(Null)} \Big/ I_{jl}^{0(Null)} \right] \qquad \textit{(1.22)}$$

If we recall both the null and alternate models are nested and since the model is fit assuming the null we can approximate these expected counts using the EM algorithm estimate of the DINA model parameters $\left\{ \hat{s}_j = \frac{\left[I_{jl}^{1} - R_{jl}^{1} \right]}{I_{jl}^{1}} \quad \hat{g}_j = \frac{R_{jl}^{0}}{I_{jl}^{0}} \right\}$ of the null model.

The alternate & null model counts for the expected examinees correctly answering item j among the total examinees with all the required attributes for item j would approximate $\approx 1 - \hat{s}_j$. However, for the counts measuring the proportion of expected examinees correctly answering item j among the total examinees that lack at least one required attribute to answer item j the null and alternate have differences. Since we are assuming that the null hypothesis ($H_0 : q_{jk} = 0$) is correct then this proportion is really the guess ($\approx g_j$) for the Null but has a slightly higher value $\approx g_j + e$ for the alternate model. Therefore, when we calculate the difference in the $\delta_{jk}^{Alt} - \delta_{jk}^{Null}$ we would find that it will approximate some small term (*e*) which really represents the overestimation of the guessing parameter for the Alternate hypothesis due to the inclusion of an unnecessary skill for item *j*.

$$\begin{aligned} \delta_{jk}^{Alt} - \delta_{jk}^{Null} &= \left[R_{jl}^{1(Alt)} \Big/ I_{jl}^{1(Alt)} - R_{jl}^{0(Alt)} \Big/ I_{jl}^{0(Alt)} \right] - \left[R_{jl}^{1(Null)} \Big/ I_{jl}^{1(Null)} - R_{jl}^{0(Null)} \Big/ I_{jl}^{0(Null)} \right] \\ &= R_{jl}^{1(Alt)} \Big/ I_{jl}^{1(Alt)} - R_{jl}^{1(Null)} \Big/ I_{jl}^{1(Null)} + R_{jl}^{0(Null)} \Big/ I_{jl}^{0(Null)} - R_{jl}^{0(Alt)} \Big/ I_{jl}^{0(Alt)} \qquad \textit{(1.23)} \\ &= \left[\left(1 - \hat{s}_j\right) - \left(1 - \hat{s}_j\right) + g - \widehat{g + e} \right] \\ &= e \end{aligned}$$

Table 7: Hypothesis Testing for a single Q-entry

	$R_{jl}^{1} \big/ I_{jl}^{1}$ Proportion of expected examinees correctly answering item j among the total examinees with all the required attributes for item j.	$R_{jl}^{0} \big/ I_{jl}^{0}$ Proportion of expected examinees correctly answering item j among the total examinees that lack at least one required attribute to answer item j.
If we assume that ***H₀***: $q_{jk} = 0$ is correct		
$\delta_{jk}^{Null} = R_{jl}^{1(null)} \big/ I_{jl}^{1(null)} - R_{jl}^{0(null)} \big/ I_{jl}^{0(null)}$	$\approx 1 - \hat{s}_j$	$\approx g_j$
$\delta_{jk}^{Alt} = R_{jl}^{1(Alt)} \big/ I_{jl}^{1(Alt)} - R_{jl}^{0(Alt)} \big/ I_{jl}^{0(Alt)}$	$\approx 1 - \hat{s}_j$	$\approx g_j + e$
If we assume that ***H_A:*** $q_{jk} = 1$ is correct		
$\delta_{jk}^{Null} = R_{jl}^{1(Null)} \big/ I_{jl}^{1(Null)} - R_{jl}^{0(Null)} \big/ I_{jl}^{0(Null)}$	$\approx 1 - \hat{s}_j$	$\approx g_j$
$\delta_{jk}^{Alt} = R_{jl}^{1(Alt)} \big/ I_{jl}^{1(Alt)} - R_{jl}^{0(Alt)} \big/ I_{jl}^{0(Alt)}$	$\approx 1 - \hat{s}_j + d$	$\approx g_j$

Next if we were to assume that the alternate hypothesis ($H_A : q_{jk} = 1$) is correct then we would find that the $\delta_{jk}^{Alt} - \delta_{jk}^{Null}$ will approximate *d*, which is really the increase in the slipping parameter due to the omission of an important skill necessary to answer item *j*.

$$\begin{aligned}
\delta_{jk}^{Alt} - \delta_{jk}^{Null} &= \left[{R_{jl}^{1(Alt)}} \Big/ {I_{jl}^{1(Alt)}} - {R_{jl}^{0(Alt)}} \Big/ {I_{jl}^{0(Alt)}} \right] - \left[{R_{jl}^{1(Null)}} \Big/ {I_{jl}^{1(Null)}} - {R_{jl}^{0(Null)}} \Big/ {I_{jl}^{0(Null)}} \right] \\
&= {R_{jl}^{1(Alt)}} \Big/ {I_{jl}^{1(Alt)}} - {R_{jl}^{1(Null)}} \Big/ {I_{jl}^{1(Null)}} + {R_{jl}^{0(Null)}} \Big/ {I_{jl}^{0(Null)}} - {R_{jl}^{0(Alt)}} \Big/ {I_{jl}^{0(Alt)}} \qquad (1.24) \\
&= \left[\left(1 - \hat{s}_j + d\right) - \left(1 - \hat{s}_j\right) + \widehat{g - g} \right] \\
&= d
\end{aligned}$$

As seen in equation 1.25 below this value of *d* is essentially the difference in the proportion of expected examinees correctly answering item j among the total examinees with all the required attributes for item j under the alternate and null models. In this study, the *d* statistic shown below is the key test statistic which will be approximated via simulation and used for identifying Q-matrix misspecification. It is important to note that both the *d* and *e* have a one to one correspondence and a summary of this is shown in Table 7. The manner in which this statistic mirrors modification indices is that all the quantities used to calculate the test statistic are calculated from the Null model. Only a single model fit is necessary.

$$\begin{aligned} d_{jk} &= R_j^{1(Alt)} \Big/ I_j^{1(Alt)} - R_j^{1(Null)} \Big/ I_j^{1(Null)} \\ d_{jk} &= R_j^{1(Alt)} \Big/ I_j^{1(Alt)} - \left(1 - s_j\right) \end{aligned} \qquad (1.25)$$

Large values of d_{jk} provide evidence to reject the null hypothesis (H_0). Since the d_{jk} statistic can be viewed as a sum of random variables it is believed to be asymptotically normal with a mean of 0 and some variance V_d under H_0 because of the central limit theorem. The proposed research will develop an estimator of the asymptotic variance V_d based on the Fisher Information Matrix of the slip and guess parameters. Simulation studies will be used to test how well the normal approximation and the estimate of the variance works for a variety of sample sizes and different model conditions. The following section describes the simulations in more detail.

Simulation Design

The present paper involves two simulation studies and one empirical study. The first simulation study is designed to find the true Type-I error rates for the test with a nominal level $\alpha = .05$ under different scenarios. The second study is a power analysis which shows how often the proposed test can correctly find that the Q-matrix in the estimated model is both under or over specified.

The simulated data sets consisted of responses from four sample sizes (500/1,000/2,000/5,000) of examinees on 16 items with each loading on one of the combinations of four fine-grained attributes. Table 8 below shows the true Q-matrix used to generate the simulated data sets. The true Q-matrix is complete (Liu, Xu & Ying, 2011), i.e., for each attribute, there exists an item only requiring that attribute.

Table 8: Q-matrix used for Simulation

Item	Attribute/Skill			
Number	1	2	3	4
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1
5	1	0	0	0
6	0	1	0	0
7	0	0	1	0
8	0	0	0	1
9	1	1	0	0
10	0	1	1	0
11	0	0	1	1
12	0	1	0	1
13	1	0	0	1
14	0	0	1	1
15	0	1	0	1
16	1	0	0	1

In the simulation study five hundred datasets are generated using the above Q-matrix and analyzed for each of the 36 conditions (4x3x3) which include four levels of sample size, three levels of correlation among attributes and three levels of item discrimination.

Factor 1: Sample size (4 levels)

The four different sample sizes are 500, 1,000, 2,000 and 5,000.

Factor 2: Item discrimination (3 levels)

For each data set the guess parameter is randomly sampled for all 16 items from a uniform distribution shown below.

$$g_j \sim U\left(0, \frac{1}{2}\right) \textit{ for all } j = 1, \dots, 16 \qquad (1.26)$$

The slip parameters for the first eight items were then generated from a $U\left(0,1-g_j\right)$ to ensure that individuals with the required skills were more likely to get the item correct than those without the skill. The slip parameters for items 8 to 16 were generated to satisfy the following:

$$\log\frac{1-s_j}{s_j}=\log\frac{g_j}{1-g_j}+\psi+\frac{z_j}{3} \qquad (1.27)$$

where z is a standard normal deviate and ψ is discrimination level. We examined three discrimination levels $\psi_1 = 1$, $\psi_2 = 2.5$ and $\psi_3 = 4$. For a given data set, the discrimination level is constant across all items.

Factor 3: Attribute associations (3 levels)

The possession of the attributes could be correlated or uncorrelated according to different situations. The artificial data in the simulation studies of the present paper were created assuming that the attributes were correlated. To create associations among the four skills, the skill vectors α were generated by sampling from a multivariate probit distribution with a common correlation r between pairs of skills. The examined correlations equal to 0, .3, and .5.

Type I Error Study

The performance of the test statistic is evaluated by assessing how often the proposed statistic correctly finds that the hypothesized model is not underspecified or over specified, thus avoiding the addition of unnecessary attributes or the deletion of a necessary attribute. Specifically, the Type I error is defined as the probability of mistakenly rejecting a true null hypothesis. In order to attain the Type I error rate, the distribution of the proposed statistic under the null hypothesis approximated through simulation. For each of the 36 conditions 500 samples generated from the null model and then the values of the proposed statistic will be calculated. Finally, a histogram is used to depict the values of the proposed statistic from these 500 samples to serve as an approximation to the true null distribution. If the theoretical null distribution is correct, the estimated Type I error rate should be close to the nominal level.

Due to the potential for a large number of hypothesis tests in this research, it becomes important that some sort of multiplicity correction be used. In performing large series of statistical tests, some will have *P* values less than 0.05 purely by chance, even when all the null hypotheses are really true. For example, if you do 100 statistical tests, and for all of them the null hypothesis is actually true, you'd expect about 5 of the tests to be significant at the P<0.05 level, just due to chance. In that case, you'd have about 5 statistically significant results, all of which were false positives. Therefore, we need a way of adjusting p-values for the number of hypothesis tests performed in order to control the Type I error rate.

There are two methods that are currently being used to control the Type I error rate. The first is the Bonferroni correction and the second is the false discovery rate using the Benjamini-Hochberg procedure. Cribbie (2007) have argued in favor of the false discovery rate (Benjamini & Hochberg, 1995) over the Bonferroni correction or familywise error rate. A major drawback for the Bonferroni correction is that it is known to lead to higher rates of Type II errors or false negatives and thereby greatly reduces the power of the tests. In addition, the Benjamini-Hochberg procedure is considered a more powerful method because it is less sensitive than the Bonferroni procedure to your decision about what is a "family" of tests. If you increase the number of tests, and the distribution of *P* values is the same in the newly added tests as in the original tests, the Benjamini-Hochberg procedure will yield the same proportion of significant results.

Table 9: False Discovery rate (FDR)

	Null True	Alternate True	Total
Not significant	U	T	m-R
Significant	V	S	R
	m_0	$m-m_0$	M

V= # of Type I errors [False positives]

The Benjamini-Hochberg procedure starts by predetermining the λ which is the value chosen for the false discovery rate (FDR). As shown in Table 9 above the FDR is designed to control the proportion of false positives among the set of rejected hypotheses (R) and is defined by the following proportion, $FDR = V/R$. Next rank the individual unadjusted p-values in ascending order: $p_1 \le p_2 \le \ldots \le p_m$. Thirdly, calculate each individual p-value's Benjamini-Hochberg critical value, using the formula $BH_{critical\_value} = \left(j/m \right) * \lambda$, where j is the individual p-value's rank, m = total number of tests, and λ the FDR. Finally, compare your original p-values to the critical BH from the last step and declare the tests of rank 1, 2, …, j as significant when $p_j \le \left(j/m \right) * \lambda$.

Power Analysis

The power of a statistical significance test is the probability of correctly rejecting a false null hypothesis. In this study, the power will look at how often the proposed statistic correctly finds that the Q-matrix in the estimated model is under or over specified. This essentially means that the proposed method successfully adds missing attributes or deletes unnecessary ones. This will be accomplished by introducing q-matrix misspecifications which involve 'switching' an entry to the wrong value during the estimation stage and later measure how often the procedure will detect these misspecifications. Detection here is defined as how often a false null hypothesis is correctly rejected.

The number of q-matrix misspecifications will also be investigated within the power study. This will involve looking at how the number of misspecifications in a Q-matrix will affect the capability of the proposed method to recover the true Q-matrix by successfully adding missing attributes or deleting unnecessary ones. As mentioned earlier this will involve creating four types of q-matrices in which the first two will each have a single entry (1/[4x16]*100=1.6%), the remaining two will have two (2/64*100=3.13%) q-matrix entries that have been 'switched' to an incorrect value during the estimation stage. This will help in understanding how well the procedure holds up to increasing Q-matrix misspecification.

Empirical Study

The proposed method will also be applied to real data for the Q-matrix estimation. The fraction subtraction dataset is a well-known dataset in Q-matrix research and is widely analyzed. Tatsuoka's fraction subtraction data set is comprised of 536 rows and 20 columns, representing the responses of 536 middle school students to each of the 20 Fraction subtraction test items. Each row in the data set corresponds to the responses of a particular student. Value "1" denotes that a correct response was recorded, and "0" denotes an incorrect response. All test items are based on 8 attributes. The Q-matrix can be found in DeCarlo (2011), and it was also used by de la Torre and Douglas (2004).

Another version of the fraction subtraction data set consists of 15 items and 536 students. The Q-matrix was defined in the de la Torre (2009). There are five required attributes, including: (1) performing basic fraction-subtraction operation (2) reduce answers to simplest form, (3) separate a whole number from a fraction, (4) borrow from a whole number part, and (5) convert a whole number to a fraction. Despite the small sample size of this dataset this study will use the 15 items version for the empirical section.

This study develops a new test statistic for the detection of diagnostic model misspecifications of the Q-matrix, which include both underfitting the Q-matrix (i.e., specifying 0s where there should be 1s) and overfitting the Q-matrix (i.e., specifying 1s where there should be 0s). In addition to the development of this new test statistic, this dissertation will also evaluate the performance of this new test statistic and develop an estimator of the asymptotic variance V_d based on the Fisher Information Matrix of the slip and guess parameters. The test statistic will be evaluated by two simulation studies and also applied to the fraction subtraction dataset, a well-known dataset in Q-matrix research. The two simulation studies will be used to test how well the normal approximation and the estimate of the variance works for a variety of sample sizes and different model conditions.

CHAPTER 4: RESULTS

Chapter 2 defined the DINA model while reviewing the existing Q-matrix misspecification and validation literature and Chapter 3 introduced a test statistic designed for the detection of diagnostic model misspecifications of the Q-matrix, which include both underfitting and overfitting the Q-matrix. However, the utility of the test statistic for the detection of Q-matrix misspecifications has yet to be investigated. In this chapter, results are presented for the two simulation studies designed to evaluate the statistical properties of the test statistic, including its Type I error rate and power across different conditions. Next the results from the Shapiro Wilk's test of Normality to determine how well the normal approximation of the test statistic and the estimator of the asymptotic variance V_d work across the 36 conditions. The V_d is based on the Fisher Information Matrix of the slip and guess parameters. Finally, we present results of applying this test statistic on to the fraction subtraction dataset.

Simulation Studies

This simulation study involved generating 500 samples for each of the 36 conditions which as shown in Table 10 represent the three factors investigated; Correlation, Sample Size and Item discrimination (3x4x3). For each of the 36 conditions 500 samples were generated from the null model and then the values of the proposed statistic were calculated.

Methodology for the Type I Error Study

The Type I error rate is defined as the probability of mistakenly rejecting a true null hypothesis or how often the proposed statistic correctly finds that the hypothesized model is not underspecified or over specified, thus avoiding the addition of unnecessary attributes or the deletion of a necessary attribute. The true null distribution of the test statistic is approximated through these simulations for the 36 conditions and then it is compared to the posited theoretical Normal distribution. If the actual sampling distribution of a test statistic is well approximated by the posited Normal distribution then the estimated Type I error rate should be close to the nominal level (.05).

Table 10: Summary of 36 conditions investigated in the simulation studies

Condition	Correlation	Sample Size	Item Discrimination	Condition	Correlation	Sample Size	Item Discrimination
1	0	500	1	19	0.3	2,000	2.5
2	0	1,000	1	20	0.3	5,000	2.5
3	0	2,000	1	21	0.5	500	2.5
4	0	5,000	1	22	0.5	1,000	2.5
5	0.3	500	1	23	0.5	2,000	2.5
6	0.3	1,000	1	24	0.5	5,000	2.5
7	0.3	2,000	1	25	0	500	4
8	0.3	5,000	1	26	0	1,000	4
9	0.5	500	1	27	0	2,000	4
10	0.5	1,000	1	28	0	5,000	4
11	0.5	2,000	1	29	0.3	500	4
12	0.5	5,000	1	30	0.3	1,000	4
13	0	500	2.5	31	0.3	2,000	4
14	0	1,000	2.5	32	0.3	5,000	4
15	0	2,000	2.5	33	0.5	500	4
16	0	5,000	2.5	34	0.5	1,000	4
17	0.3	500	2.5	35	0.5	2,000	4
18	0.3	1,000	2.5	36	0.5	5,000	4

The simulation conditions reflect the balance between showing the promise of the test statistic under a variety of ideal conditions and the desire to reflect realistic conditions. Thereby informing us on whether it can be expected to perform well in practice. In the following section, details are provided as to the rationale behind the chosen simulation conditions. A summary of these conditions is given in Table 11. The number of examinees, items, attributes, level of attribute correlation and item discrimination chosen for these simulation studies are reflective of values currently used in practice in other studies reported in the CDM literature. For instance, de la Torre (2011) used 30 items, five attributes for two sample sizes (1,000 and 2,000). Both guess and slip were assumed to be 0.2 across all items and attributes were correlated at three levels (0.15, 0.3, 0.5). de la Torre (2008) used a sample size of 5,000 with a prevalence of all the attribute patterns set at 0.5 with attribute correlation set to zero. The parameters (slip and guess) were set at 0.2 for all items. The Q-matrix had an equal number of items with 1,2 or 3 attributes implying that attributes appear equal number of times. Chiu (2013) uses sample sizes ranging from 100 to 1,000 with guess and slip ranging from 0.2 to 0.5. Liu, Xu, and Ying (2012) used a Q-matrix with 20 items and three attributes in their simulation study, with sample sizes ranging from 500 to 4,000. The simulation study in Kunina Habenicht et al. (2012) contained conditions with both three and five attributes, tests with 25 and 50 items, and samples of 1,000 and 10,000 examinees. Most of the studies choose sample size to range from 500 to 5,000 and therefore in this study four levels of sample size are investigated that fall within these ranges. Correlation is also investigated by de la Torre (2011) and de la Torre (2008) in the ranges of 0 to 0.5 therefore this study investigated attribute correlation within 0 to 0.5. Although these studies don’t explicitly measure Item discrimination they do investigate guess and slip ranging from 0.2 to 0.5 with most studies fixing both slip and guess at 0.2, which parallels to an item discrimination of 2.5 in this study.

Table 11: Summary of Simulation Conditions for Type I and Power Studies.

Characteristic	Value
Cognitive Diagnosis Model	DINA
Number of Items	16
Number of Attributes/Skills	4
Q-matrix	Complete/Identified
Prevalence of Attribute/Skills	Uniform Distribution [0.3-0.7]
Number of Examinees	500; 1,000; 2,000; 5,000
Correlation among Attributes/Skills	0,0.3,0.5
Item Discrimination	1, 2.5, 4
Number of Simulation datasets	500

All conditions in the simulation study used 16 items and four attributes and was replicated 500 times. That is, in each of the 36 conditions 500 different samples were generated. The number of replications was large enough to obtain an accurate representation of the null distributions of the test statistic in consideration and to obtain long-run estimates of Type I error and power, yet small enough to complete the simulation studies in a reasonable amount of time. Additionally, to evaluate the Type I error and Power of the test statistic a variety of sample sizes, correlation between attributes and Item level discrimination are investigated. We sampled datasets with 500, 1,000, 2,000 and 5,000 examinees; no correlation (0), moderate correlation (0.3) and high correlation between attributes/skills. Finally sampled datasets where items 8-16 had an Item discrimination of 1 (guess and slip will approximate 0.4), an item discrimination of 2.5 is reflective of using guess and slip equal to 0.2 and lastly 4 represents items with very high level of item discrimination (guess and slip will be approximate 0.1).

The prevalence of the attributes/skills is randomly sampled from a uniform distribution on (0.3,0.7) for each sampled data set. For each simulation condition, the Q-matrix of the data generating model (DINA) was the matrix given in Table 8. This true Q-matrix is complete (Liu, Xu & Ying, 2011), i.e., for each attribute, there exists an item only requiring that attribute. In addition to being complete this Q-matrix is also identified. Identification refers to the relationship between what will be estimated (the parameters and attribute profiles) and the information used to derive these estimates (Q-matrix and examinees). A key benefit of having an identified and complete Q-matrix is that it allows to estimate a unique value for every parameter. Specifically, it allows for a set of items to consistently identify all types of attribute profiles for the DINA model (Chen, Liu, Xu, & Ying, 2015). It is usually recommended to use a complete Q-matrix (Chiu et al., 2009) (Liu et al., 2013).

Results for the Type I Error Analysis

To better illustrate the findings of the Type I error rate at the nominal α=0.05 for Sample Size, Correlation and Item discrimination groups based on three different types of item averages are compared. The first group consists of all q-entries in which the true q-entry is one and belong to items that require two skills to answer the item correctly. The second group consists of all q-entries where the true q-entry is zero and belongs to items that require two skills to answer the item correctly. Finally, the third group consists of all q-entries where the true q-entry is zero and belongs to items where only one skill is required to answer the item correctly. These group averages provide some interesting insights into how the three types of q-entries behave with respect to Sample size, Correlation and Item discrimination. The results as shown in Figure 2a clearly show that as the sample size increases the Type I error significantly reduces and moves closer to the nominal value of $\alpha = .05$. For group 1 a minimum sample size of 2,000 and for group 2 and 3 a minimum sample size of 5,000 is needed in order for the Type I error rate to reduce to 0.05. Interestingly Group 3 (Zero's appearing in items requiring only one skill to answer correctly) has a much lower Type I error rate across all sample sizes when compared to other groups. Type I error seems to behave differently for q-entries of items requiring one skill verses two skills with respect to sample size. Specifically, a sample size of 500 has a Type I error rate of 0.07 for group 3, a Type I error rate of 0.11 for group 1 and a Type I error rate of 0.13 for group 2. No relationship is seen between correlation and Type I error across all three groups. Item discrimination is a function of item level parameters guess and slip where an Item discrimination of 1 corresponds to a guess and slip of approximately 0.4, an item discrimination of 2.5 is reflective of using guess and slip equal to 0.2 and lastly 4 represents items with very high level of item discrimination (guess and slip will be approximate 0.1). Surprisingly, the results in Figure 2a show that the relationship between Type I error and Item discrimination only exists for Group 2 (all q-entries where the true q-entry is zero and belongs to items that require two skills to answer the item correctly). Groups 1 and 3 show no clear relationship between Item discrimination and Type I error rates. For Group 2 the most discriminating items (Item Discrimination of 4) have the largest Type I error rates and moves the furthest away from the nominal value of $\alpha = .05$. In Group 2 the Item discrimination of 1 and 2.5 showed the lowest Type I error rate of 0.07 and as the Item discrimination increases to 4 the Type I error rate also increases to 0.11. It is important to note that the increase in the Type I error rate is seen only when Item discrimination goes to 4.

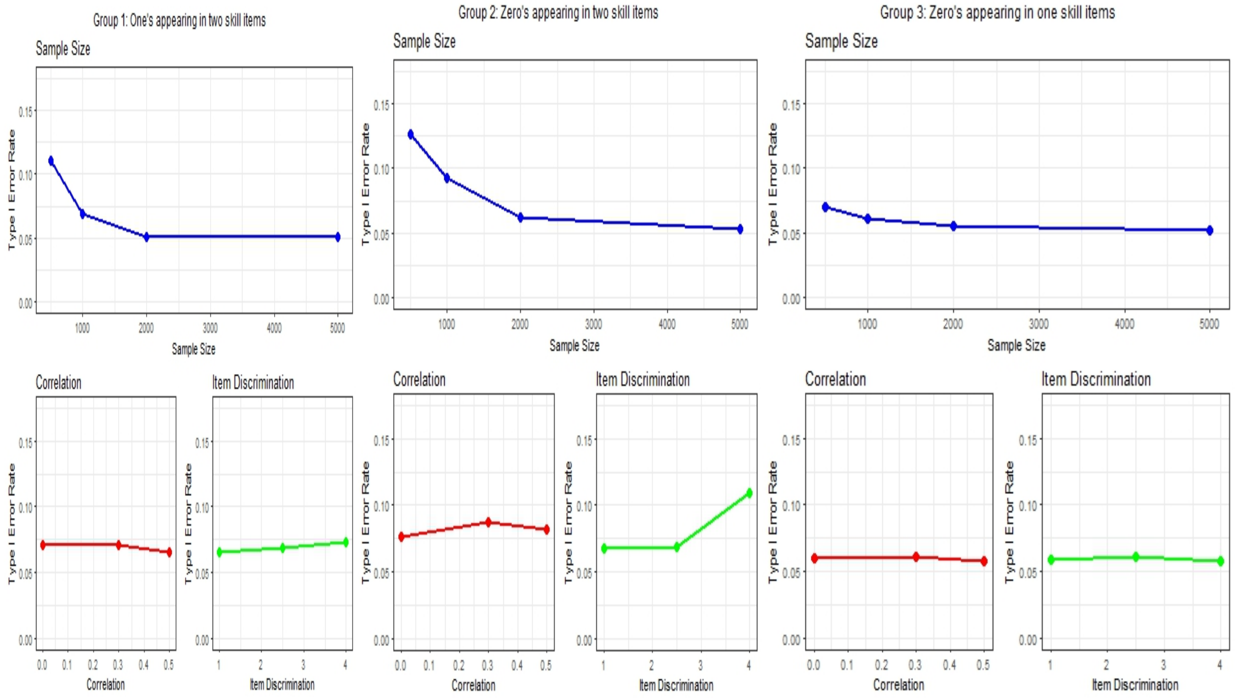


Figure 2a: *One-way Type I Error Analysis by Sample Size, Correlation, Item Discrimination for Group 1, 2 and 3*

In the next steps instead of going through each Q-vector entry individually a few representative q-vectors are chosen below to illustrate the general findings of the Type I error rate at the nominal α=0.05 for Sample Size, Correlation and Item discrimination for this discussion. Both skill 4 of Item 10 and skill 3 of Item 9 were chosen because they provide a good representation for the scenario where c. This scenario measures how well the statistic correctly finds that the Q-matrix in the hypothesized model is not underspecified, thus avoiding the addition of unnecessary attributes/skills. Whereas in the second scenario the true Q-entry is one for skill 1 of Item 9 and therefore measures how well the statistic correctly finds that the Q-matrix in the hypothesized model is not over specified, thus avoiding the exclusion of a necessary attribute/skill. It is supposed that a separate analysis of the true Q-entries that are zero verses one allows one to see whether differences exist for Type I error when it causes over specification (true q-vector is 0) verses under specification (true q-vector is 1). A complete list of all Type I Error rates by Item, Skill and factor for nominal $\alpha = \{0.01, 0.05, 0.10\}$ can be found in the appendix.

Case 1 (Item 10 Skill 4) & (Item 9 Skill 3): True Q-entry is 0

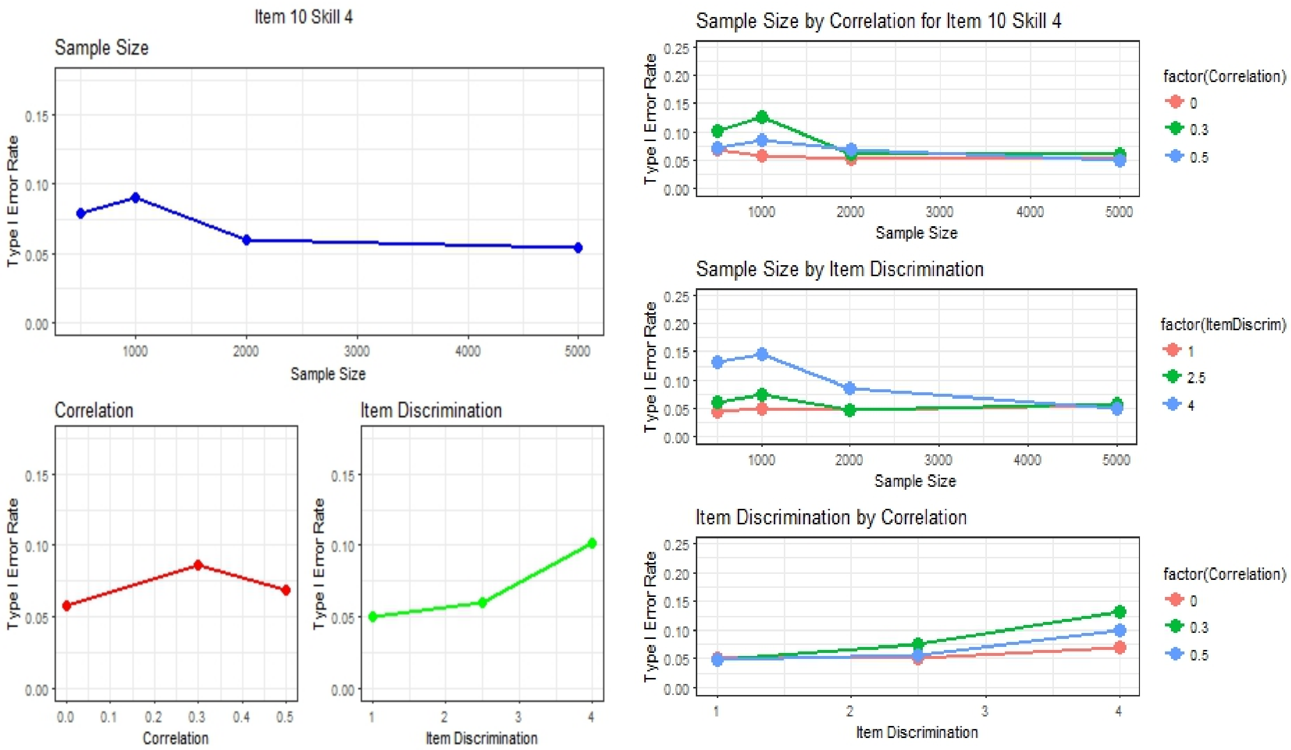


Figure 2a: *One-way & Two-way Type I error Analysis by Sample Size, Correlation, Item Discrimination for Skill 4 of Item 10*

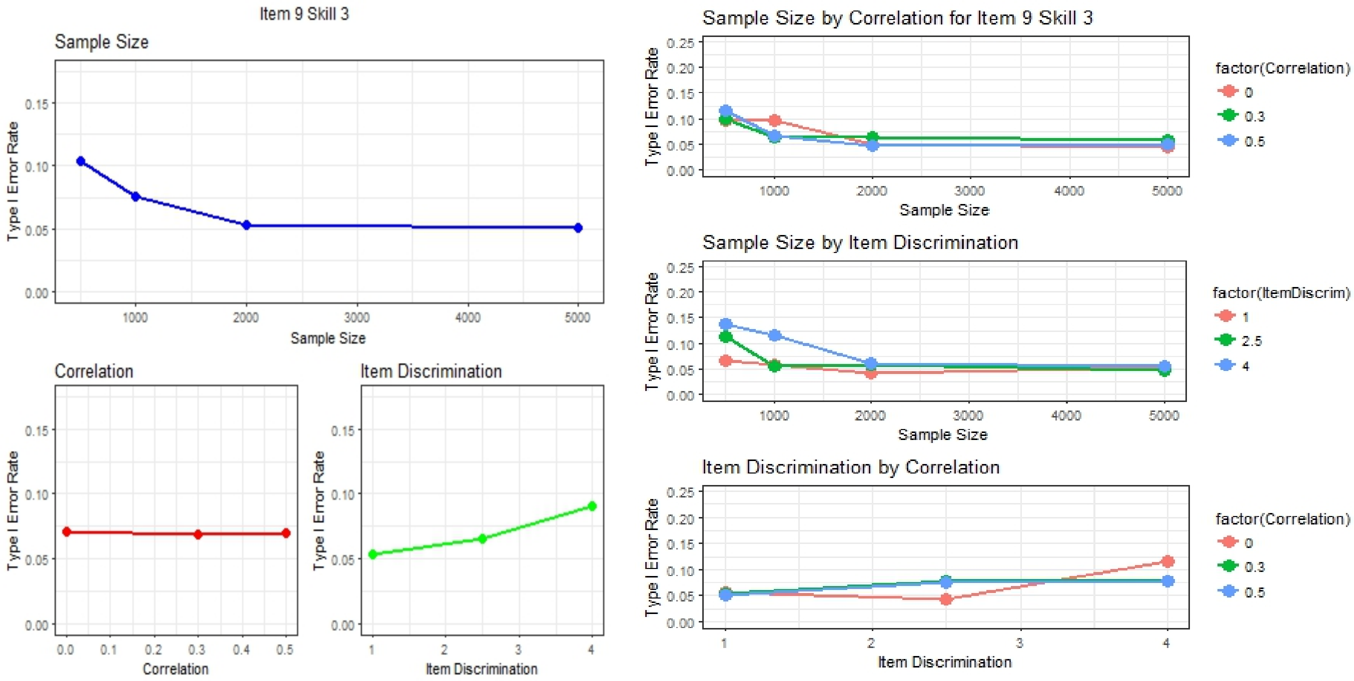


Figure 3: *One-way & Two-way Type I error Analysis by Sample Size, Correlation, Item Discrimination for Skill 3 of Item 9*

Sample Size

The results as shown in Figure 2 & Figure 3 above clearly show that as the sample size increases the Type I error significantly reduces and moves closer to the nominal value of $\alpha = .05$. This result was expected mainly because the statistic is asymptotically normal. In skill 4 of Item 10 the sample size of 500 and 1,000 behaved counter intuitively in that Type I error rate goes up slightly from 0.08 at sample size of 500 to 0.09 for sample size 1,000. However, as the sample size increases to 2,000 and 5,000 the Type I error rate continues to reduce from 0.06 to 0.05. Similarly, in skill 3 of Item 9 as sample size increases from 500 to 5,000 the Type I error rate reduces from 0.10 to 0.05. It is important to note that the most marked reduction in the Type I error rate is seen at a sample size of 2,000 in both Q-vectors. For instance, skill 4 of Item 10 the Type I error goes from 0.09 (at Sample size of 1,000) to 0.06 (at Sample Size of 2,000) and for skill 3 of Item 9 the Type I error rate goes from 0.10 (at Sample size of 500) to 0.05 (at Sample

Size of 2,000). Going from a Sample size of 2,000 to 5,000 shows a minimal reduction in Type I error rates for both q-vectors.

The two-way plots shown on the right side of Figure 3 and Figure 2 provide a deeper understanding of the relationship between Type I error rate and the two-way interaction between Sample size, Correlation and Item discrimination. Type I error rates for the interaction of Sample size and Correlation for both skill 4 of Item 10 and skill 3 of Item 9 are lowest when Sample size is large ($\geq 2,000$) and Correlation is zero. Once the sample size reaches 5,000 the Type I error rate approaches the nominal ($\alpha = .05$) across all levels of Correlation. At the two lowest Sample sizes of 500 and 1,000 and Correlation of 0.3 (medium level), skill 4 of Item 10 exhibits the highest Type I error rates of 0.10 and 0.13 respectively. However, for skill 3 of Item 9 the highest Type I error rate of 0.11 corresponds to high Correlation (0.5) and smallest Sample size (500).

Item Discrimination

As mentioned earlier Item discrimination is a function of item level parameters guess and slip where an Item discrimination of 1 corresponds to a guess and slip of approximately 0.4, an item discrimination of 2.5 is reflective of using guess and slip equal to 0.2 and lastly 4 represents items with very high level of item discrimination (guess and slip will be approximate 0.1). Interestingly the results in Figure 2 & Figure 3 above show that as the Item(s) become more discriminating the Type I error significantly increases and moves further away from the nominal value of $\alpha = .05$. In both skill 4 of Item 10 and skill 3 of Item 9 the Item discrimination of 1 showed the lowest Type I error rates of 0.05. As the Item discrimination increases from 1 to 4 the Type I error rate increase to 0.10 (skill 4 of Item 10) and 0.09 (skill 3 of Item 9). It is important to note that the most marked increase in the Type I error rate is seen when Item discrimination goes from 2.5 to 4.

Type I error rates for the interaction of Sample size and Item discrimination for both skill 4 of Item 10 and skill 3 of Item 9 are lowest when Item discrimination is 1 and Sample size is large ($\geq 2,000$). Interestingly at the lowest Item discrimination of 1 the Type I error rate approaches the nominal ($\alpha = .05$) across all levels of Sample size. As Sample size increases to 2,000 the Type I error rate approaches the nominal ($\alpha = .05$) across Item discrimination of 1 and 2.5. Yet, for an Item discrimination of 4 the Type I error rate approaches the nominal ($\alpha = .05$) only when the Sample size increases to 5,000.

Correlation

Correlation between attribute(s) can take the following values: no correlation (0), moderate correlation (0.3) and high correlation. For example, if Skills 1 and 3 have high correlation (0.5) this indicates that students who possess skill 1 tend to possess skill 3 as well. The results in Figure 2 & Figure 3 shows no clear relationship between Type I error rates and increasing attribute(s) correlation. For instance, in skill 4 of Item 10 having no correlation between attribute(s) had the lowest Type I error rate of 0.06 but this trend does not continue as correlation between attributes increases. However, for skill 3 of Item 9 the Type I error rate is 0.07 across the three correlation values showing no variation. The two-way graphs comparing Type I error rates by Item discrimination and Correlation actually show contradictory findings between the two q-vectors (skill 3 of Item 9 & skill 4 of Item 10). They show that increasing Item discrimination leads to higher Type I error rates but no clear trend exists with respect to Correlation.

Case 2 (Item 9 Skill 1): True Q-entry is 1

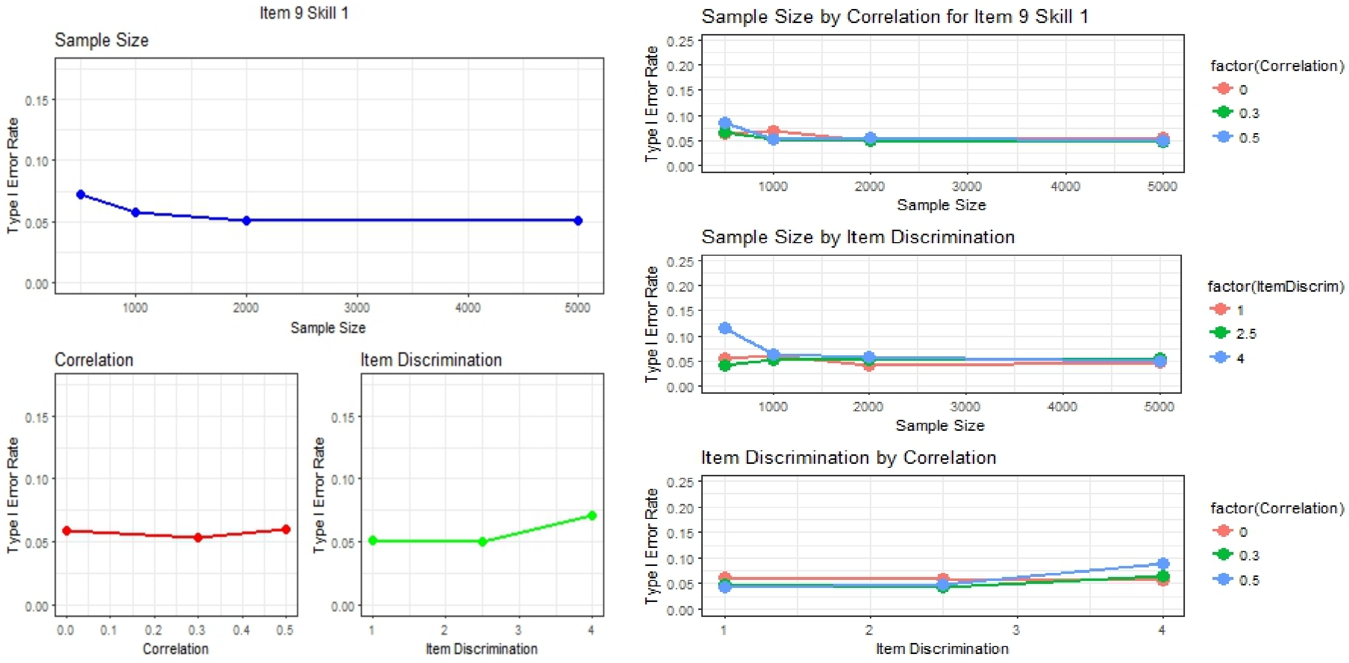


Figure 4: *One-way & Two-way Type I error Analysis by Sample Size, Correlation, Item Discrimination for Skill 1 of Item 9*

Sample Size

As expected the results in Figure 4 clearly show that as the sample size increases the Type I error significantly reduces and moves closer to the nominal value of $\alpha = .05$. In skill 1 of Item 9 the sample size of 500 had the largest Type I error rate of 0.07. However, as the sample size increases to 2,000 and 5,000 the Type I error rate attains the nominal value of $\alpha = .05$. It is important to note that the most marked reduction in the Type I error rate is seen at a sample size of 1,000. Going from a Sample size of 2,000 to 5,000 shows a minimal change in Type I error rates.

The two-way plots shown on the right side of Figure 4 provide a deeper understanding of the relationship between Type I error rate and the two-way interaction between Sample size, Correlation and Item discrimination. Type I error rates for the interaction of Sample size and

Correlation for skill 1 of Item 9 are lowest when Sample size is large ($\geq 2,000$) irrespective of Correlation. For medium (0.3) to large (0.5) degree of attribute Correlation the Type I error rate approaches the nominal ($\alpha = .05$) at Sample size of 1,000 or higher. However, for the group with no attribute Correlation the two smaller Sample sizes (500, 1,000) behave counter intuitively in that the Type I error goes up slightly from 0.6 to 0.7 when sample size increases from 500 to 1,000. This group (no attribute correlation) exhibits normally once Sample size is large $(\geq 2,000)$.

Item Discrimination

As stated earlier Item discrimination is a function of guess and slip where an Item discrimination of 1 corresponds to a guess and slip of approximately 0.4, an item discrimination of 2.5 is reflective of using guess and slip equal to 0.2 and lastly 4 represents items with very high level of item discrimination (guess and slip will be approximate 0.1). The results in Figure 4 show that as the Item(s) become more discriminating the Type I error significantly increases and moves further away from the nominal value of $\alpha = .05$. In Skill 1 of Item 9 the Item discrimination of 1 and 2.5 showed the lowest Type I error rates of 0.05. As the Item discrimination increases from 2.5 to 4 the Type I error rate increases to 0.07 (skill 1 of Item 9). It is important to note that the most marked increase in the Type I error rate is seen when Item discrimination goes from 2.5 to 4.

Type I error rate for the interaction of Sample size and Item discrimination for skill 1 of Item 9 are lowest when Item discrimination is 1 and Sample size is 2,000. Remarkably at the Item discrimination of 2.5 the Type I error rate approaches the nominal ($\alpha = .05$) across all levels of Sample size. While for an Item discrimination of 4 only as Sample size increases to

5,000 the Type I error rate approaches the nominal ($\alpha = .05$). For Item discrimination of 4 the Type I error rate shows the steepest drop when the Sample size goes from 500 to 1,000.

Correlation

Correlation between attribute(s) can take the following values: no correlation (0), moderate correlation (0.3) and high correlation. For example, if Skills 1 and 3 have high correlation (0.5) this indicates that students who possess skill 1 tend to possess skill 3 as well. The results in Figure 4 shows no clear relationship between Type I error rates and increasing attribute(s) correlation as shown by the flat nature of the graphed line. Specifically, in skill 1 of Item 9 having some correlation (0.3) between attribute(s) had the lowest Type I error rate of 0.05 but this trend does not continue as correlation between attributes increases. The two-way graphs comparing Type I error rates by Item discrimination and Correlation show that the group having no attribute correlation has little to no change across all Item discriminations. However, a high Item discrimination of 4 leads to higher Type I error rates as Correlation between attributes increases high. Specifically, when Item discrimination is 4 the group with no Correlation has a Type I error rate of 0.06, some Correlation (0.3) has a Type I error rate of 0.06 and high Correlation (0.5) has a Type I error rate of 0.09.

Multiplicity Correction for the Family-wise Type I error

In performing large series of statistical tests, some will have P values less than 0.05 purely by chance, even when all the null hypotheses are really true therefore a need for a multiplicity correction was investigated for the Family-wise Type I error by Sample Size, correlation and item discrimination. In order to investigate the severity of the Type I error inflation this study compared the Benjamini Hochberg and Bonferroni corrections for the Family-wise Type I error rates to the uncorrected using the p.adjust function in R. Both the

Benjamini Hochberg and Bonferroni corrections control the Type I error rate by adjusting the p-values for the number of hypothesis tests performed in slightly different ways. These adjusted p-values are compared across the 36 (4x3x3) conditions for 500 Samples with alpha at 0.05 level. The Family-wise Type I error rate represents the probability of making at least one or more false conclusion(s) in a series of hypothesis tests. The term “Family-wise” comes from a family of tests, which is really a series of tests on the data. In this study Family-wise represents all the q-entry elements in the Q-matrix (4 skills x 16 items = 64 q-vectors). Since the number of q-entry elements is large it is expected that the Type I error rates will be inflated due to chance and therefore a strong need exists to determine the degree of the Family-wise Type I error inflation.

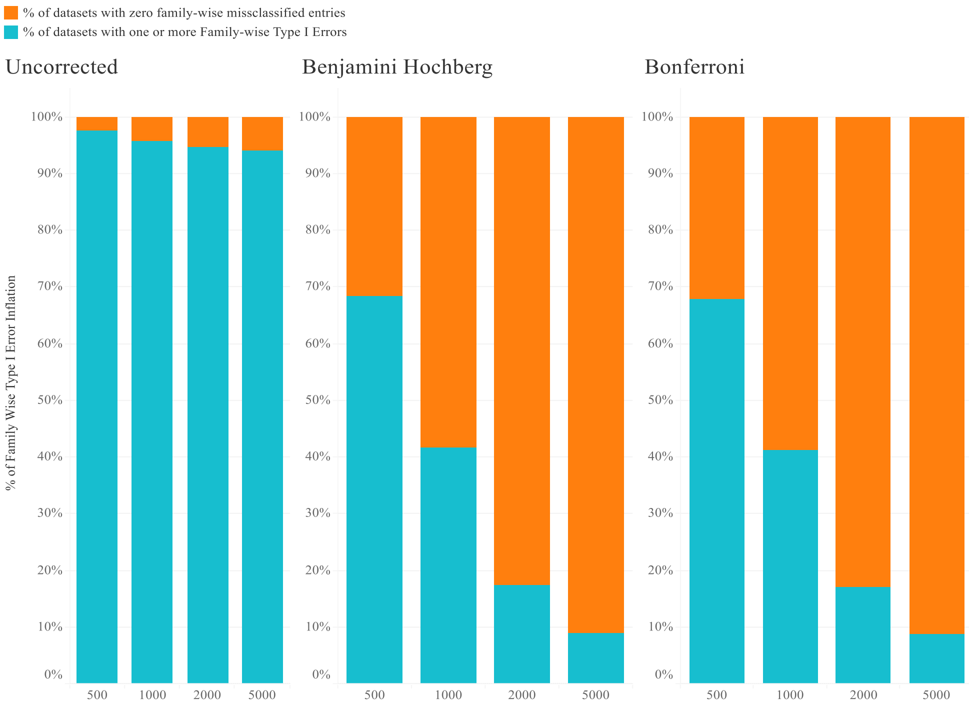


Figure 5: *Sample Size Family-wise Type I Error Analysis*

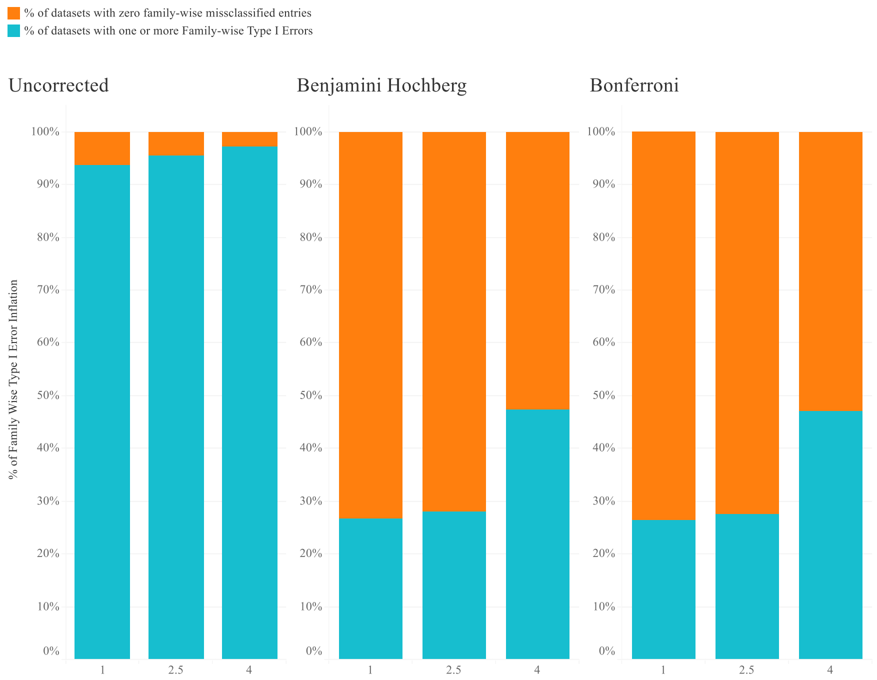


Figure 6: *Item Discrimination Family-wise Type I Error Analysis*

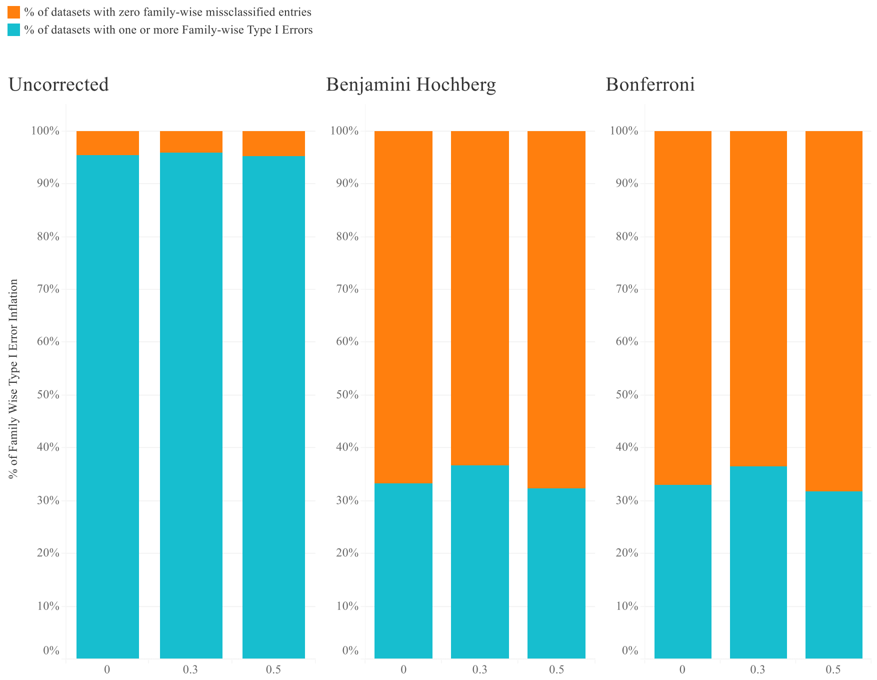


Figure 7: *Correlation Family-wise Type I Error Analysis*

In Figure 5, 6 and 7 the Family-wise Type I error is compared between the uncorrected, Benjamini Hochberg and Bonferroni adjustments with respect to Sample size, Item discrimination and Correlation. The first group represented graphically in orange is the % of the samples where zero q-entries were misclassified Family-wise implying that there was no Family-wise Type I error. The second group represented graphically in blue is the % of the samples where at least one or more of the Family-wise q-entries had a Type I error (Null was rejected incorrectly at least once for the whole Q-matrix). All three Figures clearly show the need for the corrections in controlling the Family-wise Type I error inflation. Clearly using the Benjamini Hochberg and Bonferroni corrections yields marked improvements in reducing Family-wise Type I error when compared to the uncorrected. Interestingly the Benjamini Hochberg and Bonferroni both yield similar improvements in reducing the Family-wise Type I error inflation across Sample size, Correlation and Item discrimination. Increasing Sample size clearly reduces the Family-wise Type I error for the Benjamin Hochberg and Bonferroni corrections considerably more than the uncorrected. Having low or medium Item discrimination (1 or 2.5) yields no difference in reducing the Family-wise Type I error however having high Item discrimination (4) clearly increases the Family-wise Type I error. Increasing attribute Correlation does not show any difference in reducing the Family-wise Type I error.

Another way to compare and contrast the Benjamini Hochberg, Bonferroni and uncorrected would be understand how they impact the number of samples that have zero Family-wise misclassifications across Sample Size, Correlation and Item discrimination. The percentage of samples which have all Family-wise q-entries correctly identified goes up significantly from the Uncorrected when the Benjamini Hochberg or Bonferroni adjustment is applied. Figures 8a, 8b and 8c show these relationships with respect to Sample Size, Correlation and Item discrimination. The percentage of samples that have zero Family-wise misclassifications goes up with increasing Sample Size. At a Sample size of 5,000 for both the Benjamini Hochberg and Bonferroni adjusted nearly 91% of the 500 samples have zero Family-wise misclassifications meaning only 9% exhibit some level of Type I error. No clear relationship is seen for item correlation. However, for Item discrimination we find that having low or medium Item discrimination (1 or 2.5) yields no difference in percentage of samples with zero Family-wise missclassifications however having high Item discrimination (4) clearly reduces the percentage of samples with zero Family-wise misclassifications from 72% to 53% and thereby increases the Family-wise Type I error from 28% to 47%.

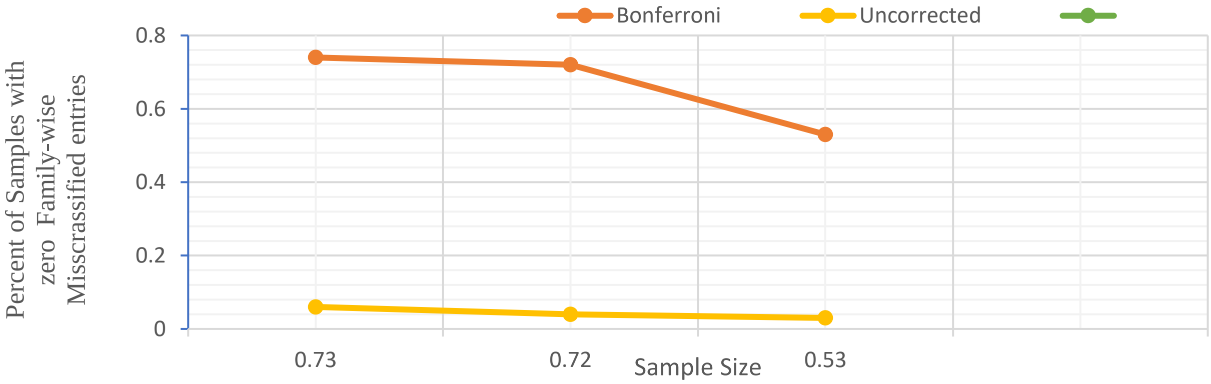


Figure 8a: *Percentage of Samples with Zero Family-wise Misclassifications by Sample Size*

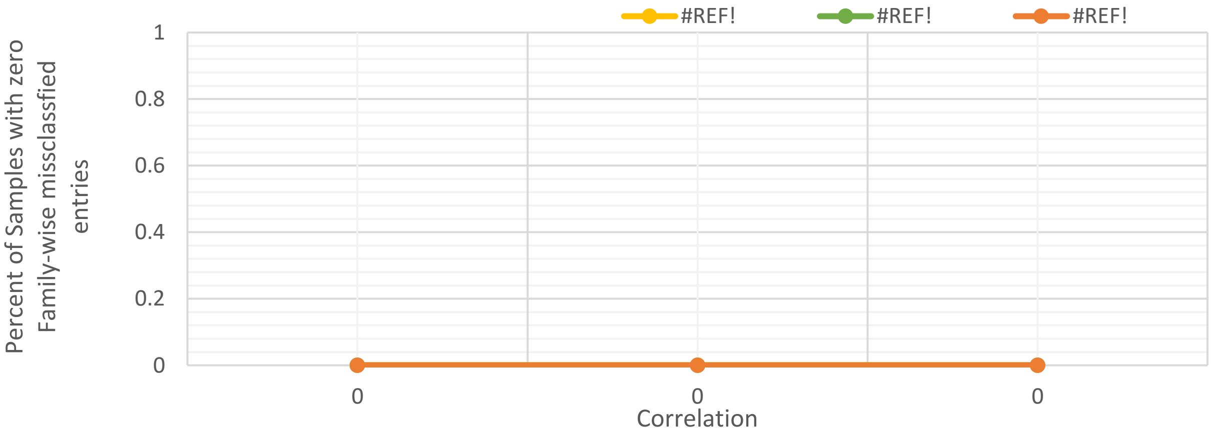


Figure 8b: *Percentage of Samples with Zero Family-wise Misclassifications by Correlation*

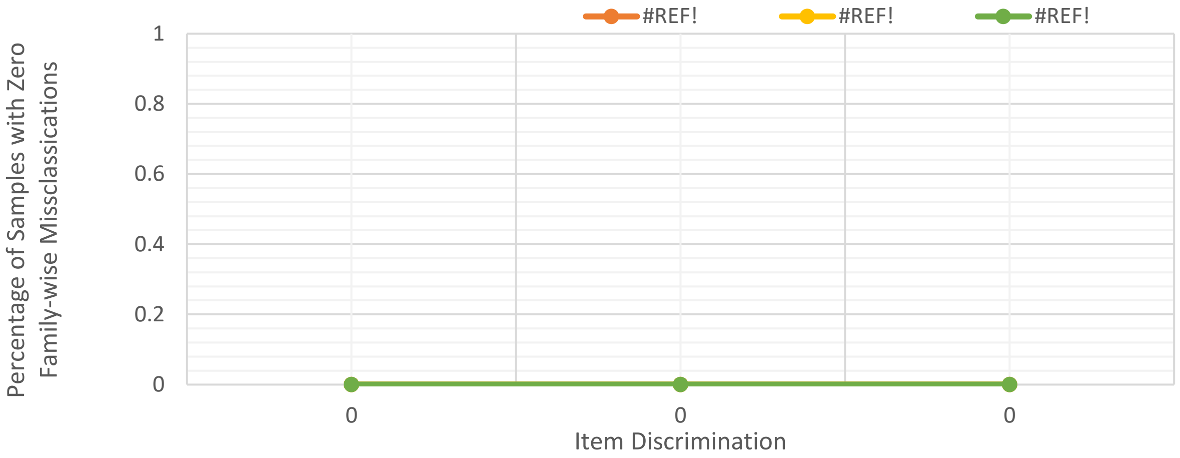


Figure 8c: *Percentage of Samples with Zero Family-wise Misclassifications by Item Discrimination*

In Figure 9, 10 and 11 the Family-wise Type I error is compared between the uncorrected, Benjamini Hochberg and Bonferroni adjustments for the two-way interaction between Sample size, Item discrimination and Correlation. As shown earlier these Figures also show the need for the corrections in controlling the Family-wise Type I error inflation. Using the Benjamini Hochberg (BH) and Bonferroni (BF) corrections yields marked improvements in reducing Family-wise Type I error over the uncorrected. In a manner similar to the one-way, two-way interactions of Sample size, Correlation and Item discrimination for the Benjamini Hochberg and Bonferroni adjustments show similar results in reducing the Family-wise Type I error inflation. Overall Family-wise Type I error rates $(>10\%)$ are larger for high Item Discrimination (4) even at large Sample sizes $(\geq 2,000)$. Having low or medium Item discrimination (1 or 2.5) in addition to large Sample sizes $(\geq 5,000)$ reduces the Family-wise Type I error $(<10\%)$ for both BH and BF. Large Sample sizes $(\geq 5,000)$ reduces the Family-wise Type I error rates $(<10\%)$ for both BH and BF across all Correlations. The interaction of

Correlation and Item discrimination doesn’t show any trend. Sample size seems to be the key driver in reducing Family-wise Type I error rates both BH and BF.

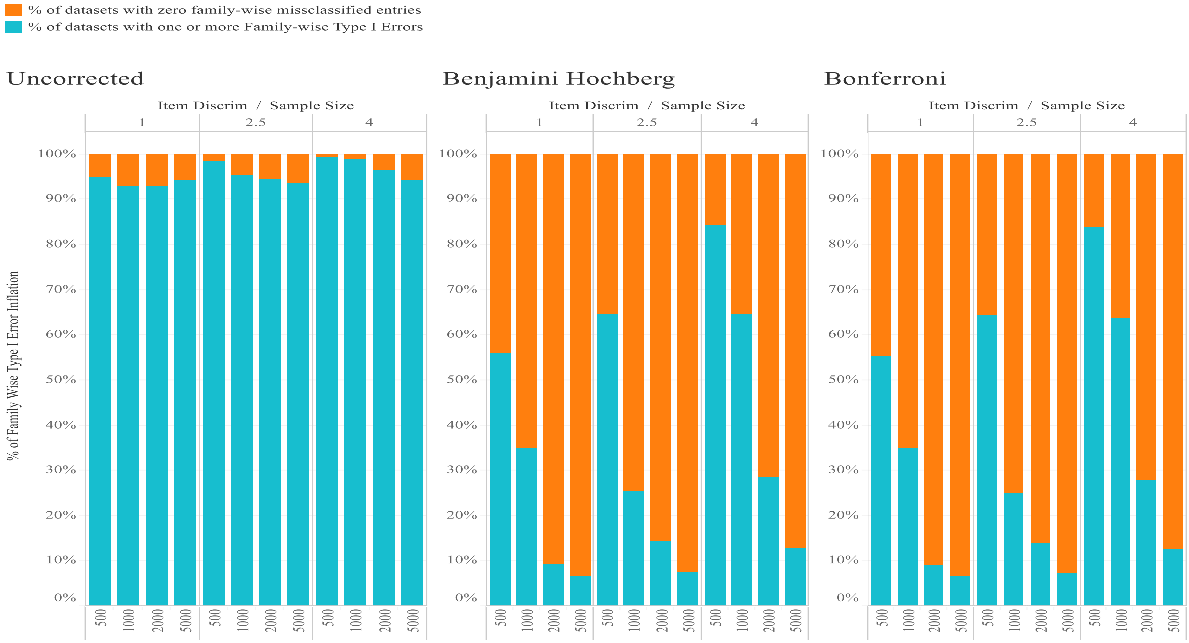


Figure 9: *Item Discrimination and Sample Size Family-wise Type I Error Analysis*

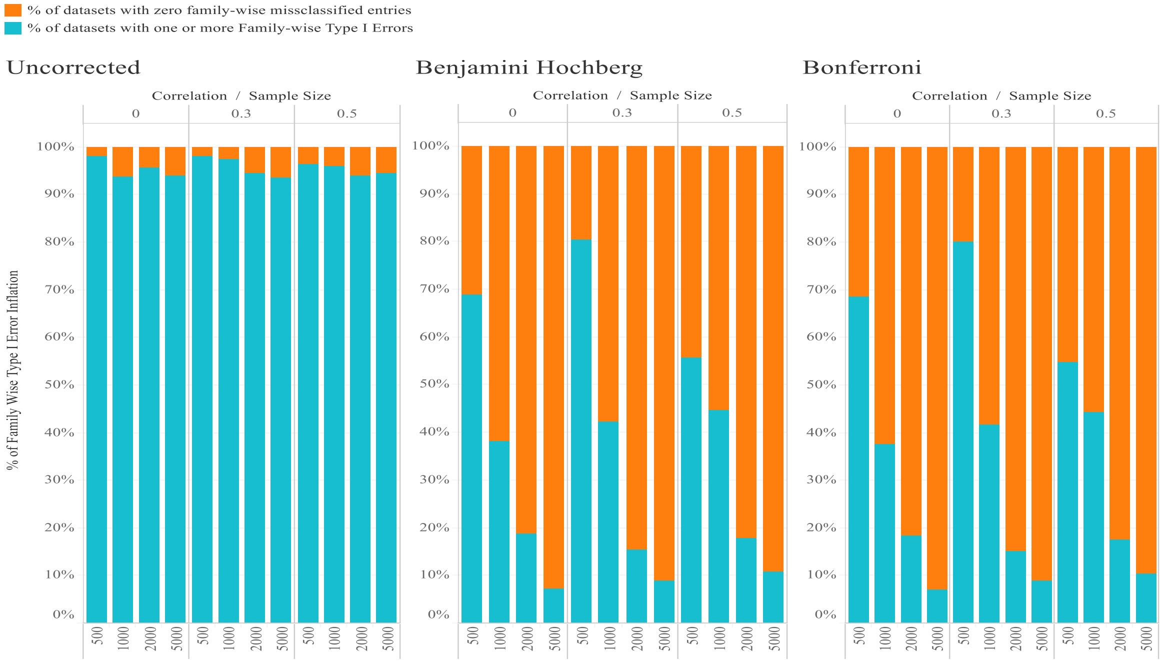


Figure 10: *Sample Size and Correlation Family-wise Type I Error Analysis*

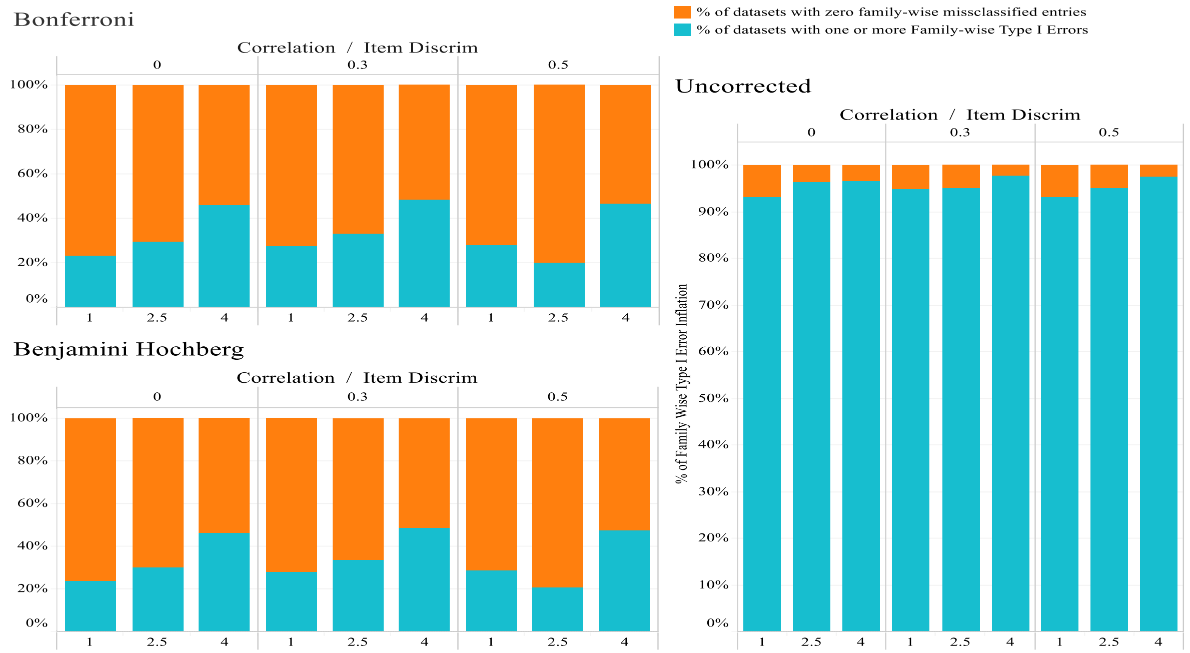


Figure 11: *Item Discrimination and Correlation Family-wise Type I Error Analysis*

Methodology for the Power Study

The second study is a Power analysis in which the performance of the test statistic is evaluated by assessing how often the proposed statistic correctly finds that the hypothesized model is underspecified or over specified, thus correctly identifying the addition of a necessary attribute or the deletion of an unnecessary attribute. The test is considered powerful when the test statistic correctly identifies when the null model is incorrectly specified across various Sample Sizes, Correlation and Item discriminations. The main purpose of this analysis was to determine, across a variety of sample size, attribute correlations, and Item discriminations, the ability of the statistic to correctly indicate when a Q-matrix modification is in fact needed. This simulation study involved generating 500 samples for each of the 36 conditions which as shown in Table 10 represent the three factors investigated; Correlation, Sample Size and Item discrimination (3x4x3). For each of the 36 conditions 500 samples were generated by introducing q-matrix misspecification(s) which involved 'switching' an entry to the wrong value during the estimation stage and later measuring how often the procedure detected these misspecifications. Detection here is defined as how often a false null hypothesis is correctly rejected. As shown in the Table below four cases of misspecifications were investigated in this study. The first was item 9 in which skill 1 (α_1) was switched from 1 to 0 and therefore represents a underfitting misspecification. The second involved item 10 in which skill 1(α_1) is switched from 0 to 1 and therefore represents an overfitting misspecification. These first two denote the identification of a single miss-specified q-vector whereas case 3 & 4 involve the identification of two miss-specified Q-vectors. Specifically, case 3 involved skill 1 (α_1) of item 9 to be switched from 1 to 0 and skill 1 (α_1) of item 10 to be switched from 0 to 1, so representing under and over fitting misspecifications across two items for the same skill. In case 4 skill 1 (α_1) of item 9 was switched from 1 to 0 and skill (α_3) of item 9 was switched from 0 to 1, thereby representing under and over fitting misspecifications together in a single item. These four cases represent the main types of misspecification's that would be of interest in CDM's.

Table 12: *Summary of Q-vector Misspecifications*

Case	Type	Item Altered	Q-vector before Altered Skill				Q-vector after Altered Skill				Number of Alterations
			α_1	α_2	α_3	α_4	α_1	α_2	α_3	α_4	
1	US	9	1	1	0	0	0	1	0	0	1
2	OS	10	0	1	1	0	1	1	1	0	1
3	OUS	9	1	1	0	0	0	1	0	0	2
		10	0	1	1	0	1	1	1	0	
4	OUS	9	1	1	0	0	0	1	1	0	2

Note. OS= Over fitting q-vector; US= Under fitting q-vector; OUS= Over- and under fitting q-vector.

Results for the Power Analysis

Case 1 (Item 9 Skill 1): Q matrix underspecified by a single Q-entry

The simulation results indicate that the statistic is very powerful in the detection of an under-specified Q-matrix and has ample power in large sample sizes $(\geq 2,000)$ or when item 9 is highly discriminating (≥ 2.5) between students that have mastered or have not mastered skill 1 (α_1). Specifically, the statistic achieves perfect power for item 9 when it was highly discriminating (4) between students that have mastered or have not mastered skill $1(\alpha_1)$. However, when item 9 is not very discriminating $(=1)$ between masters and non-masters of the measured skill 1 (α_1), this statistic was not as powerful (Power = 0.57) in the detection of model under-specification. It is important to note that the improvement in power when an item is highly discriminating (4) comes at a cost of higher Type I error. For the two smaller sample sizes (500 and 1,000), the power of the test never exceeded 0.75. In the condition with Sample size of 2,000 examinees, power was above 0.90. As expected, the statistic was most powerful (Power = 0.95) in the largest sample size (5,000) condition.

When skill 1 (α_1) had no correlation with other skills for Item 9 the statistic was most powerful (Power = 0.92). Yet if any correlation (≥ 0.3) between skill 1 (α_1) and other skills exists the power drops to 0.80. A possible reason for this finding is that high correlations between two or more skills creates redundant information and thus a smaller amount of information is known about each individual skill. In the extreme case of perfect correlation between skill $1(\alpha_1)$ and another skill, skill $1(\alpha_1)$ can be used to predict the other but then it becomes difficult to figure out which skill among the correlated skills is actually required for Item 9. This problem is commonly found in Regression analysis and is named Multicollinearity.

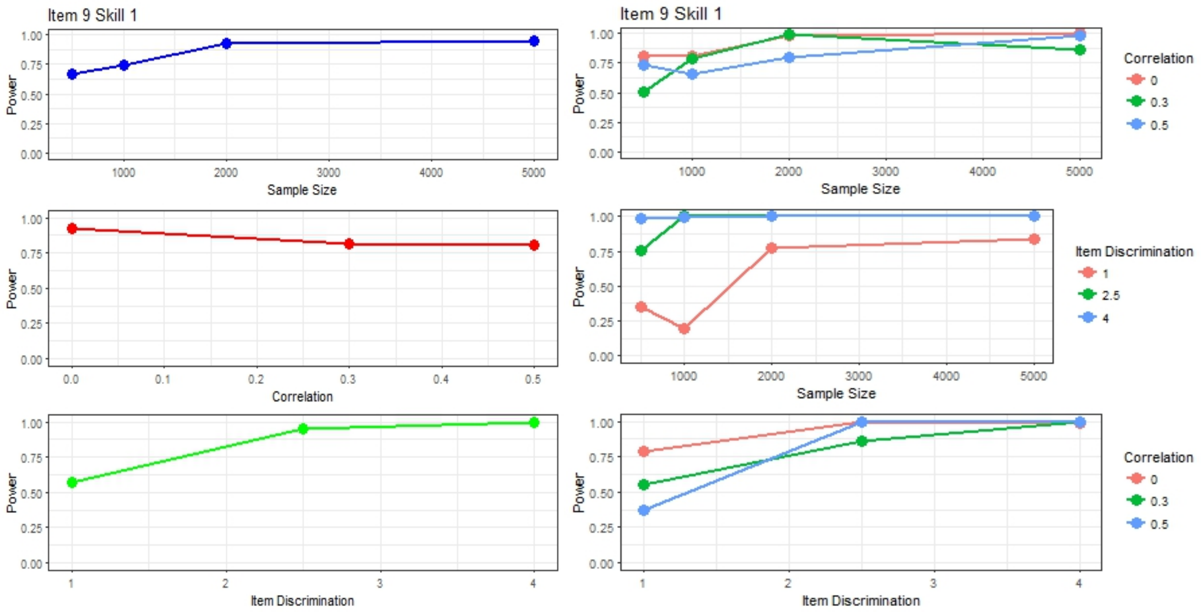


Figure 12: *Power Analysis for Case 1(Item 9 Skill 1) by Factor*

The relationship between sample size and having higher correlation (≥ 0.3) between skill $1(\alpha_1)$ and other skills is unclear with respect to its impact on power for Item 9. Remarkably the statistic achieves perfect power for item 9 at very small sample sizes when it was highly discriminating (≥ 2.5) between students that have mastered or have not mastered skill $1(\alpha_1)$. When item 9 is not very discriminating $(=1)$ between masters and non-masters of the measured skill $1(\alpha_1)$ a minimum sample size of 2,000 only yields a Power = 0.77 in the detection of model under-specification. Lastly the statistic achieves perfect power for item 9 when it was highly discriminating $(=4)$ between students that have mastered or have not mastered skill $1(\alpha_1)$ irrespective of whether skill $1(\alpha_1)$ is correlated to other skills. This nearly perfect power comes at a cost, an increased Type I error rate.

Case 2 (Item 10 Skill 1): Q matrix over specified by a single Q-entry

The simulation results indicate that the statistic is very powerful in the detection of an over-specified Q-matrix and has sufficient power in large sample sizes $(\geq 5,000)$ or when item

10 is highly discriminating (≥ 2.5) between students that have mastered or have not mastered skill $1(\alpha_1)$. Moreover, the statistic achieves nearly perfect power for item 10 when it was highly discriminating (≥ 2.5) between students that have mastered or have not mastered skill $1(\alpha_1)$. This nearly perfect power seen for highly discriminating comes with an increased Type I error rate. Still, when item 10 is not very discriminating $(=1)$ between masters and non-masters of the measured skill 1 (α_1), this statistic was not as powerful (Power = 0.48) in the detection of model over-specification. For the three smaller sample sizes (500, 1,000 and 2,000), the power of the test never exceeded 0.78. In the condition with Sample size of 5,000 examinees, power is 0.96.

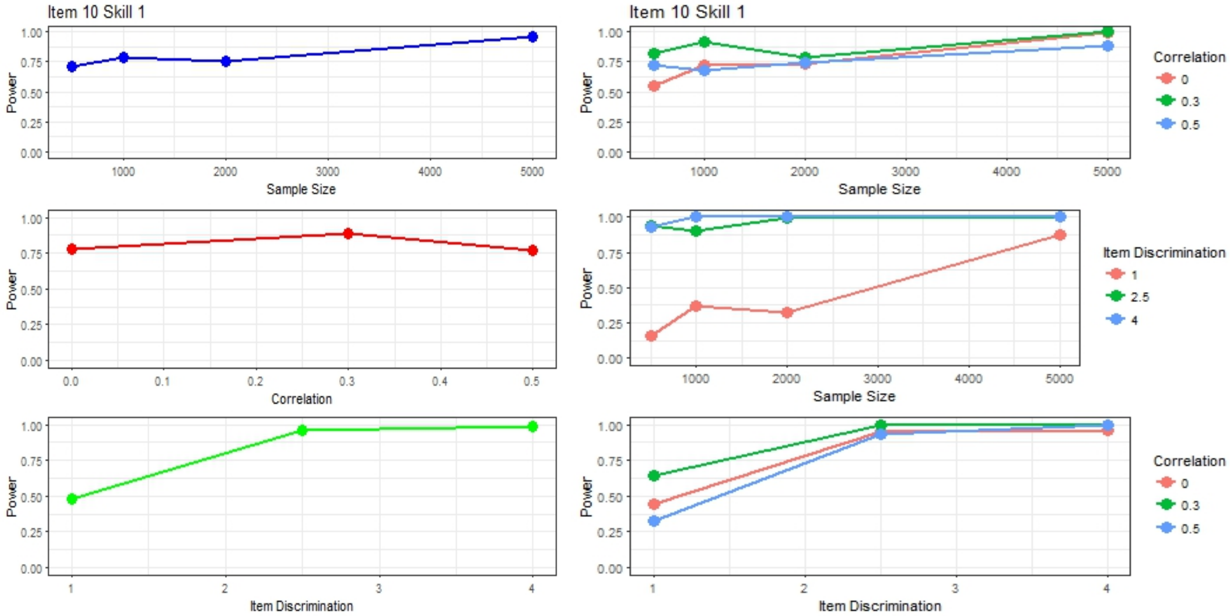


Figure 13: *Power Analysis for Case 2(Item 10 Skill 1) by Factor*

Having higher correlation (≥ 0.3) between skill $1(\alpha_1)$ and other skills is unclear with respect to its impact on power for Item 10. The statistic achieves near perfect power for item 10 at very small sample sizes when it was highly discriminating (≥ 2.5) between students that have mastered or have not mastered skill $1(\alpha_1)$. When item 10 is not very discriminating $(=1)$

between masters and non-masters of the measured skill $1(\alpha_1)$ a minimum sample size of 5,000 only yields a Power = 0.88 in the detection of model over-specification. Lastly, the statistic achieves perfect power for item 10 when it was highly discriminating (≥ 2.5) between students that have mastered or have not mastered skill $1(\alpha_1)$ irrespective of whether skill $1(\alpha_1)$ is correlated to other skills.

Case 3 (Item 10 & Item 9 Skill 1): Q matrix with two misspecifications across two items

Case 3 analyzes the under and over-specification caused by adding two simultaneous misspecifications across two items but for the same skill. Specifically, skill 1 (α_1) of item 9 was switched from 1 to 0 and skill 1 (α_1) of item 10 to be switched from 0 to 1. The proportion of samples with statistically significant rejections of the two Null(s) for each factor of the analysis is listed in Table 13. The first column is the proportion of samples where both Null(s) for Case 3 were correctly rejected using this statistic. The second column is the proportion of samples where the Null for Skill 1 of Item 9 is correctly rejected, basically this group is a sum of column 1 (samples where both Null(s) were correctly rejected) and column 4 (only Skill 1 of Item 9 Null was correctly rejected). The third column is the proportion of samples where the Null for Skill 1 of Item 10 is correctly rejected, basically this group is a sum of column 1 (samples where both Null(s) were correctly rejected) and column 5 (only Skill 1 of Item 10 Null was correctly rejected). Column 4 and 5 represent the situation where only one of the two Null(s) is correctly rejected by the statistic and therefore means that one incorrect Null was accepted leading to either over-specification (Skill 1 of Item 9) or under-specification (Skill 1 of Item 10) error. Column 6 is the proportion of samples where both Null(s) were incorrectly accepted or failed to reject an incorrect Null using this statistic.

Table 13: *Proportion Correctly Rejected Null(s) by Factor for Case 3 (Item 10 & Item 9 Skill1)*

Factor	Both Q-entries Correctly Rejected	Q-entry correctly rejected (Item 9 Skill 1)	Q-entry correctly rejected (Item 10 Skill 1)	Only one Q-entry correctly rejected (Item 9 Skill 1)	Only one Q-entry correctly rejected (Item 10 Skill 1)	Both Q-entries incorrectly accepted
Sample Size						
500	46.5%	67.6%	53.5%	21.0%	6.9%	25.4%
1,000	52.2%	65.5%	61.5%	13.0%	9.3%	25.1%
2,000	71.9%	85.8%	80.3%	13.3%	8.4%	5.7%
5,000	86.8%	93.6%	87.2%	6.8%	0.4%	6.0%
Correlation						
0	64.5%	77.9%	71.0%	12.8%	6.4%	15.6%
0.3	81.7%	88.6%	87.3%	6.8%	5.6%	5.8%
0.5	46.8%	67.9%	53.6%	21.0%	6.8%	25.3%
Item Discrimination						
1	35.1%	49.3%	46.7%	14.0%	11.6%	39.1%
2.5	72.8%	87.8%	79.1%	15.0%	6.2%	6.0%
4	85.1%	97.3%	86.2%	11.5%	1.0%	1.7%

As sample size increases the proportion of samples correctly rejecting both Nulls (Column 1) increases from 46.5% to 86.8%, the proportion of samples correctly rejecting at least Skill 1 of Item 9 (Column 2) increases from 67.6% to 93.6% and the proportion of samples correctly rejecting at least Skill 1 of Item 10 (Column 3) increases from 53.5% to 87.2%. Instead the proportion of samples where only one of the two Nulls was correctly rejected (Column 4 & 5) decreased with increasing sample size primarily due to the increase in the number of samples where both Nulls are correctly rejected (Column 1) since Column 2 is really a sum of Column 1 and 4 and Column 3 is a sum of Column 1 and 5. Fewer of the two q-entries (Skill 1 of Item 9 & Item 10) are erroneously accepted with larger sample sizes as seen by the proportion of samples where both Nulls are incorrectly accepted (Column 6) decreasing dramatically from 25.4% at a sample size of 500 to 6.0% for a sample size of 5,000.

As both item 9 and 10 become more discriminating (≥ 2.5) between students that have mastered or have not mastered skill $1(\alpha_1)$ the proportion of samples correctly rejecting both Nulls (Column 1) goes from 35.1% to 85.1%, the proportion of samples where at least the Null for skill 1 of Item 9 is correctly rejected goes from 49.3% (at low discrimination) to 97.3% (at high discrimination) and the proportion of samples where at least the Null for skill 1 of Item 10 is correctly rejected goes from 46.7% (at low discrimination) to 86.2% (at high discrimination). Its noteworthy that detection of under-specification (skill 1 of Item 9) is better than for over-specification (skill 1 of Item 10) when both are being tested simultaneously. Fewer of the two q-entries (Skill 1 of Item 9 & Item 10) are erroneously accepted when both item 9 and 10 become more discriminating (≥ 2.5) as seen by the proportion of samples where both Nulls are incorrectly accepted (Column 6) decreasing dramatically from 39.1% at an Item discrimination of 1 to only 1.7% for an Item discrimination of 4. Important to remember that the increased power seen for high item discrimination is tempered with higher Type I error rates. Finally, the relationship between the proportion of samples with correctly accepted Nulls is unclear for having higher correlation (≥ 0.3) between skill $1\left(\alpha_1\right)$ and other skills.

Case 4 (Item 9 Skill 1 & 3): Q matrix with two misspecifications within the same item

Case 4 introduced two simultaneous misspecifications of the Q-matrix for item 9 across skill 1 (α_1) and skill 3 (α_3). For item 9 skill 1 (α_1) was switched from 1 to 0 and skill 3 (α_3) of item 9 was switched from 0 to 1, thereby causing under and over specification misspecification in a single item but across two skills. The proportion of samples with statistically significant rejections of the two Null(s) for each factor of the analysis is listed in Table 14. The first column is the proportion of samples where both Null(s) for Case 4 were correctly rejected using this statistic. The second column is the proportion of samples where the Null for Skill 1 of Item 9 is correctly rejected, basically this group is a sum of column 1 (samples where both Null(s) were correctly rejected) and column 4 (only Skill 1 of Item 9 Null was correctly rejected). The third column is the proportion of samples where the Null for Skill 3 of Item 9 is correctly rejected, basically this group is a sum of column 1 (samples where both Null(s) were correctly rejected) and column 5 (only Skill 3 of Item 9 Null was correctly rejected). Column 4 and 5 represent the situation where only one of the two Null(s) is correctly rejected by the statistic and therefore means that one incorrect Null was accepted leading to either over-specification (Skill 1 of Item 9) or under-specification (Skill 3 of Item 9) error. Column 6 is the proportion of samples where both Null(s) were incorrectly accepted or failed to reject an incorrect Null using this statistic.

Table 14: *Proportion Correctly Rejected Nulls by Factor for Case 4 (Item 9 Skill 1 & Skill 3)*

Factor	**Both Q-entries Correctly Rejected**	**Q-entry correctly rejected (Item 9 Skill 1)**	**Q-entry correctly rejected (Item 9 Skill 3)**	**Only one Q-entry correctly rejected (Item 9 Skill 1)**	**Only one Q-entry correctly rejected (Item 9 Skill 3)**	**Both Q-entries incorrectly accepted**
Sample Size						
500	43.3%	69.9%	48.6%	26.3%	5.0%	24.6%
1,000	52.4%	74.8%	57.8%	21.6%	5.4%	19.8%
2,000	65.8%	89.2%	70.0%	23.4%	4.2%	6.6%
5,000	82.4%	91.7%	85.1%	9.2%	2.7%	5.6%
Correlation						
0	57.4%	84.4%	60.8%	26.3%	3.3%	12.2%
0.3	66.6%	77.2%	74.5%	10.3%	7.5%	14.9%
0.5	59.0%	82.8%	61.1%	23.8%	2.1%	15.2%
Item Discrimination						
1	31.3%	53.8%	40.8%	22.0%	9.3%	36.4%
2.5	53.8%	90.8%	56.7%	37.0%	2.9%	6.3%
4	97.8%	99.2%	98.6%	1.4%	0.8%	0.0%

As sample size increases the proportion of samples correctly rejecting both Nulls (Column 1) increases from 43.3% to 82.4%, the proportion of samples correctly rejecting at least Skill 1 of Item 9 (Column 2) increases from 69.9% to 91.7% and the proportion of samples correctly rejecting at least Skill 3 of Item 9 (Column 3) increases from 48.6% to 85.1%. The proportion of samples where only one of the two Nulls was correctly rejected (Column 4 & 5) decreased with increasing sample size primarily due to the increase in the number of samples where both Nulls are correctly rejected (Column 1) since Column 2 is really a sum of Column 1 and 4 and Column 3 is a sum of Column 1 and 5. Fewer of the two q-entries (Skill 1 of Item 9 & Skill 3 of Item 9) are erroneously accepted with larger sample sizes as seen by the proportion of samples where both Nulls are incorrectly accepted (Column 6) decreasing dramatically from 25.6% at a sample size of 500 to 5.6% for a sample size of 5,000.

As item 9 becomes more discriminating (≥ 2.5) the proportion of samples correctly rejecting both Nulls (Column 1) goes from 31.3% to 97.8%, the proportion of samples where at least the Null for skill 1 of Item 9 is correctly rejected goes from 53.8% (at low discrimination) to 99.2% (at high discrimination) and the proportion of samples where at least the Null for skill 3 of Item 9 is correctly rejected goes from 40.8% (at low discrimination) to 98.6% (at high discrimination). This tremendous increase in power seen for high discrimination is also associated with increased Type I error rates. The detection of under-specification (skill 1 of Item 9) is similar for over-specification (skill 3 of Item 9) when both are being tested simultaneously. Fewer of the two q-entries (Skill 1 and Skill 3 of Item 9) are erroneously accepted when item 9 becomes more discriminating (≥ 2.5) as seen by the proportion of samples where both Nulls are incorrectly accepted (Column 6) decreasing dramatically from 36.4% at an Item discrimination of 1 to only 0% for an Item discrimination of 4. Finally, the relationship between the proportion of samples with correctly accepted Nulls is unclear for having higher correlation (≥ 0.3) between skills.

Estimator for the Asymptotic Variance V$_d$

Since the d_{jk} statistic can be viewed as a sum of random variables it is believed to be asymptotically normal with a mean of 0 and some variance V_d under H$_0$ because of the central limit theorem. The estimator of the asymptotic variance V_d was derived by (Johnson, 2015). Johnson (2015) accounts for the statistic being evaluated at the MLE and therefore approximated it with the equation below.

$$d_{jk}(\hat{\psi}) \approx d_{jk}(\hat{\psi}) + (d')^T (\hat{\psi} - \psi)$$

$$d_{jk}(\hat{\psi}) \approx d_{jk}(\hat{\psi}) + (d')^T \mathrm{V}_{\hat{\psi}} \Delta l$$

$$\textit{Variance} \text{ is } V_d \approx V_d - C^T \mathrm{V}_{\hat{\psi}}^{-1}$$

Where $\hat{\psi}$ denotes the maximum likelihood estimator of the entire parameter vector ψ. Johnson (2015) derives that $d_{jk}(\hat{\psi})$ is asymptotic normal with mean 0 and variance $V_{\hat{d}} \approx V_d - C^T V_{\hat{\psi}}^{-1}$ and C is the covariance between $d(\psi)$ and ∇l.

Normal Approximation Tested using the Shapiro Wilk's Test of Normality

The normal approximation of $d(\psi)$ was tested using the Shapiro Wilk's test of normality and results are shown in Table 15. The results for the Shapiro and Wilk (1965) tests of normality are summarized for the all the q-vectors (4 skills x 16 items = 64) across all the conditions (4 Sample sizes x 3 Correlations x 3 Item discriminations=36 conditions). This test for normality calculates the probability that the sample was drawn from a normal population.

The hypothesis tested is:

H_0: The sample data are not significantly different than a normal population.

H_A: The sample data are significantly different than a normal population.

The Null hypothesis was rejected with P-values $\alpha < .05$. (Field, 2009) points out that this test is biased by sample size in such a way that for large samples, the P-value could be low even though the deviations from normality are negligible. It is therefore important to supplement the test results with a visual tool called Q-Q plots in order to fully conclude whether data is normal (Field, 2009). Individual Q-Q plots are provided for each q-vector by condition in the appendix. Briefly a normal Q-Q plot has sample quantiles on the vertical axis and theoretical (normal) quantiles on the horizontal axis. If the data is drawn from a Gaussian distribution, all data points should to be located on a straight line.

Table 15: *Shapiro Wilk Test of Normality by Factor*

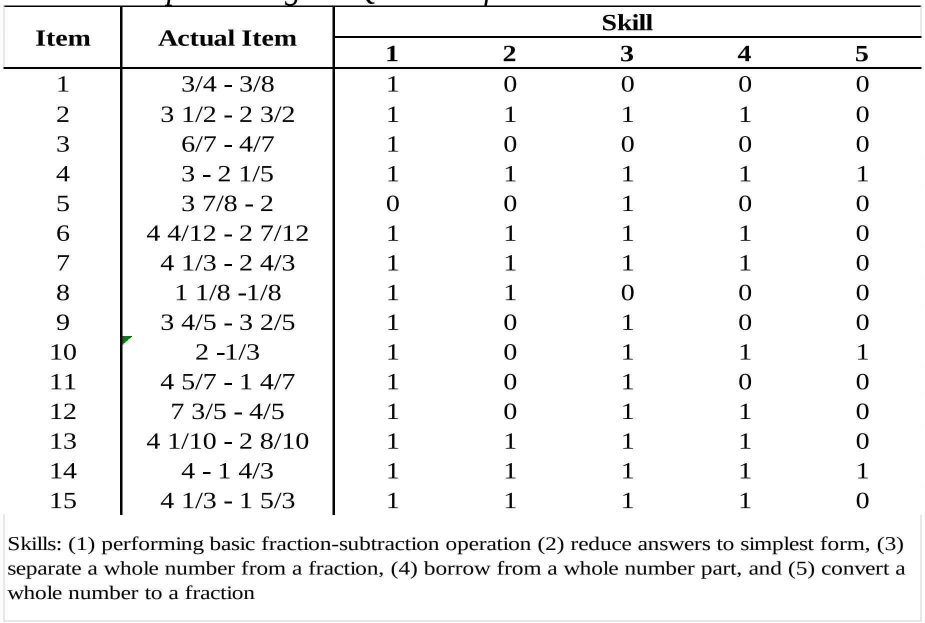
	Shapiro Wilk Test of Normality	
Factor	Number Rejected	% of Total
Item Discrimination		
1	295	38.4%
2.5	328	42.7%
4	355	46.2%
Correlation		
0	282	36.7%
0.3	332	43.2%
0.5	354	46.7%
Sample Size		
500	350	60.8%
1000	275	47.7%
2000	193	33.5%
5000	160	27.8%

The results from the Shapiro-Wilks test of Normality show that as sample size increases the number of tests rejected because enough evidence exists to reject the Null that the data comes from a Gaussian or Normal distribution reduces from 60.8% at Sample size of 500 to 27.8% at a Sample size of 5,000. This result is very much in line with the central limit theorem (CLT) states that the sum of identically and independently distributed (i.i.d.) random variables with well-defined expected value and variance will approach normality when the number of observations gets sufficiently large, regardless of the underlying distribution. The inspection of the Q-Q plots for each q-vector for each of the conditions also reinforces the results found in the Shapiro Wilks test of normality. The Q-Q plots found in the Appendix showed that deviations from normality were mainly found for smaller sample sizes (500) and high item discrimination (=4).

Fraction Subtraction Methodology

In order to provide practical value to the simulation results, a real data example is needed to see how the methods hold up to real response data. The estimated Q-matrix is compared with the expert designed Q-matrix and to the results obtained by DeCarlo (2012) in order to evaluate the current method in its capacity to identify when Q-matrix elements are poorly designed by experts. Tatsuoka (1990) fraction subtraction dataset is widely analyzed and has been used for the analysis of Q-matrix validation by de la Torre (2008) and DeCarlo (2012). This study uses the version which consists of 15 items and 536 middle school students. The Q-matrix used here is the same as the one used by de la Torre (2008) who adapted it from Mislevy (1996) and is shown in Table 16. There are five required attributes, including: (1) performing basic fraction-subtraction operation (2) reduce answers to simplest form, (3) separate a whole number from a fraction, (4) borrow from a whole number part, and (5) convert a whole number to a fraction. Despite the small sample size of this dataset this study will use the 15 items version for this empirical section.

Table 16: *Expert Designed Q-Matrix for Fraction Substraction Dataset*



Item	**Actual Item**	**Skill**				
		1	**2**	**3**	**4**	**5**
1	3/4 - 3/8	1	0	0	0	0
2	3 1/2 - 2 3/2	1	1	1	1	0
3	6/7 - 4/7	1	0	0	0	0
4	3 - 2 1/5	1	1	1	1	1
5	3 7/8 - 2	0	0	1	0	0
6	4 4/12 - 2 7/12	1	1	1	1	0
7	4 1/3 - 2 4/3	1	1	1	1	0
8	1 1/8 -1/8	1	1	0	0	0
9	3 4/5 - 3 2/5	1	0	1	0	0
10	2 -1/3	1	0	1	1	1
11	4 5/7 - 1 4/7	1	0	1	0	0
12	7 3/5 - 4/5	1	0	1	1	0
13	4 1/10 - 2 8/10	1	1	1	1	0
14	4 - 1 4/3	1	1	1	1	1
15	4 1/3 - 1 5/3	1	1	1	1	0

Skills: (1) performing basic fraction-subtraction operation (2) reduce answers to simplest form, (3) separate a whole number from a fraction, (4) borrow from a whole number part, and (5) convert a whole number to a fraction

Fraction Subtraction Results

The analysis of the fraction subtraction dataset using the test statistic found that 11 out of the 75 (15%) q-entries had enough evidence to reject the Null hypothesis that the original expert designed Q-entry is correct. This clearly indicates that the test finds misfit in the original expert designed Q-matrix, a finding which has been found by various other researchers (DeCarlo, 2012;

Chung, 2014; Chen-miao, 2013). The 11 q-entries that this test rejects are $q_{jk} = \left(q_{11}, q_{33}, q_{34}, q_{41}, q_{61}, q_{64}, q_{10,1}, q_{11,4}, q_{14,1}, q_{14,2}, q_{14,3}\right)$, where j=item and k=skill. The results as shown in Table 17 display that the rejected 11 q-entries are from 7 items (item 1,3,4,6,10,11 and 14) whereas the remaining 8 items {2, 5, 7, 8, 9, 12, 13 and 15} remain unchanged from the original expert designed Q-matrix.

Table 17: *A comparison of Expert designed vs. current hypothesis test(s)*

		Expert Designed						**Current Method**				
Item	**Actual Item**	**Skill**					**Item**	**Skill**				
		1	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**
1	3/4 - 3/8	1	0	0	0	0	1	1	0	0	0	0
2	3 1/2 - 2 3/2	1	1	1	1	0	2	1	1	1	1	0
3	6/7 - 4/7	1	0	0	0	0	3	1	0	**1**	**1**	0
4	3 - 2 1/5	1	1	1	1	1	4	**0**	1	1	1	1
5	3 7/8 - 2	0	0	1	0	0	5	0	0	1	0	0
6	4 4/12 - 2 7/12	1	1	1	1	0	6	**0**	1	1	**0**	0
7	4 1/3 - 2 4/3	1	1	1	1	0	7	1	1	1	1	0
8	1 1/8 -1/8	1	1	0	0	0	8	1	1	0	0	0
9	3 4/5 - 3 2/5	1	0	1	0	0	9	1	0	1	0	0
10	2 -1/3	1	0	1	1	1	10	**0**	0	1	1	1
11	4 5/7 - 1 4/7	1	0	1	0	0	11	1	0	1	**1**	0
12	7 3/5 - 4/5	1	0	1	1	0	12	1	0	1	1	0
13	4 1/10 - 2 8/10	1	1	1	1	0	13	1	1	1	1	0
14	4 - 1 4/3	1	1	1	1	1	14	**0**	**0**	**0**	1	1
15	4 1/3 - 1 5/3	1	1	1	1	0	15	1	1	1	1	0

Skills: (1) performing basic fraction-subtraction operation (2) reduce answers to simplest form, (3) separate a whole number from a fraction, (4) borrow from a whole number part, and (5) convert a whole number to a fraction

To begin with Item 1 does not require any skill in the Q-matrix suggested by the current hypothesis test. The test shows that skill 1 (performing basic fraction-subtraction operation) seems to be especially problematic because in the original expert designed Q-matrix all items except item 5 require skill 1 to be mastered but this test clearly finds that it is not necessary to master for items 1, 4, 6, 10 and 14. Results for items 1, 3, 6 and 11 are a bit unclear and may indicate that more than one q-entry is misfit. In the current method each test really involves assuming that the rest of the Q-matrix is correct. These unclear results suggest that the assumption that the rest of the Q-matrix is correct may be false. A possible approach to try would be a sequential test.

It’s important to keep in mind that this dataset has some inherent differences that may make the application of this hypothesis test problematic. To begin with its relatively small sample size is a concern, because some attribute patterns might be too sparse to estimate accurately. Most importantly the original Q-matrix designed by experts is not identified or complete. Lui, Xu & Ying (2012) define a Q-matrix is complete if and only if for each skill there exists an item only requiring that skill. They also show that completeness is among the sufficient conditions to identify Q. The simulations for the hypothesis test were based on a complete and identified Q-matrix.

Table 18: *Comparison of current method with DeCarlo's Bayesian Approach*

Item	**Actual item**	**Expert Designed Skills**						**Current Method Skills**						**DeCarlo's Bayesian Approach Skills**				
		1	**2**	**3**	**4**	**5**	**Item**	**1**	**2**	**3**	**4**	**5**	**Item**	**1**	**2**	**3**	**4**	**5**
4	3 - 2 1/5	1	1	1	1	1	4	0	1	1	1	1	4	1	0	1	1	1
5	3 7/8 - 2	0	0	1	0	0	5	0	0	1	0	0	5	0	1	1	0	1
10	2 -1/3	1	0	1	1	1	10	0	0	1	1	1	10	1	0	1	1	1
14	4 - 1 4/3	1	1	1	1	1	14	0	0	0	1	1	14	1	1	1	1	1

Skills: (1) performing basic fraction-subtraction operation (2) reduce answers to simplest form, (3) separate a whole number from a fraction, (4) borrow from a whole number part, and (5) convert a whole number to a fraction

Next, the results from the current test are compared to DeCarlo (2012) who used the Bayesian Framework to test misfit for the same fraction subtraction dataset. DeCarlo (2012) investigates four items (items 4, 5, 10 and 14) that involve whole numbers (without fractions) with respect to three skills (skill 2, 4 and 5). A major difference exists between the current method and DeCarlo’s approach in that DeCarlo estimates these 12 uncertain elements simultaneously using the Bayesian framework while the current approach conducts individual q-entry tests that assume all remaining Q-entries are correct. Both methods differ with respect to 4 of the 12 q-entries compared and shown in Table 18. Striking the current method matches the expert designed q-entries for 11 of the 12 q-entries. There is complete agreement between DeCarlo's and the current method for Item 10 and the four q-entries where the current method differs from DeCarlo's are $q_{jk} = \left(q_{42}, q_{52}, q_{55}, q_{14,2}\right)$ where j=item and k=skill. DeCarlo (2012) suggests that skill 2 and 5 should be included for Item 5 but our results agree with the expert designed q-matrix in which skill 2 and 5 are not necessary to master to answer Item 5. In a similar manner DeCarlo (2012) found that skill 2 was not necessary to answer item 4 whereas our results agree with the expert designed q-matrix in which skill 2 is necessary to answer item 4. Lastly, the results by DeCarlo (2012) matched the expert designed q-matrix in which skill 2 is necessary to answer item 14 but the current method finds that skill 2 is not necessary to answer item 14.

CHAPTER 5: DISCUSSION

The primary purpose of the dissertation is to develop a new test statistic for the detection of diagnostic model misspecifications of the Q-matrix, which include both underfitting the Q-matrix (i.e., specifying 0s where there should be 1s) and overfitting the Q-matrix (i.e., specifying 1s where there should be 0s). In addition to the development of this new test statistic, this dissertation evaluated the performance of this new test statistic and developed an estimator of the asymptotic variance V_d based on the Fisher Information Matrix of the slip and guess parameters. The test statistic was evaluated by two simulation studies and also applied to the fraction subtraction dataset. The two simulation studies were used to test how well the normal approximation and the estimate of the variance works for a variety of sample sizes, attribute correlations and Item discriminations. Results were very much in line with the central limit theorem (CLT) which states that the sum of identically and independently distributed (i.i.d.) random variables with well-defined expected value and variance will approach normality when the number of observations gets sufficiently large, regardless of the underlying distribution. However when sample size is small and item discrimination (=4) is highest the Q-Q plots show departures from normality.

The simulation studies conducted in this dissertation made important strides in better understanding the conditions in which the inferential test will be most useful. The results of the simulation study showed the test to be very powerful in the detection of an incomplete Q-matrix. Though the conditions considered are in no way exhaustive, they were carefully chosen so as to be reflective of those encountered in practice. However, one possible limitation for this dissertation is that the number of items per skill was not investigated which may prove to be an interesting factor to investigate for future work.

The first simulation study investigated the true Type-I error rates for the test under 36 conditions (4x3x3) which include four levels of sample size, three levels of correlation among attributes and three levels of item discrimination. Results showed that as the sample size increases to 2,000 the Type I error reduces to the nominal value of $\alpha = .05$. No relationship was seen between skill correlation and Type I error rates. Surprisingly, the results for the relationship between Type I error and Item discrimination show that the most discriminating items (Item Discrimination of 4) have the largest Type I error rates and move the furthest away from the nominal value of $\alpha = .05$. This finding is most pronounced for the group of q-entries for which the true q-entry is zero and belongs to items that require two skills to answer the item correctly.

Another point of emphasis in this dissertation has been the need for some kind of multiplicity control when conducting a large number of hypothesis tests. This study looked at the effect of a Bonferroni as well as Benjamini Hochberg correction. Found strong evidence to support the need for these corrections which yield marked improvements in reducing Family-wise Type I error when compared to the uncorrected. Interestingly the Benjamini Hochberg and Bonferroni both yield similar improvements in reducing the Family-wise Type I error inflation across Sample size, Correlation and Item discrimination. Increasing Sample size clearly reduces the Family-wise Type I error for the Benjamin Hochberg and Bonferroni corrections considerably more than the uncorrected.

The second study is a power analysis in which the performance of the test statistic is evaluated by assessing how often the proposed statistic correctly finds that the hypothesized model is underspecified or over specified, thus identifying correctly the need for an addition of a necessary attribute or the deletion of an unnecessary attribute. This involved creating four types of q-matrices in which the first two will have a single misclassified q-entry and the third and fourth q-matrix will have 2 q-matrix entries that have been ‘switched’ to an incorrect value during the estimation stage. The main purpose of this analysis was to determine, across a variety of sample sizes, attribute correlations, and Item discriminations, the ability of the statistic to correctly indicate when a Q-matrix modification is in fact needed, thus showing its power.

The power study showed that for either Case 1 where the Q matrix underspecified by a single Q-entry or Case 2 where the Q matrix is over specified by a single Q-entry the statistic is very powerful in the detection of under-specification or over-specification of the Q-matrix with large sample sizes $(\geq 2,000)$ and/or when items are highly discriminating (≥ 2.5) between students that have mastered or have not mastered a skill. The results for both Case 3 in which the Q matrix has two misspecifications across two items and Case 4 in which the Q matrix has two misspecifications within the same item show that as sample size increases the proportion of samples correctly rejecting both Nulls increases significantly. Its noteworthy that detection of under-specification is better than for over-specification when two misclassifications are being tested simultaneously.

When both Type I error and Power results are assessed together it is found that item discrimination (>=2.5) is very powerful in the detection of under or over specification however this increased power comes at a price of increased Type I error for items that have the highest item discrimination (=4) only. This increase in Type I error is not seen when item discrimination is 1 or 2.5. Additionally, increasing sample size provides both a reduction in Type I error and a significant improvement in Power. No relationship is found for correlation.

The analysis of the fraction subtraction dataset using the test statistic found that 11 out of the 75 (15%) q-entries had enough evidence to reject the Null hypothesis that the original expert designed Q-entry is correct. This clearly indicates that the test finds misfit in the original expert designed Q-matrix, a finding which has been found by various other researchers (DeCarlo, 2012; Chung, 2014; Chen C.-M. , 2013).

The results from this dissertation can provide insight on the conditions necessary to properly apply this testing procedure in the field. This inference based method ideally requires having a minimum sample size of 2,000 examinees, a complete q-matrix (when for each skill there exists an item only requiring that skill), the underlying cognitive diagnosis model is DINA, Q-matrix has been initially designed by domain experts, items on the test show lower discrimination (<4) and lastly the overall number of Q-matrix misspecifications expected to be small.

Nevertheless meeting all these conditions in the field may not always be feasible and hence a few guidelines are provided for the following scenarios; when sample size is small and/or item discriminations are large (=4), overall number of Q-matrix misspecifications are expected to be large, and finally what to do when the Q-matrix is expected to be completely wrong. In the situation of small sample sizes and/or test items that have high discrimination the data can no longer assume normality consequently it may work better to calculate the p-values for the inference test(s) using parametric bootstrapping. Parametric bootstrapping will allow for better estimation of the p-values specifically for small sample sizes. When the Q-matrix is expected to contain a large number (10%-25%) of Q-matrix misspecifications then Tu et al. (2012) uses a γ method which does a better job for greater number of Q-matrix misspecifications. If the Q-matrix may be altogether wrongly specificed then it would be better to avoid using this inference procedure and choose instead to employ purely empirical methods of Q-matrix estimation. Researchers like Chung (2014) extended DeCarlo's (2012) Bayesian approach to estimate the whole Q-matrix of the DINA model and the reduced reparameterized unified model (rRUM). Additional exploratory research on estimating the Q-matrix without any expert input was done by Xiang (2013). The research evaluates the method through simulation studies and applies it to estimate Q matrix from real item response data.

Future Directions for Research

De la Torre (2008) and Liu et al. (2012) both proposed empirically based methods for modifying a Q-matrix within the DINA framework. However, neither of these methods are inferential procedures that could be used to test statistical hypotheses. This dissertation presented an inferential test as a method for detecting and correcting model under or over specification, and the work in this dissertation represents an initial step in understanding their usefulness in practical applications. Additionally, the intended use of this method should complement and in the best-case scenario work in conjunction with rather than replace domain experts' opinions.

Clearly, the conditions considered in the simulation study could not be exhaustive. Additional limitations of this work are that the method require large sample sizes (>= 2,000) and has only been applied to the DINA framework. Additional research needs to be done in order to understand the behavior of this test for DCMs across a broader range of assessment design conditions, diagnostic model specifications and incidences of Q-matrix misspecification. The current method could perhaps be extended for non-DINA models using de la Torre & Chiu (2016) extension of the sequential EM based delta method to the generalized DINA (G-DINA) model. In de la Torre & Chiu (2016) ς_j^2 is the generalization of the discrimination index δj and can be used on the G-DINA model which uses a more flexible parameterization. The G-DINA model can be converted to a class of reduced CDMs such as DINA and DINO models, additive CDMs, linear logistic models, and reparameterized unified models.

In the current work both Bonferroni and Benjamini Hochberg corrections were found to be necessary for controlling the inflation of the Type I Error(s). Another meaningful next step for future work would be to conduct a omnibus test as an alternate to the conducted correction methods. An omnibus test will provide a way to test whether the family-wise Q-matrix is correctly specified but a rejection of the Null hypothesis does not provide any information about the specific problematic Q-entries.

Moreover, the simulations for the hypothesis test were based on a complete and identified Q-matrix and therefore another possible area of investigation would be to test the performance of the test when the Q-matrix is no longer identified or complete. This in particular will be reflective of something that is often encountered in practice such that it will help make this test more amenable to becoming more generalizable to empirical applications. Additional area of investigation for future work lies in understanding how varying the number of items per skill can impact results. Another possible application for future research of this inferential testing would be for the identification of attributes which may have hierarchical structure such that mastery of basic attributes is prerequisite for mastering more complex attributes (Leighton Gierl, & Hunka, 2004; Templin & Bradshaw, 2014). Currently CDM examples assume independent cognitive skills and, in such cases, CDMs need to take the hierarchical structure into account; otherwise, they may not be appropriate and useful (Templin & Bradshaw, 2014).

REFERENCES

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

Barnes, T. (2010). Novel derivation and application of skill matrices: The q-matrix method. In *Handbook on educational data mining* (pp. 159-172). Boca Raton: CRC Press.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B, 57*, 289-300.

Chen, J., & de la Torre, J. (2013). A General Cognitive Diagnosis Model for Expert-Defined Polytomous Attributes. *Applied Psychological Measurement, 37*(6), 419-437.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical Analysis of Q-Matrix Based Diagnostic Classification Models. *Journal of the American Statistical Association, 110*(510).

Chen-Miao, C. (2013). Examining Uncertainty and Misspecifications of Attributes in Cognitive DIagnostic Models. (Doctoral dissertation).

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

Chung, M. -t. (2014). Estimating the Q-matrix for Cognitive Diagnosis Models in a Bayesian Framework. New York: Columbia University.

Cribbie, R. A. (2007). Multiplicity Control in Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(1), 98-112.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applicatons. *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., & Chiu, C.-Y. (2016). A General Method of Empirical Q-matrix Validation. *Psychometrika, 81*(2), 253-273.

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

DeCarlo, L. (2012). Recognizing Uncertainty in the Q-Matrix via a Bayesian Extension of the DINA Model. *Applied Psychological Measurement, 36*(6), 447-468.

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

Field, A. (2009). *Discovering Statistics Using SPSS. Third edition.* Los Angeles: Sage Publications.

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

Henson, R. 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.

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

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*(74), 191-210.

Im, S., & Corter, J. E. (2011). Statistical consequences of attribute misspecification in the Rule Space method. *Educational and Psychological Measurement, 57*(3), 712-731.

Johnson, M. (2015). A hypothesis testing procedure for Q-matrix entries. *Psychometric Society*.

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.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*(49), 59-81.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement, 49*, 59-81.

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

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuokas rule-space approach. *Journal of Educational Measurement, 41*, 205-237.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36*, 609-618.

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

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*(2), 99-120.

Muthen, L. K., & Muthen, B. O. (1998-2010). *Mplus User's Guide. Sixth Edition.* Los Angeles, CA: Muthen & Muthen.

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

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

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

Sorbom, D. (1989). Model modification. *Psychometrika, 54*, 371-384.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*(51), 337-350.

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

Templin, J. (2004). Estimation of the RUM without alpha tilde: a general model for the proficiency space of examinee ability.

Templin, J. L., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psyhometrika, 79,* 317-339.

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

Tu, D. B., Cai, Y., & Dai, H. Q. (2012). A new method of Q-matrix validation based on DINA model. *Acta Psychologica Sinica, 44*(4), 558-568.

Von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the genearlized partial credit model. *Applied Psychological Measurement*(28), 389-406.

Xiang, R. (2013). Nonlinear Penalized Estimation of True Q-Matrix in Cognitive Diagnostic Models. New York: Columbia Uiversity.

APPENDIX A

Type I Error for the test with a nominal level α=0.05 by Sample Size, Correlation and Item Discrimination

Item	Skill	Correlation			Item Discrimination			Sample Size			
		0.0	0.3	0.5	1	2.5	4	500	1,000	2,000	5,000
1	1	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
2	1	0.07	0.06	0.05	0.06	0.06	0.06	0.07	0.05	0.05	0.06
3	1	0.07	0.06	0.06	0.05	0.06	0.07	0.07	0.07	0.07	0.05
4	1	0.05	0.07	0.05	0.06	0.05	0.06	0.07	0.05	0.05	0.05
5	1	0.02	0.00	0.00	0.01	0.01	0.01	0.02	0.00	0.01	0.00
6	1	0.06	0.06	0.06	0.06	0.06	0.06	0.07	0.07	0.05	0.05
7	1	0.05	0.07	0.06	0.06	0.06	0.06	0.07	0.06	0.05	0.05
8	1	0.07	0.06	0.06	0.06	0.07	0.06	0.07	0.06	0.07	0.05
9	1	0.06	0.05	0.06	0.05	0.05	0.07	0.07	0.06	0.05	0.05
10	1	0.09	0.11	0.08	0.04	0.07	0.16	0.12	0.12	0.08	0.06
11	1	0.08	0.14	0.09	0.09	0.09	0.13	0.20	0.08	0.08	0.05
12	1	0.10	0.08	0.07	0.06	0.05	0.13	0.11	0.12	0.05	0.05
13	1	0.07	0.06	0.09	0.10	0.06	0.05	0.10	0.09	0.05	0.05
14	1	0.09	0.09	0.09	0.06	0.07	0.13	0.14	0.10	0.07	0.06
15	1	0.06	0.08	0.07	0.06	0.06	0.10	0.12	0.06	0.06	0.06
16	1	0.07	0.07	0.08	0.05	0.05	0.12	0.12	0.09	0.04	0.05
1	2	0.07	0.06	0.06	0.06	0.08	0.06	0.07	0.08	0.05	0.05
2	2	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
3	2	0.07	0.06	0.06	0.06	0.07	0.07	0.06	0.07	0.07	0.06
4	2	0.05	0.06	0.06	0.06	0.05	0.06	0.08	0.05	0.05	0.05
5	2	0.06	0.06	0.07	0.06	0.06	0.06	0.07	0.06	0.06	0.05
6	2	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00
7	2	0.06	0.07	0.06	0.06	0.07	0.06	0.07	0.07	0.05	0.05
8	2	0.07	0.05	0.06	0.06	0.07	0.05	0.07	0.05	0.07	0.05
9	2	0.06	0.05	0.06	0.05	0.05	0.06	0.07	0.06	0.05	0.05
10	2	0.06	0.07	0.06	0.06	0.07	0.06	0.08	0.06	0.05	0.05
11	2	0.06	0.14	0.09	0.09	0.10	0.11	0.20	0.08	0.07	0.05
12	2	0.10	0.08	0.06	0.07	0.08	0.10	0.15	0.07	0.05	0.05
13	2	0.10	0.07	0.13	0.12	0.06	0.12	0.16	0.14	0.06	0.05
14	2	0.08	0.08	0.09	0.06	0.08	0.12	0.13	0.09	0.07	0.06
15	2	0.06	0.06	0.08	0.06	0.06	0.07	0.09	0.06	0.05	0.05
16	2	0.06	0.07	0.07	0.06	0.07	0.08	0.08	0.08	0.06	0.05

Type I Error for the test with a nominal level α=0.05 by Sample Size, Correlation and Item Discrimination

Item	Skill	Correlation			Item Discrimination			Sample Size			
		0.0	0.3	0.5	1	2.5	4	500	1,000	2,000	5,000
1	3	0.06	0.06	0.06	0.06	0.07	0.05	0.06	0.08	0.05	0.05
2	3	0.07	0.05	0.05	0.06	0.06	0.05	0.08	0.05	0.05	0.05
3	3	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0.02	0.00	0.00
4	3	0.05	0.07	0.06	0.06	0.05	0.06	0.08	0.05	0.06	0.05
5	3	0.05	0.06	0.06	0.06	0.06	0.05	0.06	0.06	0.05	0.06
6	3	0.05	0.06	0.06	0.06	0.06	0.05	0.08	0.06	0.05	0.05
7	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	3	0.06	0.07	0.05	0.06	0.07	0.05	0.08	0.06	0.06	0.05
9	3	0.07	0.07	0.07	0.05	0.07	0.09	0.10	0.08	0.05	0.05
10	3	0.05	0.07	0.05	0.05	0.07	0.05	0.09	0.06	0.05	0.04
11	3	0.06	0.11	0.05	0.08	0.09	0.05	0.15	0.06	0.05	0.05
12	3	0.09	0.08	0.07	0.06	0.07	0.11	0.12	0.10	0.06	0.05
13	3	0.10	0.07	0.12	0.11	0.06	0.11	0.16	0.14	0.05	0.05
14	3	0.09	0.07	0.05	0.05	0.09	0.06	0.12	0.05	0.05	0.05
15	3	0.07	0.09	0.07	0.06	0.07	0.09	0.14	0.06	0.06	0.05
16	3	0.07	0.07	0.07	0.06	0.06	0.09	0.09	0.08	0.06	0.05
1	4	0.06	0.06	0.05	0.05	0.07	0.06	0.06	0.07	0.06	0.05
2	4	0.06	0.06	0.05	0.06	0.06	0.06	0.08	0.05	0.05	0.06
3	4	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.05	0.05
4	4	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00
5	4	0.05	0.06	0.06	0.06	0.05	0.06	0.07	0.06	0.05	0.05
6	4	0.06	0.06	0.06	0.06	0.06	0.06	0.07	0.06	0.06	0.05
7	4	0.06	0.06	0.06	0.06	0.05	0.07	0.06	0.07	0.05	0.05
8	4	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00
9	4	0.06	0.06	0.07	0.05	0.06	0.07	0.08	0.06	0.05	0.06
10	4	0.06	0.09	0.07	0.05	0.06	0.10	0.08	0.09	0.06	0.05
11	4	0.06	0.12	0.06	0.08	0.09	0.06	0.17	0.05	0.05	0.05
12	4	0.10	0.08	0.06	0.06	0.08	0.10	0.16	0.06	0.05	0.05
13	4	0.07	0.05	0.09	0.11	0.06	0.06	0.09	0.10	0.05	0.05
14	4	0.09	0.06	0.05	0.06	0.10	0.06	0.13	0.05	0.05	0.05
15	4	0.06	0.06	0.06	0.05	0.06	0.07	0.07	0.07	0.05	0.05
16	4	0.07	0.08	0.08	0.05	0.05	0.12	0.12	0.10	0.05	0.05

Type I Error for the test with a nominal level α=0.01 by Sample Size, Correlation and Item Discrimination

Item	Skill	Correlation			Item Discrimination			Sample Size			
		0.0	0.3	0.5	1	2.5	4	500	1,000	2,000	5,000
1	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	1	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.01	0.01	0.01
3	1	0.03	0.02	0.02	0.01	0.02	0.03	0.02	0.02	0.03	0.01
4	1	0.01	0.03	0.01	0.02	0.01	0.02	0.03	0.01	0.01	0.01
5	1	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.00
6	1	0.01	0.02	0.01	0.02	0.01	0.02	0.02	0.02	0.01	0.01
7	1	0.01	0.02	0.02	0.02	0.02	0.02	0.03	0.02	0.01	0.01
8	1	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.01
9	1	0.01	0.02	0.02	0.01	0.01	0.02	0.03	0.02	0.01	0.01
10	1	0.04	0.06	0.04	0.01	0.03	0.10	0.07	0.07	0.03	0.01
11	1	0.02	0.09	0.04	0.04	0.04	0.07	0.15	0.04	0.03	0.01
12	1	0.05	0.04	0.03	0.02	0.01	0.08	0.07	0.07	0.02	0.01
13	1	0.03	0.01	0.04	0.05	0.02	0.01	0.05	0.04	0.01	0.01
14	1	0.04	0.05	0.04	0.02	0.03	0.08	0.08	0.05	0.03	0.02
15	1	0.02	0.03	0.03	0.01	0.02	0.05	0.06	0.02	0.01	0.02
16	1	0.03	0.04	0.04	0.01	0.01	0.07	0.07	0.05	0.01	0.01
1	2	0.02	0.02	0.02	0.02	0.03	0.01	0.02	0.03	0.01	0.01
2	2	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
3	2	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01
4	2	0.01	0.02	0.02	0.02	0.01	0.02	0.03	0.02	0.01	0.01
5	2	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.02	0.01	0.01
6	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	2	0.01	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.01	0.01
8	2	0.02	0.01	0.02	0.01	0.02	0.01	0.01	0.01	0.02	0.01
9	2	0.01	0.01	0.02	0.01	0.01	0.02	0.02	0.02	0.01	0.01
10	2	0.01	0.02	0.02	0.02	0.02	0.02	0.04	0.02	0.01	0.01
11	2	0.02	0.10	0.04	0.04	0.06	0.06	0.16	0.03	0.03	0.01
12	2	0.06	0.04	0.02	0.02	0.04	0.06	0.12	0.03	0.01	0.01
13	2	0.05	0.03	0.08	0.08	0.02	0.07	0.11	0.09	0.02	0.01
14	2	0.03	0.04	0.04	0.02	0.03	0.06	0.08	0.04	0.02	0.02
15	2	0.02	0.01	0.03	0.02	0.02	0.02	0.04	0.02	0.01	0.01
16	2	0.02	0.02	0.03	0.02	0.02	0.03	0.03	0.03	0.02	0.01

Type I Error for the test with a nominal level α=0.01 by Sample Size, Correlation and Item Discrimination

Item	Skill	Correlation			Item Discrimination			Sample Size			
		0.0	0.3	0.5	1	2.5	4	500	1,000	2,000	5,000
1	3	0.02	0.02	0.02	0.02	0.03	0.01	0.02	0.03	0.01	0.01
2	3	0.02	0.01	0.01	0.02	0.01	0.01	0.03	0.01	0.01	0.01
3	3	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00
4	3	0.01	0.02	0.01	0.02	0.01	0.02	0.03	0.01	0.01	0.01
5	3	0.01	0.02	0.02	0.01	0.02	0.01	0.03	0.01	0.01	0.01
6	3	0.01	0.02	0.01	0.02	0.02	0.01	0.03	0.01	0.01	0.01
7	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	3	0.02	0.02	0.01	0.02	0.02	0.02	0.03	0.01	0.02	0.01
9	3	0.03	0.02	0.03	0.02	0.02	0.04	0.05	0.03	0.01	0.01
10	3	0.01	0.02	0.01	0.01	0.02	0.01	0.03	0.01	0.01	0.01
11	3	0.02	0.07	0.01	0.04	0.05	0.01	0.11	0.01	0.01	0.01
12	3	0.05	0.03	0.02	0.02	0.01	0.07	0.06	0.05	0.01	0.01
13	3	0.05	0.03	0.07	0.07	0.02	0.07	0.11	0.09	0.01	0.01
14	3	0.04	0.03	0.01	0.01	0.05	0.01	0.08	0.01	0.01	0.01
15	3	0.02	0.04	0.03	0.02	0.02	0.05	0.08	0.02	0.02	0.01
16	3	0.02	0.02	0.03	0.02	0.02	0.04	0.04	0.04	0.02	0.01
1	4	0.02	0.02	0.01	0.01	0.02	0.01	0.01	0.03	0.01	0.01
2	4	0.02	0.02	0.01	0.02	0.01	0.01	0.03	0.01	0.01	0.01
3	4	0.01	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.01	0.01
4	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	4	0.01	0.01	0.02	0.02	0.01	0.02	0.02	0.02	0.01	0.01
6	4	0.01	0.01	0.02	0.01	0.01	0.02	0.02	0.02	0.01	0.01
7	4	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.02	0.01	0.01
8	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
9	4	0.02	0.02	0.02	0.01	0.02	0.02	0.03	0.01	0.01	0.02
10	4	0.02	0.03	0.03	0.01	0.02	0.05	0.04	0.04	0.02	0.01
11	4	0.02	0.08	0.02	0.04	0.05	0.02	0.12	0.01	0.01	0.01
12	4	0.06	0.04	0.01	0.02	0.03	0.06	0.12	0.02	0.01	0.01
13	4	0.03	0.01	0.05	0.06	0.02	0.01	0.05	0.05	0.01	0.01
14	4	0.05	0.02	0.01	0.02	0.06	0.01	0.08	0.01	0.01	0.01
15	4	0.02	0.02	0.02	0.01	0.02	0.02	0.03	0.02	0.01	0.01
16	4	0.03	0.04	0.04	0.02	0.02	0.08	0.08	0.05	0.01	0.01

Type I Error for the test with a nominal level α=0.10 by Sample Size, Correlation and Item Discrimination

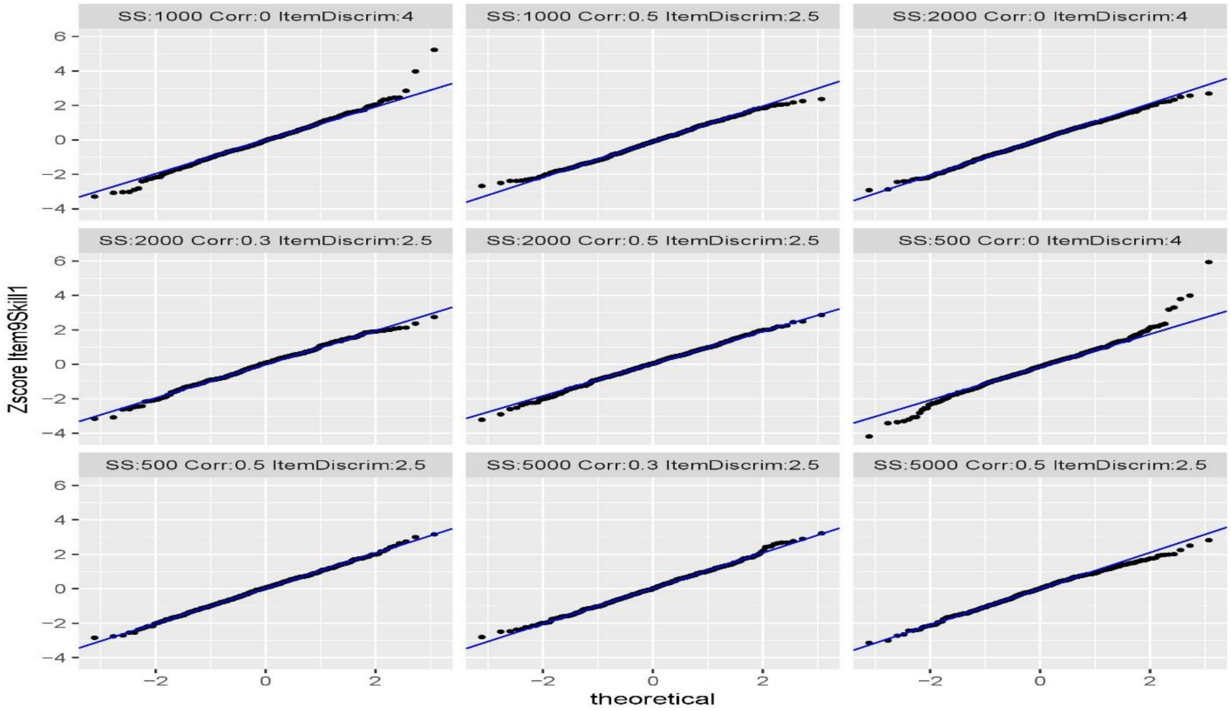
Item	Skill	Correlation			Item Discrimination			Sample Size			
		0.0	0.3	0.5	1	2.5	4	500	1,000	2,000	5,000
1	1	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.00
2	1	0.12	0.11	0.10	0.11	0.11	0.11	0.13	0.11	0.10	0.10
3	1	0.11	0.11	0.11	0.10	0.11	0.12	0.12	0.12	0.12	0.09
4	1	0.10	0.12	0.10	0.11	0.10	0.11	0.12	0.10	0.11	0.11
5	1	0.02	0.00	0.00	0.01	0.01	0.01	0.02	0.00	0.01	0.00
6	1	0.11	0.12	0.11	0.11	0.11	0.11	0.13	0.11	0.11	0.11
7	1	0.10	0.12	0.11	0.11	0.11	0.10	0.12	0.12	0.10	0.10
8	1	0.12	0.11	0.11	0.11	0.12	0.11	0.12	0.12	0.12	0.10
9	1	0.10	0.10	0.11	0.10	0.10	0.12	0.12	0.11	0.10	0.10
10	1	0.14	0.17	0.13	0.09	0.13	0.21	0.17	0.17	0.13	0.11
11	1	0.13	0.19	0.14	0.14	0.13	0.18	0.25	0.13	0.13	0.11
12	1	0.15	0.13	0.12	0.11	0.11	0.18	0.18	0.16	0.10	0.10
13	1	0.13	0.11	0.13	0.16	0.11	0.11	0.15	0.14	0.10	0.10
14	1	0.13	0.14	0.14	0.10	0.12	0.18	0.19	0.14	0.11	0.10
15	1	0.11	0.13	0.13	0.11	0.11	0.15	0.18	0.11	0.10	0.11
16	1	0.12	0.12	0.13	0.10	0.10	0.17	0.17	0.15	0.09	0.10
1	2	0.12	0.11	0.11	0.11	0.13	0.11	0.12	0.14	0.10	0.10
2	2	0.01	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.00
3	2	0.12	0.11	0.11	0.10	0.12	0.12	0.11	0.12	0.12	0.10
4	2	0.10	0.12	0.11	0.11	0.10	0.11	0.14	0.10	0.10	0.10
5	2	0.11	0.11	0.12	0.11	0.11	0.10	0.12	0.11	0.11	0.10
6	2	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.01
7	2	0.11	0.12	0.11	0.11	0.12	0.11	0.13	0.12	0.10	0.11
8	2	0.11	0.11	0.12	0.11	0.12	0.10	0.12	0.10	0.12	0.10
9	2	0.10	0.10	0.11	0.10	0.10	0.11	0.11	0.11	0.10	0.09
10	2	0.11	0.12	0.11	0.11	0.11	0.11	0.13	0.12	0.10	0.10
11	2	0.12	0.19	0.14	0.13	0.15	0.16	0.25	0.13	0.12	0.10
12	2	0.15	0.13	0.11	0.11	0.13	0.14	0.20	0.13	0.11	0.09
13	2	0.15	0.13	0.18	0.17	0.12	0.17	0.21	0.20	0.11	0.10
14	2	0.14	0.13	0.14	0.11	0.13	0.17	0.18	0.14	0.12	0.11
15	2	0.11	0.11	0.13	0.11	0.12	0.12	0.15	0.12	0.11	0.10
16	2	0.11	0.12	0.12	0.11	0.11	0.12	0.13	0.12	0.11	0.10

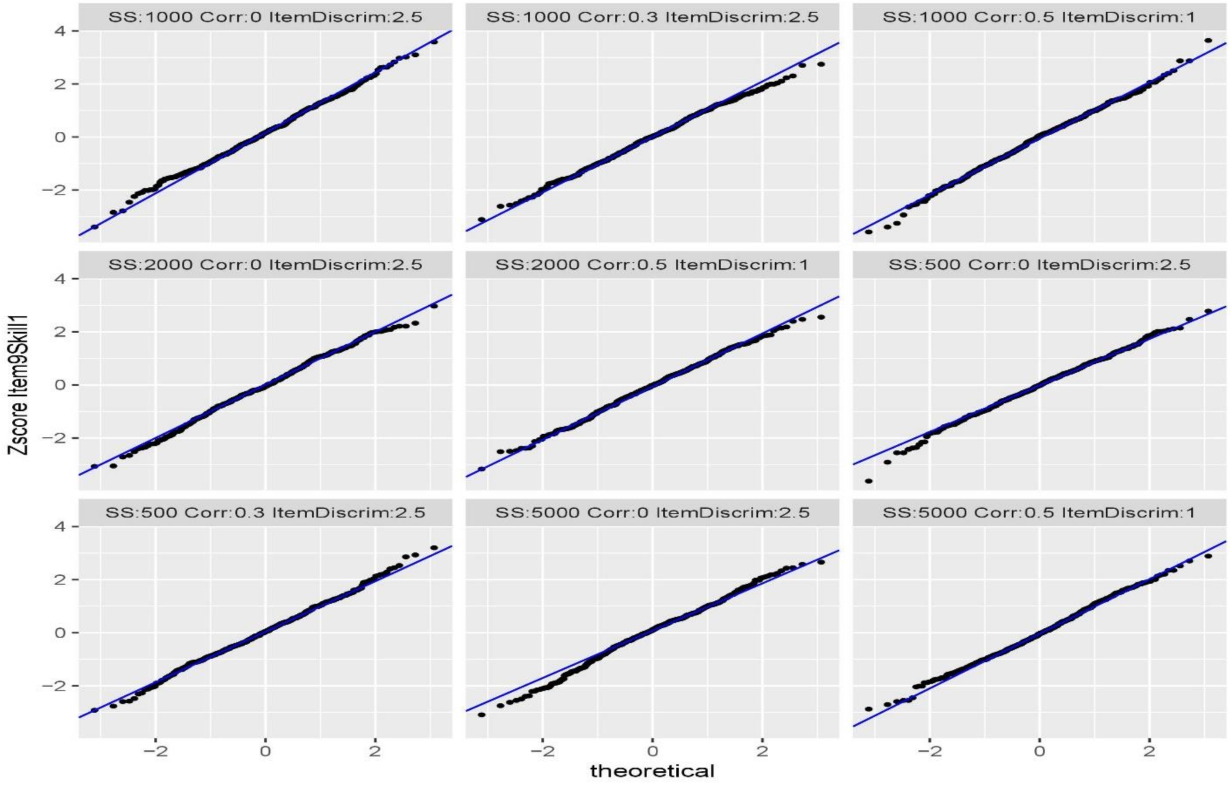
Type I Error for the test with a nominal level α=0.10 by Sample Size, Correlation and Item Discrimination

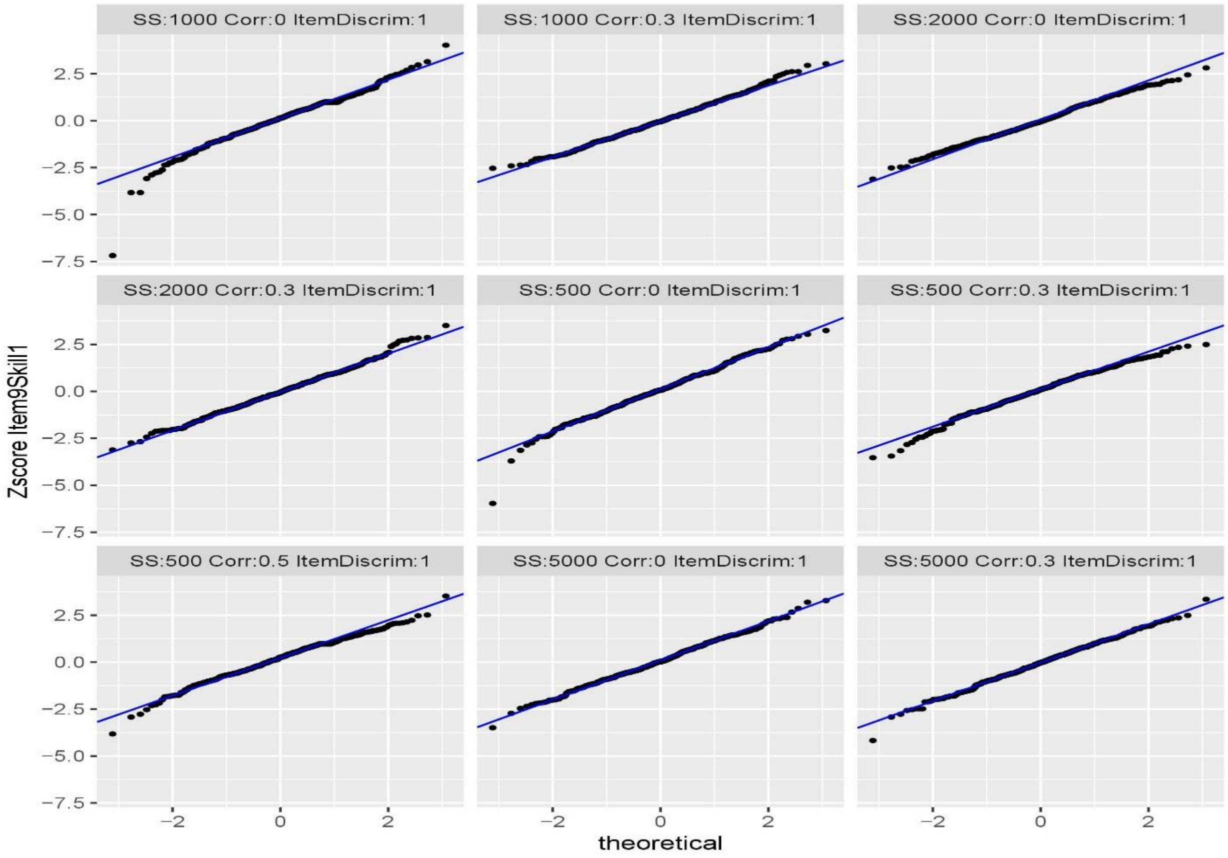
Item	Skill	Correlation			Item Discrimination			Sample Size			
		0.0	0.3	0.5	1	2.5	4	500	1,000	2,000	5,000
1	3	0.11	0.11	0.11	0.10	0.12	0.10	0.11	0.13	0.10	0.09
2	3	0.12	0.11	0.10	0.12	0.11	0.10	0.13	0.11	0.10	0.11
3	3	0.02	0.01	0.00	0.00	0.01	0.02	0.01	0.02	0.00	0.00
4	3	0.11	0.11	0.11	0.12	0.11	0.11	0.14	0.10	0.10	0.10
5	3	0.11	0.11	0.10	0.11	0.12	0.10	0.12	0.11	0.10	0.10
6	3	0.10	0.11	0.10	0.11	0.11	0.10	0.13	0.11	0.09	0.09
7	3	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
8	3	0.12	0.12	0.10	0.11	0.12	0.11	0.13	0.11	0.12	0.10
9	3	0.13	0.12	0.12	0.10	0.12	0.14	0.16	0.12	0.10	0.10
10	3	0.11	0.12	0.10	0.11	0.12	0.10	0.13	0.11	0.10	0.10
11	3	0.11	0.17	0.11	0.13	0.14	0.11	0.19	0.11	0.11	0.11
12	3	0.14	0.13	0.12	0.11	0.12	0.16	0.17	0.15	0.11	0.10
13	3	0.14	0.13	0.16	0.16	0.11	0.16	0.21	0.18	0.10	0.10
14	3	0.13	0.12	0.10	0.10	0.14	0.11	0.18	0.10	0.10	0.09
15	3	0.12	0.15	0.12	0.11	0.12	0.15	0.18	0.11	0.11	0.11
16	3	0.12	0.12	0.12	0.11	0.11	0.14	0.13	0.13	0.11	0.10
1	4	0.12	0.11	0.10	0.10	0.12	0.11	0.11	0.12	0.11	0.10
2	4	0.11	0.11	0.10	0.11	0.11	0.10	0.13	0.10	0.10	0.11
3	4	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.11	0.09	0.10
4	4	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00
5	4	0.10	0.11	0.11	0.11	0.10	0.11	0.12	0.11	0.10	0.10
6	4	0.11	0.11	0.11	0.11	0.11	0.11	0.12	0.10	0.11	0.10
7	4	0.10	0.11	0.10	0.10	0.10	0.11	0.11	0.11	0.09	0.10
8	4	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.00	0.00
9	4	0.11	0.10	0.12	0.11	0.11	0.12	0.13	0.11	0.11	0.11
10	4	0.10	0.14	0.12	0.10	0.11	0.16	0.13	0.14	0.11	0.10
11	4	0.11	0.17	0.11	0.13	0.14	0.11	0.22	0.10	0.11	0.10
12	4	0.15	0.13	0.10	0.11	0.13	0.15	0.20	0.11	0.10	0.10
13	4	0.13	0.10	0.14	0.16	0.10	0.11	0.14	0.15	0.11	0.11
14	4	0.15	0.11	0.10	0.11	0.15	0.11	0.18	0.10	0.10	0.10
15	4	0.12	0.11	0.11	0.11	0.11	0.12	0.13	0.12	0.11	0.10
16	4	0.11	0.13	0.14	0.10	0.10	0.17	0.17	0.14	0.10	0.09

Appendix B

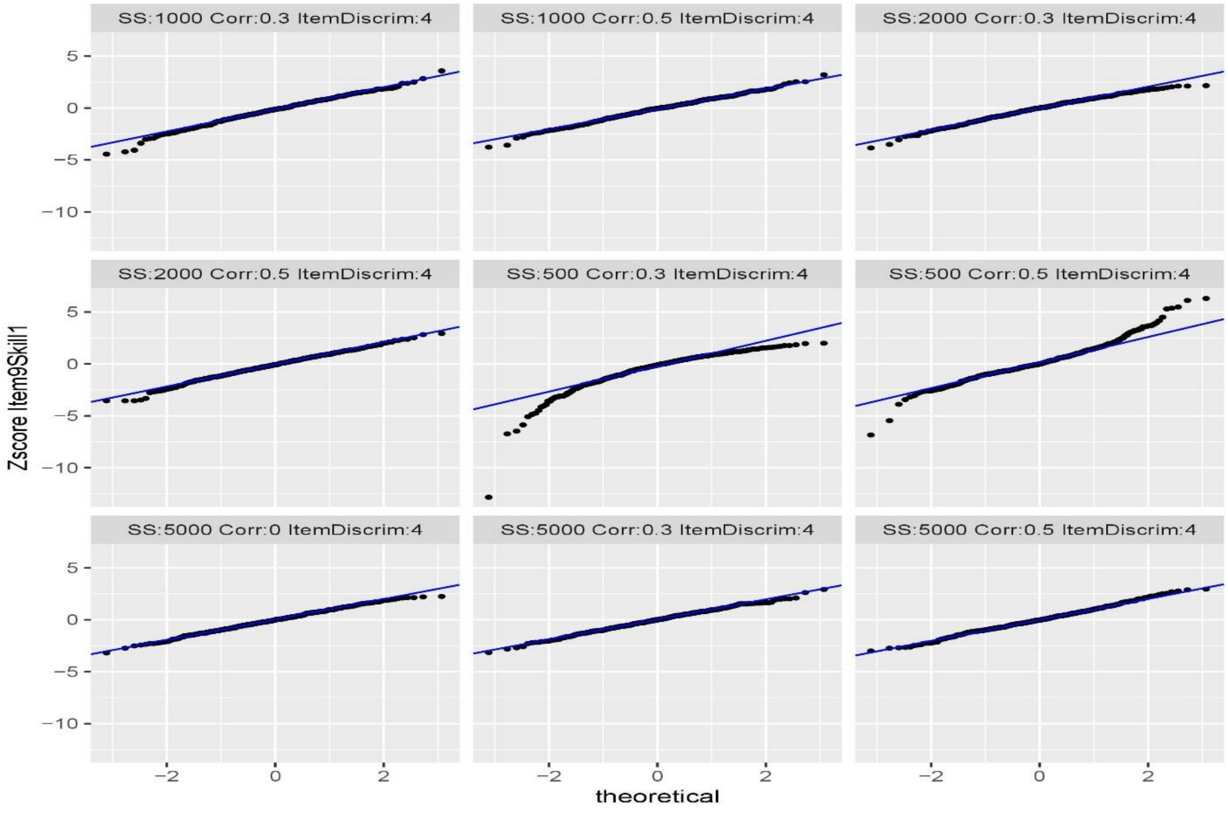
Q-Q Plots by Sample Size, Correlation and Item Discrimination for Item 9 Skill 1





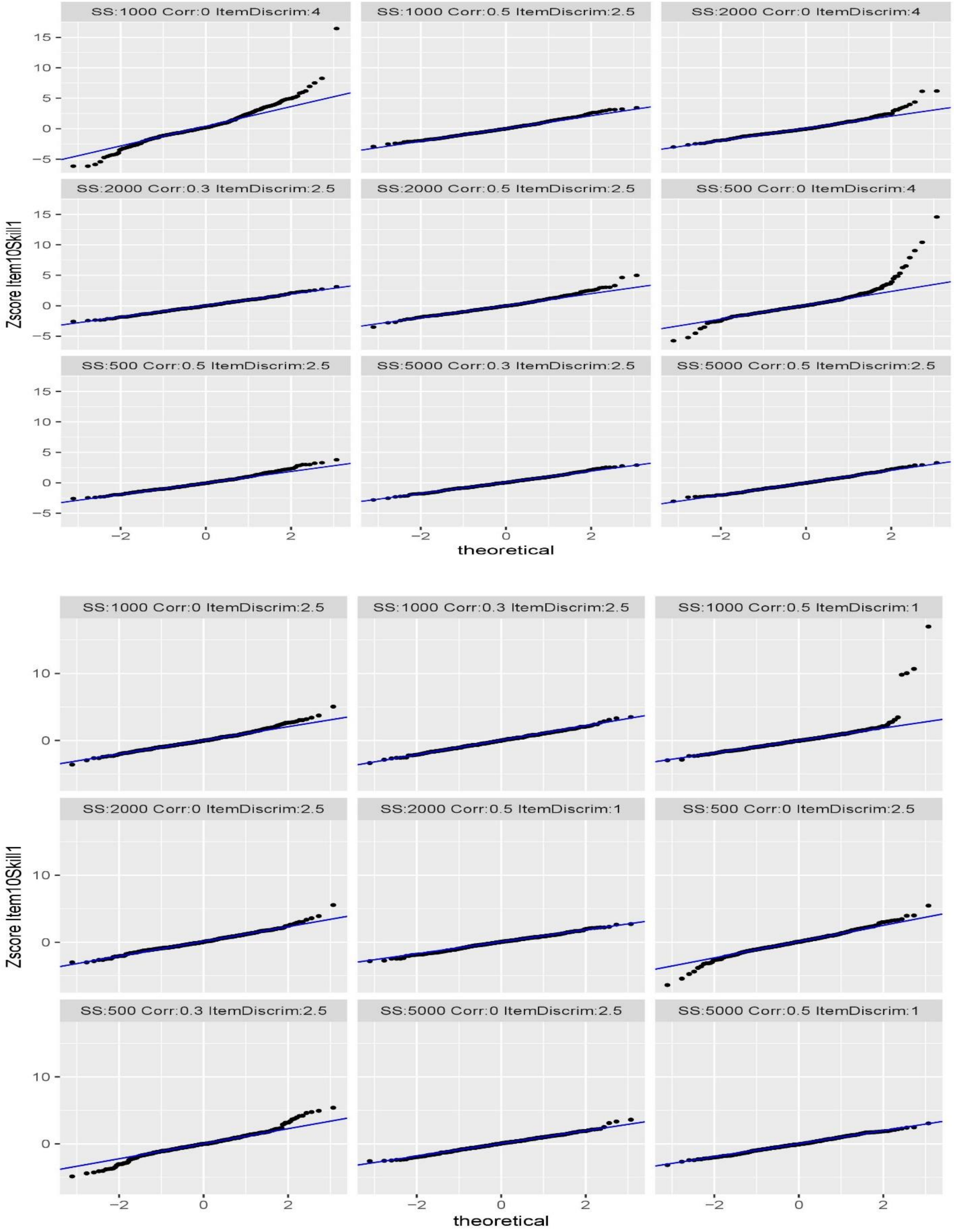


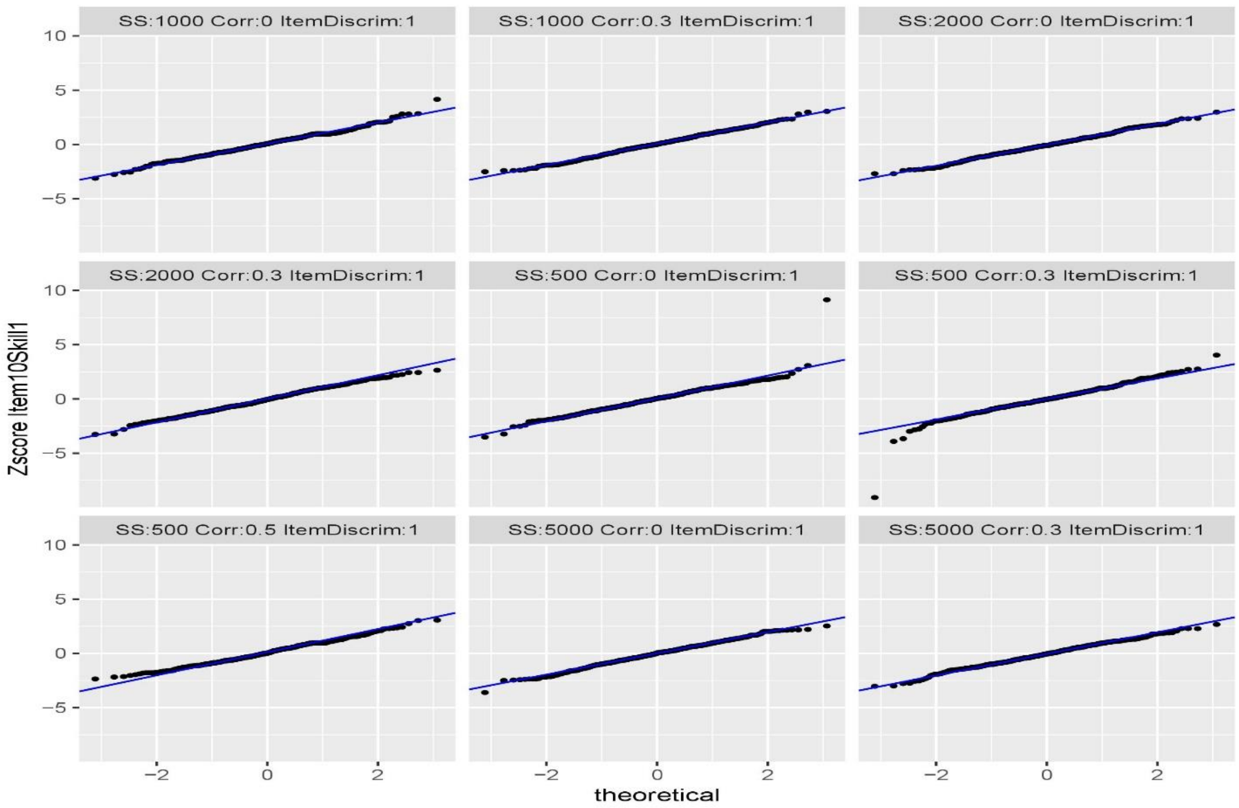
SS:1000 Corr:0 ItemDiscrim:1
SS:1000 Corr:0.3 ItemDiscrim:1
SS:2000 Corr:0 ItemDiscrim:1
SS:2000 Corr:0.3 ItemDiscrim:1
SS:500 Corr:0 ItemDiscrim:1
SS:500 Corr:0.3 ItemDiscrim:1
SS:500 Corr:0.5 ItemDiscrim:1
SS:5000 Corr:0 ItemDiscrim:1
SS:5000 Corr:0.3 ItemDiscrim:1
Zscore Item9Skill1
2.5
0.0
−2.5
−5.0
−7.5
−2
0
2
theoretical

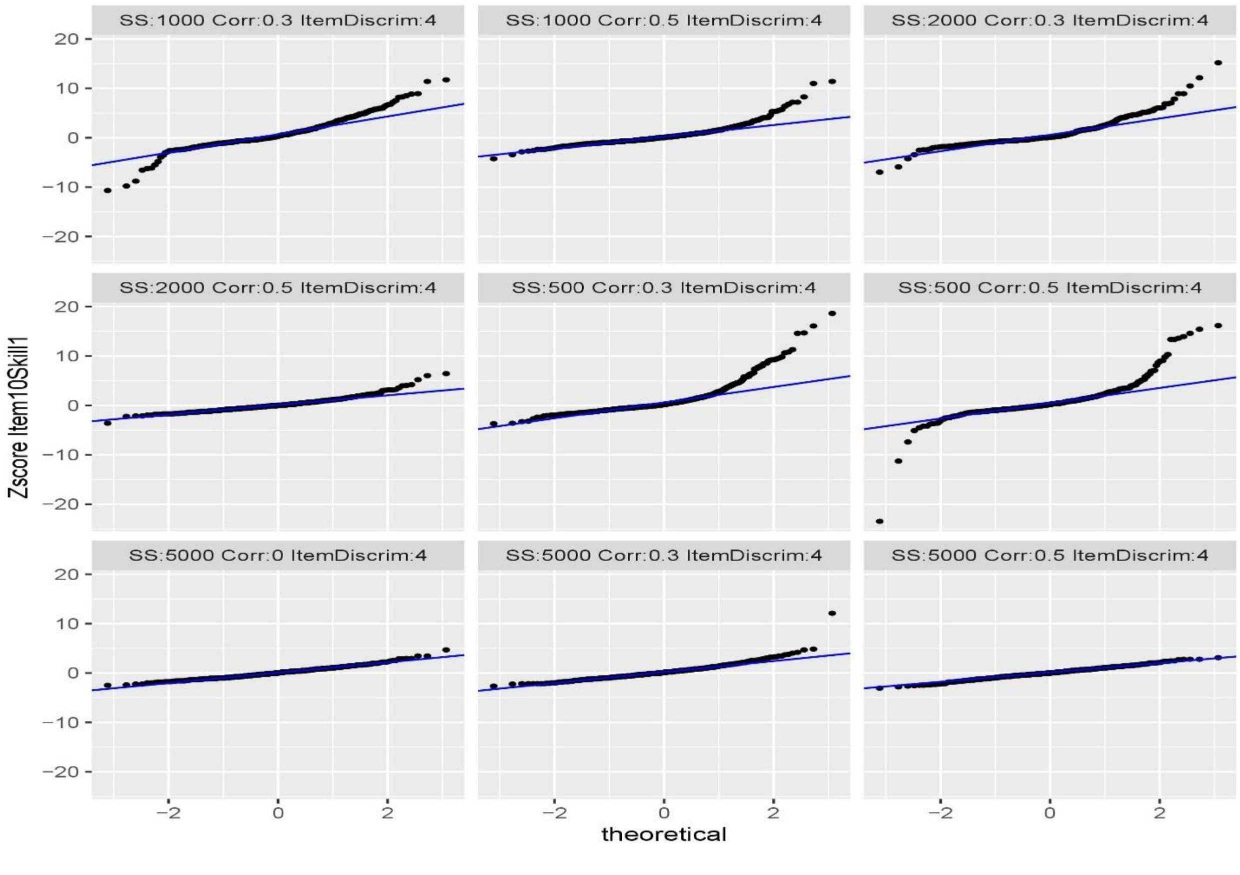


SS:1000 Corr:0.3 ItemDiscrim:4
SS:1000 Corr:0.5 ItemDiscrim:4
SS:2000 Corr:0.3 ItemDiscrim:4
SS:2000 Corr:0.5 ItemDiscrim:4
SS:500 Corr:0.3 ItemDiscrim:4
SS:500 Corr:0.5 ItemDiscrim:4
SS:5000 Corr:0 ItemDiscrim:4
SS:5000 Corr:0.3 ItemDiscrim:4
SS:5000 Corr:0.5 ItemDiscrim:4
Zscore Item9Skill1
5
0
−5
−10
−2
0
2
theoretical

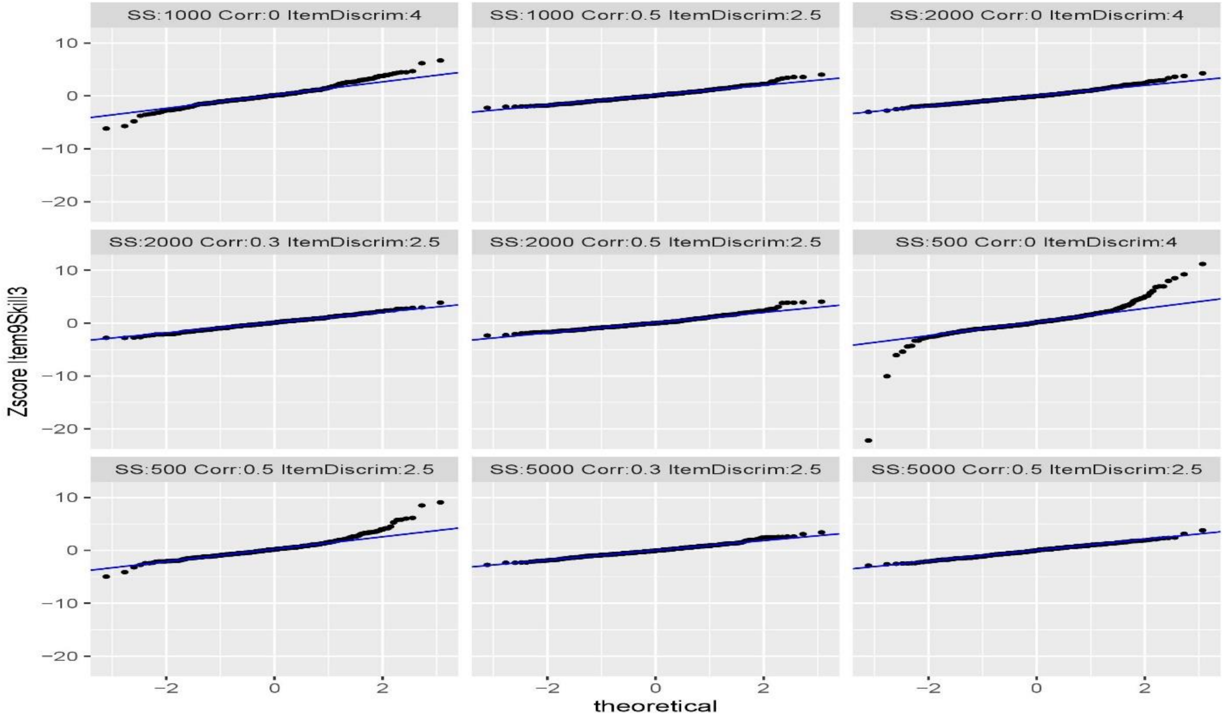
Q-Q Plots by Sample Size, Correlation and Item Discrimination for Item 10 Skill 1

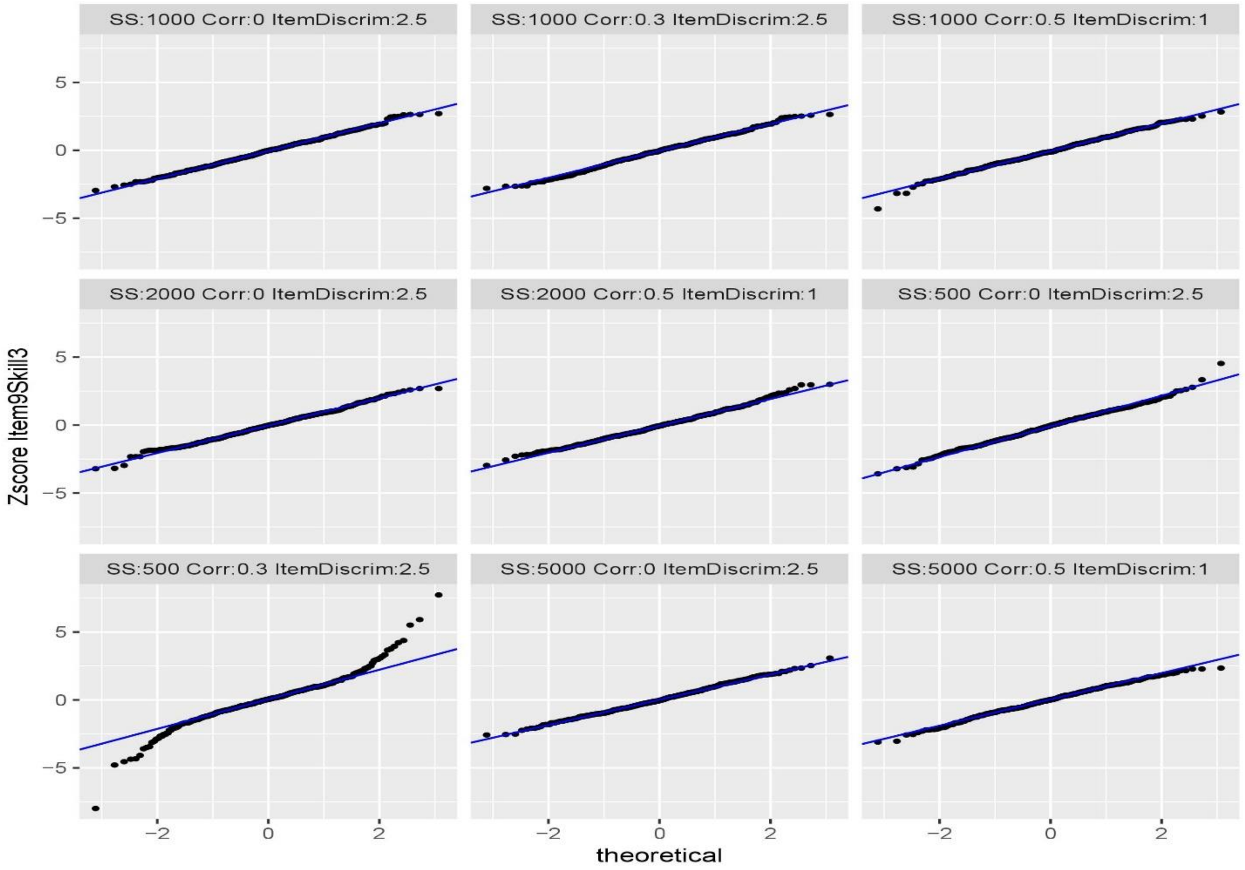


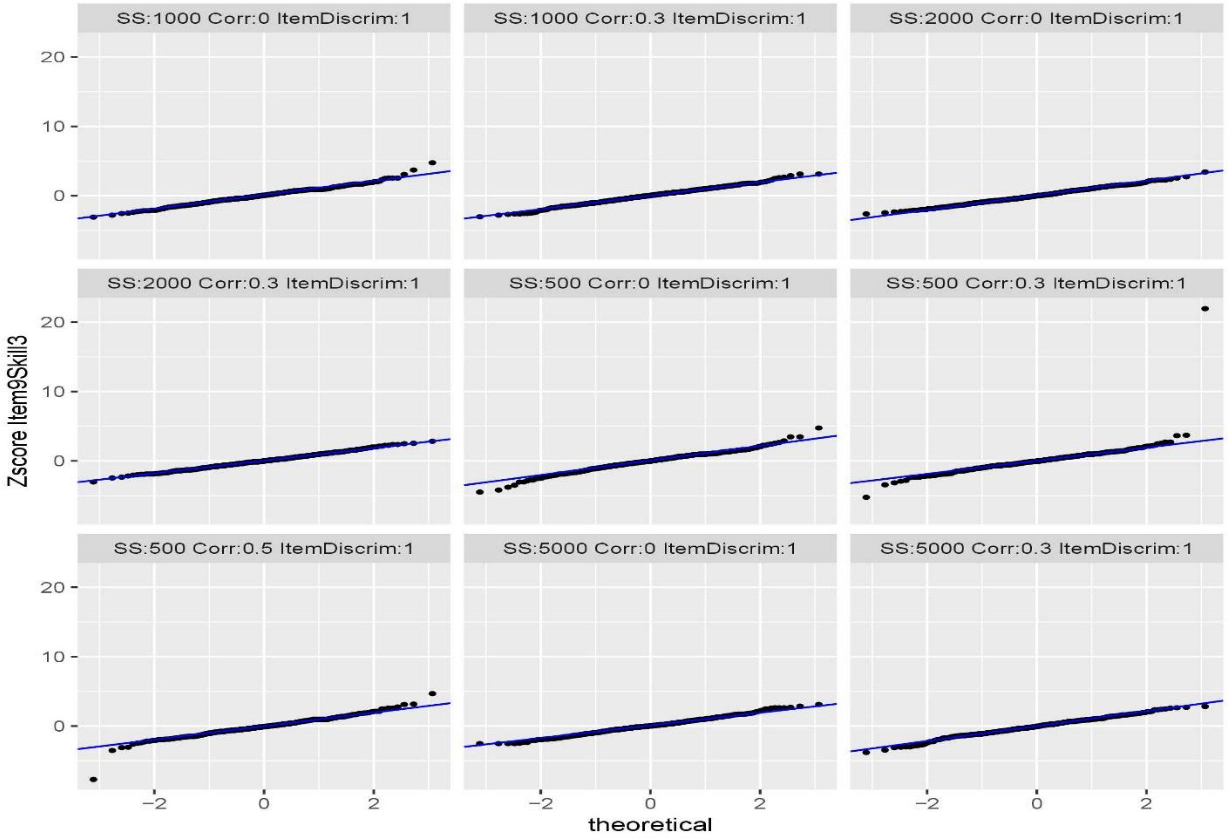

SS:1000 Corr:0 ItemDiscrim:1
SS:1000 Corr:0.3 ItemDiscrim:1
SS:2000 Corr:0 ItemDiscrim:1
SS:2000 Corr:0.3 ItemDiscrim:1
SS:500 Corr:0 ItemDiscrim:1
SS:500 Corr:0.3 ItemDiscrim:1
SS:500 Corr:0.5 ItemDiscrim:1
SS:5000 Corr:0 ItemDiscrim:1
SS:5000 Corr:0.3 ItemDiscrim:1
Zscore Item10Skill1
10
5
0
−5
−2
0
2
theoretical

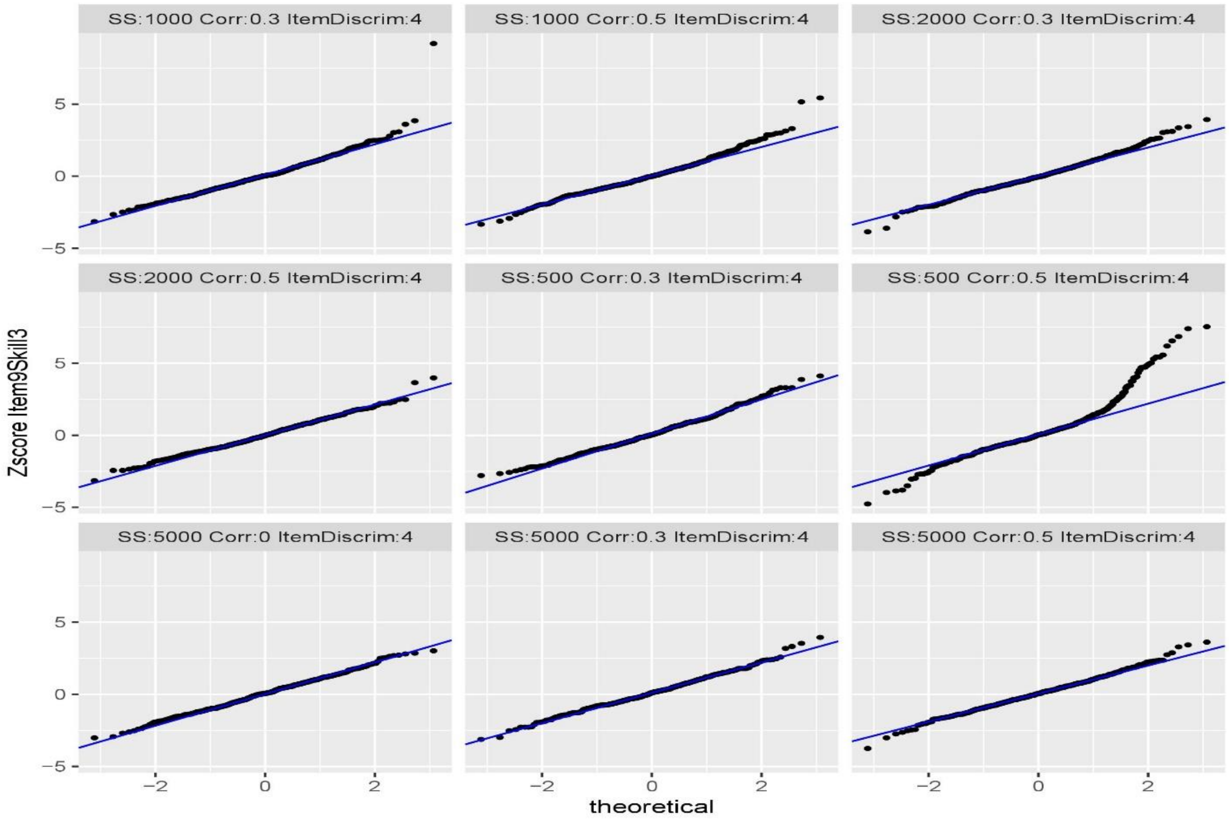

SS:1000 Corr:0.3 ItemDiscrim:4
SS:1000 Corr:0.5 ItemDiscrim:4
SS:2000 Corr:0.3 ItemDiscrim:4
SS:2000 Corr:0.5 ItemDiscrim:4
SS:500 Corr:0.3 ItemDiscrim:4
SS:500 Corr:0.5 ItemDiscrim:4
SS:5000 Corr:0 ItemDiscrim:4
SS:5000 Corr:0.3 ItemDiscrim:4
SS:5000 Corr:0.5 ItemDiscrim:4
Zscore Item10Skill1
20
10
0
−10
−20
−2
0
2
theoretical

Q-Q Plots by Sample Size, Correlation and Item Discrimination for Item 9 Skill 3

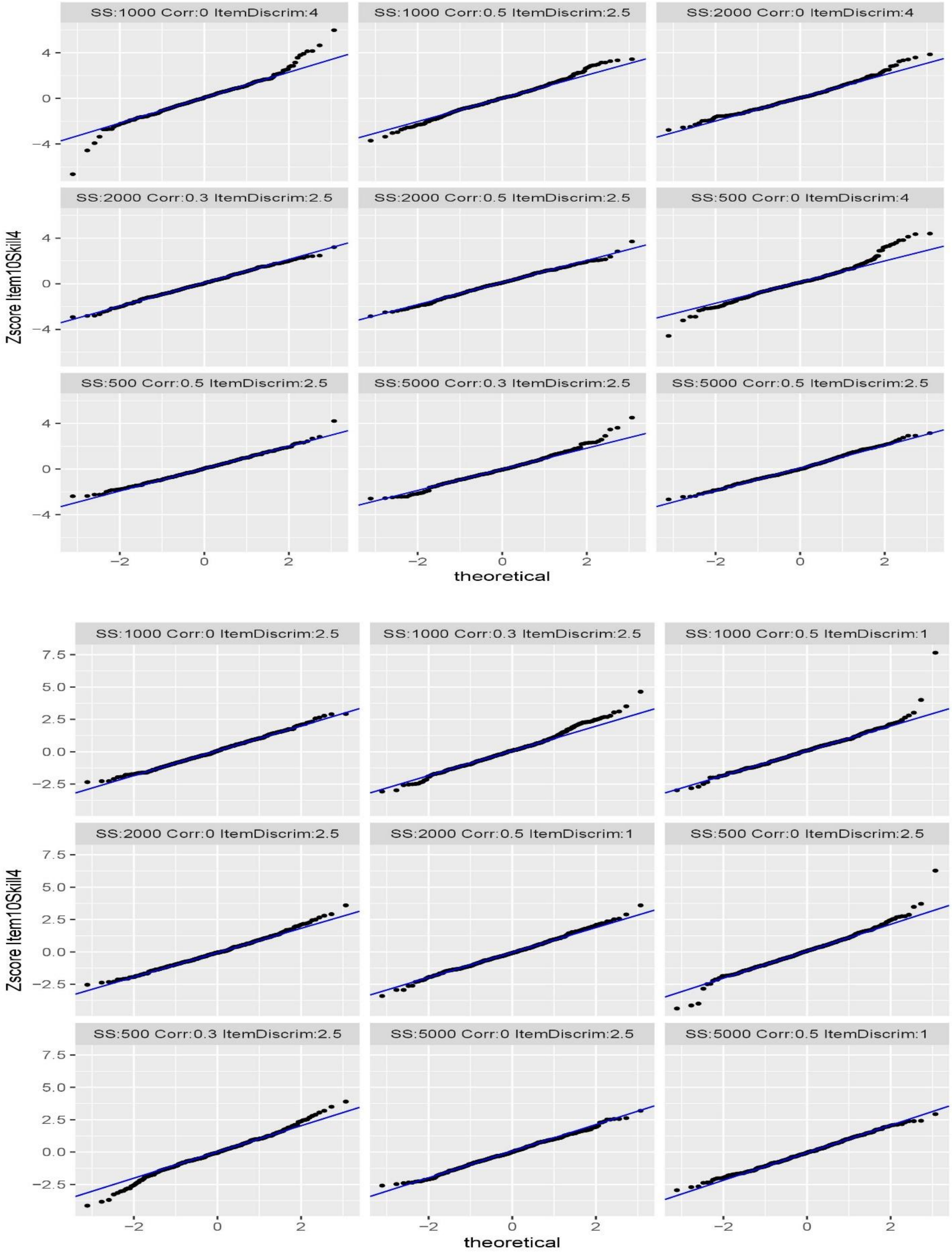


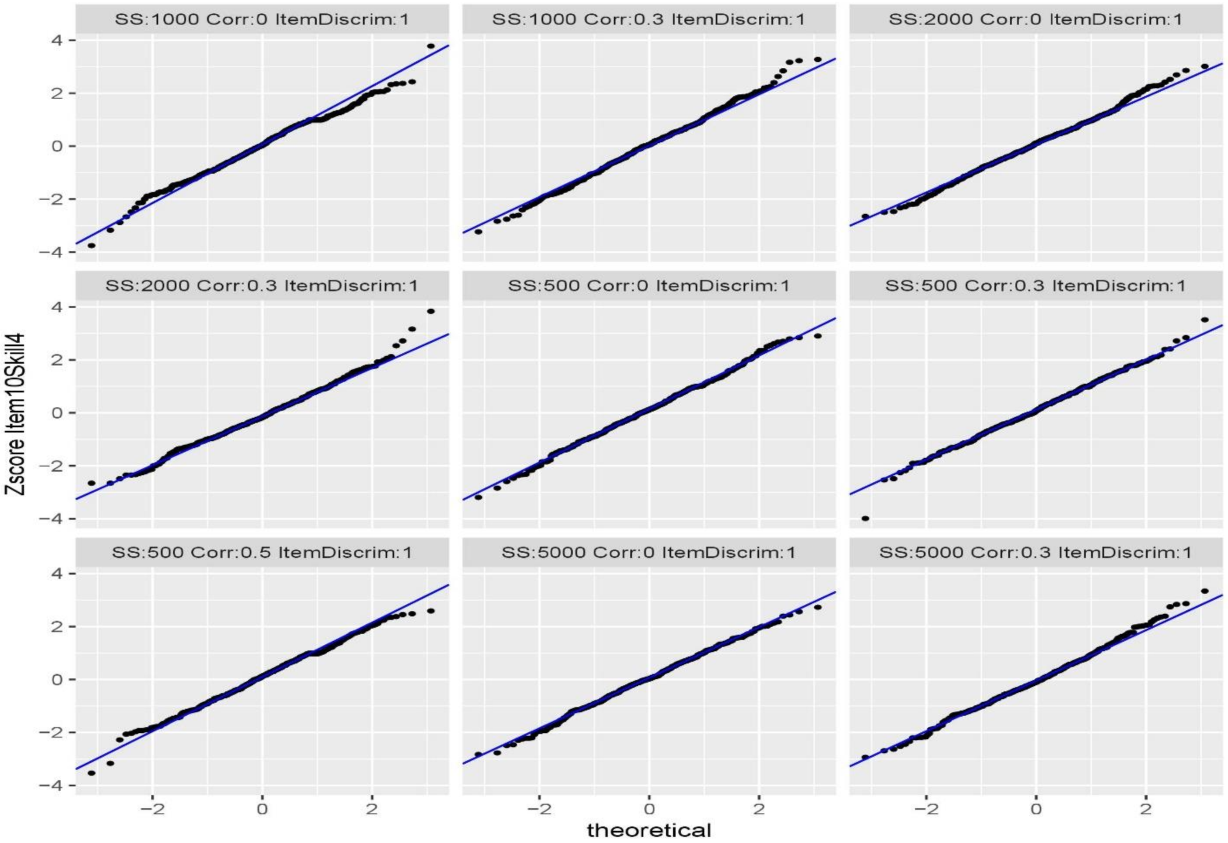


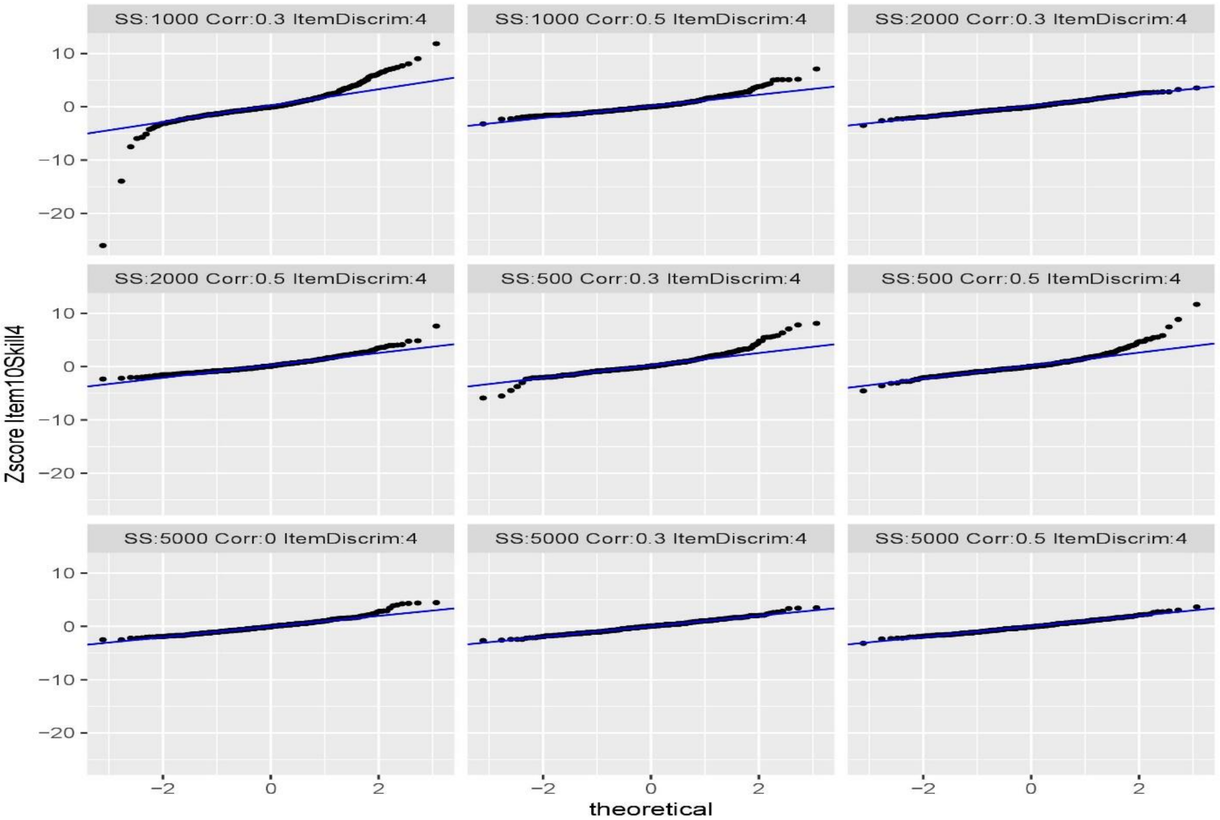

SS:1000 Corr:0 ItemDiscrim:1
SS:1000 Corr:0.3 ItemDiscrim:1
SS:2000 Corr:0 ItemDiscrim:1
SS:2000 Corr:0.3 ItemDiscrim:1
SS:500 Corr:0 ItemDiscrim:1
SS:500 Corr:0.3 ItemDiscrim:1
SS:500 Corr:0.5 ItemDiscrim:1
SS:5000 Corr:0 ItemDiscrim:1
SS:5000 Corr:0.3 ItemDiscrim:1
20
10
0
−2
0
2
Zscore Item9Skill3
theoretical


SS:1000 Corr:0.3 ItemDiscrim:4
SS:1000 Corr:0.5 ItemDiscrim:4
SS:2000 Corr:0.3 ItemDiscrim:4
SS:2000 Corr:0.5 ItemDiscrim:4
SS:500 Corr:0.3 ItemDiscrim:4
SS:500 Corr:0.5 ItemDiscrim:4
SS:5000 Corr:0 ItemDiscrim:4
SS:5000 Corr:0.3 ItemDiscrim:4
SS:5000 Corr:0.5 ItemDiscrim:4
5
0
−5
−2
0
2
Zscore Item9Skill3
theoretical

Q-Q Plots by Sample Size, Correlation and Item Discrimination for Item 10 Skill 4




SS:1000 Corr:0 ItemDiscrim:1
SS:1000 Corr:0.3 ItemDiscrim:1
SS:2000 Corr:0 ItemDiscrim:1
SS:2000 Corr:0.3 ItemDiscrim:1
SS:500 Corr:0 ItemDiscrim:1
SS:500 Corr:0.3 ItemDiscrim:1
SS:500 Corr:0.5 ItemDiscrim:1
SS:5000 Corr:0 ItemDiscrim:1
SS:5000 Corr:0.3 ItemDiscrim:1
Zscore Item10Skill4
4
2
0
−2
−4
−2
0
2
theoretical


SS:1000 Corr:0.3 ItemDiscrim:4
SS:1000 Corr:0.5 ItemDiscrim:4
SS:2000 Corr:0.3 ItemDiscrim:4
SS:2000 Corr:0.5 ItemDiscrim:4
SS:500 Corr:0.3 ItemDiscrim:4
SS:500 Corr:0.5 ItemDiscrim:4
SS:5000 Corr:0 ItemDiscrim:4
SS:5000 Corr:0.3 ItemDiscrim:4
SS:5000 Corr:0.5 ItemDiscrim:4
Zscore Item10Skill4
10
0
−10
−20
−2
0
2
theoretical