2019-12-11

Recommendations Regarding Q-Matrix Design and Missing Data Treatment in the Main Effect Log-Linear Cognitive Diagnosis Model

Rui Ma
Brigham Young University

Recommendations Regarding Q-Matrix Design and Missing Data Treatment

in the Main Effect Log-Linear Cognitive Diagnosis Model

Rui Ma

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Ross Allen Andrew Larsen, Chair
Laine Percell Bradshaw
Troy L. Cox
Joseph A. Olsen
Richard R. Sudweeks

Educational Inquiry, Measurement, and Evaluation

Brigham Young University

ABSTRACT

Recommendations Regarding Q-Matrix Design and Missing Data Treatment in the Main Effect Log-Linear Cognitive Diagnosis Model

Rui Ma
Educational Inquiry, Measurement, and Evaluation, BYU
Doctor of Philosophy

Diagnostic classification models used in conjunction with diagnostic assessments are to classify individual respondents into masters and nonmasters at the level of attributes. Previous researchers (Madison & Bradshaw, 2015) recommended items on the assessment should measure all patterns of attribute combinations to ensure classification accuracy, but in practice, certain attributes may not be measured by themselves. Moreover, the model estimation requires large sample size, but in reality, there could be unanswered items in the data. Therefore, the current study sought to provide suggestions on selecting between two alternative Q-matrix designs when an attribute cannot be measured in isolation and when using maximum likelihood estimation in the presence of missing responses. The factorial ANOVA results of this simulation study indicate that adding items measuring some attributes instead of all attributes is more optimal and that other missing data treatments should be sought if the percent of missing responses is greater than 5%.

Keywords: diagnostic classification model, log-linear cognitive diagnostic model, Q-matrix, missing data, classification accuracy, attribute reliability

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Diagnostic classification models (DCMs) are statistical models used to classify respondents into one of two categories (mastery vs. nonmastery) for each unique attribute. In educational settings, their potential use is to give feedback to individual examinees on their strengths and weaknesses at the level of measured subskills or attributes.

Applications of DCMs have been concentrated in using assessments that were developed for purposes other than diagnosis, such as the Test of English as a Foreign Language (TOEFL; von Davier, 2005) and the Michigan English Language Assessment Battery (MELAB; Li & Suen, 2013). This procedure is a reversed process insomuch that the Q-matrix, which specifies which attributes should be measured by which items, is constructed after the questions have been developed for another purpose. Nevertheless, this is a common practice as (a) the underlying cognitive process or strategies for specific tasks may not be clearly supported by theory; (b) development of assessments for diagnostics is a demanding process, as described in Bradshaw (2017); and (c) the large sample size requirement (Kunina-Habenicht, Rupp, & Wilhelm, 2012) for model estimation is hard to obtain, while using already developed standardized tests provides a convenient large sample size.

As the first issue relates to specific disciplines and should be discussed case by case, this study intends to make informed recommendations for the last two of the three issues mentioned above: assessment development and model estimation regards to sample size.

Although research findings from standard testing offer insight, in order to fully realize the power of DCMs, diagnostic assessments have to be used. A diagnostic assessment usually takes place at the beginning of a course or instructional treatment. The assessment results are then used to inform the respondents about their strengths and weaknesses and where they need to focus their study efforts. To optimize the classification accuracy, researchers have made recommendations on the Q-matrix design (Madison & Bradshaw, 2015). However, these recommendations may be impossible to follow in practice, especially with attributes that cannot be measured in isolation. For example, in solving linear algebra, removing parentheses is almost always followed by combining like terms.

The other issue in practice is the presence of unanswered items, also referred to as missing responses. The patterns of these missing responses can be missing completely at random (MCAR), missing at random (MAR), missing not at random (MNAR), or any combinations of them. It is expected that respondents are not all masters of all measured attributes. Given the nature of diagnostic assessments, these tests are also supposed to be low-stakes. As a result, as students skip questions which they consider to be too hard and not worth the effort, responses that are missing may occur. In order to meet the requirement of large sample size, researchers always need to make decisions regarding these missing responses rather than deleting the entire case containing missing responses. In fact, one of the obstacles of DCM application is classification accuracy in the presence of missing responses (Rupp & Templin, 2008). A handful of published DCM studies have reported how missing responses were handled (Ayers, Nugent, & Dean, 2009; Gu, 2012; Hansen, 2013; Harrison, Bradshaw, Naqvi, Paff, & Campbell, 2017; Lee, Park, & Taylan, 2011; Sheehan, Tatsuoka, & Lewis, 1993; Skaggs, Wilkins, & Hein, 2016; Templin & Hoffman, 2013; Xin & Zhang, 2015). Among them is the demonstration of using Mplus to estimate the log-linear cognitive diagnostic model (LCDM) by Templin and Hoffman (2013), and the authors suggested letting missing responses be handled automatically by the maximum likelihood estimation. While their demonstration serves as a guide for practitioners, how well maximum likelihood estimation handles missing responses in terms of classification accuracy and attribute reliability and when practitioners should seek out advanced missing data treatments described in the above studies remain unclear.

Statement of the Purpose

To help future researchers and practitioners make informed decisions about designing Q-matrices with attributes that cannot be measured in isolation, and what to do about missing responses, the current simulation study investigates how well the full information maximum likelihood (FIML) estimation procedure in Mplus estimates the main effect LCDM in the presence of missing responses across several Q-matrix designs and with varying sample sizes in terms of classification accuracy and attribute reliability.

Research Questions

The research questions were as follows:

1. Under the condition of a balanced Q-matrix design, what is the effect of different percentages of missing responses (5%, 10%, 15%, 20%, and 30%) on attribute classification accuracy, profile classification accuracy, and attribute reliability across sample sizes (500; 1,000; and 2,000) in all missing data mechanism conditions (MCAR, MAR, and MNAR), compared with complete data?

2. When an attribute cannot be measured in isolation (unbalanced Q-matrix), what is the effect of different percentages of missing responses (5%, 10%, 15%, 20%, and 30%) on attribute classification accuracy, profile classification accuracy, and attribute reliability across sample sizes (500; 1,000; and 2,000) in all missing data mechanism conditions (MCAR, MAR, and MNAR), compared with complete data?

3. Between the two alternative Q-matrix designs, which one obtains higher attribute classification accuracy, profile classification accuracy, and attribute reliability?

CHAPTER 2

Review of Literature

In this section, I briefly introduce different diagnostic classification models (DCMs) and the role of the Q-matrix design in DCM applications. The common phenomenon of missing responses in educational research, missing data mechanisms, and common missing data treatments (MDTs) will also be discussed. In the end, I will present missing response studies in the DCM context.

Diagnostic Classification Models

Diagnostic classification models (DCMs), also known as cognitive diagnostic models (CDMs; Henson & Douglas, 2005), are a series of statistical models that assume that mastery of certain skills or attributes contribute to a correct response to an item. Some DCMs are compensatory (or disjunctive) models, and some are noncompensatory (or conjunctive) models (Rupp & Templin, 2008). An assumption associated with the use of compensatory models is that not all attributes mapped onto the item have to be present to reach a correct answer. That is, the lack of some attributes can be compensated by mastering other attributes. The assumption of noncompensatory models is that all attributes measured by the item have to be mastered for producing the correct answer. A correct answer under this assumption is produced only by examinees who master all attributes other than guessing, while a correct answer under the compensatory assumption can be produced by examinees with mastery of different combinations of the attributes. Although hierarchical relations among attributes based on substantive theory have been integrated into some models, these studies are rarer, so these models are not studied here.

Several DCMs have appeared in the literature. Commonly studied general models include (a) the generalized deterministic input noisy *and* gate (G-DINA) model (de la Torre, 2011), (b) the general diagnostic model (GDM; von Davier, 2005), and (c) the log-linear cognitive diagnostic model (LCDM; Henson, Templin, & Willse, 2009). These general models can be constrained to either noncompensatory models or compensatory models, which rely on more restricted assumptions. Common dichotomous noncompensatory models are the deterministic input noisy *and* gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model and the reduced reparameterized unified model (R-RUM; Chiu & Köhn, 2016; Henson, Templin, & Douglas, 2007); common dichotomous compensatory models are the deterministic input noisy o*r* gate (DINO; Templin & Henson, 2006) model and compensatory reparameterized unified model (C-RUM; Hartz, 2002). Each of these models will be discussed subsequently.

General models. The G-DINA model is a general model which can be constrained to several reduced models. For every item, it assumes the existence of a baseline probability, which is the probability of a correct response when there is no mastery of any of the required attributes. It further allows a different additive effect of mastering an additional attribute, of mastering different level of interaction effect, and of mastering all the required attributes.

The G-DINA model with dichotomously scored items uses an identity link function, according to de la Torre (2011):

$$\mathrm{P}(X_j{=}1|\alpha_{lj}^*) = \delta_{j0}+ \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k\prime=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk\prime}\alpha_{lk}\,\alpha_{lk\prime} \ldots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \quad (1)$$

Where

X_j=1 is the correct response on item j;

α_{lj}^* is a vector whose elements indicate actual measured attributes by item j;

δ_{j0} is the intercept for item j;

K_j^* is the number of attributes actually measured by item j;

δ_{jk} is the main effect coefficient due to α_k;

$\delta_{jkk'}$ is the interaction effect coefficient due to $\alpha_k$and $\alpha_{k'}$; and

$\delta_{j12\ldots K_j^*}$ is the interaction effect coefficient due to α_1,..., $\alpha_{K_j^*}$.

In the G-DINA model, δ_{j0} and δ_{jk} are usually nonnegative, but the interaction effect coefficients can take on any values. It is also reasonable to impose a monotonicity constraint where the mastery of increased number of attributes does not associate with a decrease in correct response probability.

General diagnostic models (GDMs; von Davier, 2005) is another general model framework. Models in that category are based on extensions of latent class models, the Rasch model, item response theory models, and skill profile models. The class of GDMs allows arbitrary attribute mastery levels and levels of an attribute required for the item. The log ratio of the probability of getting a certain score to the probability of producing a completely wrong answer for an item is determined by the item difficulty of obtaining the score and a linear combination of attribute discriminations of each attribute measured by the item. de la Torre (2011) stated that with dichotomous items, GDM uses a log link function instead of an identity function of the G-DINA model. GDM not only encompasses compensatory as well as noncompensatory models but also allows partial credit attribute entries in the Q-matrix and polytomously scored items. However, GDM is the least studied general CDM, and its link to other reduced models is unclear (Henson et al., 2009).

On the contrary, a special case of GDM, the LCDM (de la Torre, 2011; Henson et al., 2009; Jurich & Bradshaw, 2014), is more commonly studied than the GDM. Similar with item response modeling, the LCDM uses a logit link function, as introduced by de la Torre (2011):

$$\operatorname{logit}[P(X_j = 1 \mid \alpha^*_{lj})] = \lambda_{j0} + \sum_{k=1}^{K_j^*} \lambda_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \lambda_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \lambda_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}, \qquad (2)$$

The LCDM assumes both responses and latent attributes are dichotomous rather than allowing them to be polytomous as the GDM does. That is, respondents are assumed to either possess an attribute or not, and items are either correct or incorrect. With an attribute-item mapping called a Q-matrix (discussed in the next section), and with a reference group consisting of respondents who have not mastered any of the attributes, the LCDM needs an additional constraint, which is the probability of a correct answer is equal or greater with additional mastered attributes (monotonicity). The logit of the probability of answering an item correct is expressed as the mastery of the number the attributes measured by the item compared with the reference group consisting of respondents who have mastered none of the attributes. Because of this simple interpretation, the LCDM is the focus of this article, and the following will introduce four common reduced models and how the LCDM can be constrained to DINA, R-RUM, DINO, and C-RUM (Henson et al., 2009).

The deterministic input noisy *and* gate model. The DINA model is a simple noncompensatory model that only estimates two parameters per item (Henson et al., 2009; Roussos, Templin, & Henson, 2007), assuming the probability of a correct answer for an item depends on how easy it is to guess and how easy it is to carelessly produce a wrong response. One parameter is called *guessing*; it estimates the probability of producing the correct answer without mastering all the required attributes. The other parameter is *slipping;* it estimates the probability of not producing the correct answer with all the required attributes mastered. The DINA model assumes the probability of answering the item correctly having mastered all required attributes to be higher than the probability of guessing the correct answer without mastering all the required attributes (Henson et al., 2009). The LCDM can be constrained to the DINA model by setting the discrimination of the interaction term of all required attributes positive and setting other discrimination parameters to 0, and then by making a transformation of the parameters (Henson et al., 2009).

The reduced reparameterized unified model. The R-RUM is a reduced version of the reparameterized unified model (RUM; Hartz, 2002). The RUM is also called the Fusion model (Hartz & Roussos, 2008; Li, Hunter, & Lei, 2016). With both the R-RUM and the Fusion model, the probability of a correct response to an item decreases with an increased number of unmastered attributes (Henson et al., 2009). The Fusion model and R-RUM have been commonly used in analyzing data from language assessments. For example, Kim (2015) applied the R-RUM to an English as a Second Language (ESL) placement test to provide feedback, Jang, Dunlop, Wagner, Kim, and Gu (2013) used the R-RUM to analyze an elementary level reading achievement test to classify Grade 6 students' reading skill mastery, and Jang (2009a) used the Fusion model on data collected from a second language reading comprehension test and examined the use of diagnostic feedback. The Fusion model assumes that the Q-matrix does not cover all attributes involved to reach a correct answer for the item and that the model involves a continuous residual parameter indicating any skill unspecified by the Q-matrix (Hartz & Roussos, 2008). However, Roussos, DiBello, Henson, Jang, and Templin (as cited in Jang, 2009a) indicated, this continuous parameter "soaks up" most of the variance in the item response if a test has a single dominant dimension, so the R-RUM which does not have this parameter is usually used to analyzed real data (Henson et al., 2007; Jang, 2009b; Rupp, Templin, & Henson, 2010). The R-RUM can be constrained from LCDM by redefining parameters and mathematical transformations (Chiu & Köhn, 2016; Henson et al., 2009).

The deterministic input noisy *or* gate model. In addition to the two commonly used noncompensatory models, the DINO model is a simple compensatory model. Similar to the DINA model, the DINO model also estimates the guessing and slipping parameters. However, the DINO model does not require mastery of all related attributes, but rather it treats the attributes as equivalent alternatives. The assumption is that the probability of answering the item correctly having mastered at least one attribute is higher than the probability of guessing the correct answer without mastering any of the attributes (Henson et al., 2009). Therefore, the probability of a correct response per item is modeled to be the same for examinees who master any number of the attributes. To derive the DINO model from the LCDM, all the absolute values of the main effects and interaction effects are constrained to be the same and the only one of these effects contributes to the model when there is mastery of more than one of the measured attributes, following which parameter transformation needs to be performed (Köhn & Chiu, 2016).

The compensatory reparameterized unified model. The last compensatory model introduced here is the C-RUM model. The C-RUM is the compensatory counterpart of R-RUM. Similar with the R-RUM, the C-RUM does not have a latent continuous parameter to account for any ability not specified by the Q-matrix (Yi, 2012). For each item, the log odds of the probability of a correct answer is estimated by an intercept parameter at the item level and unique contributions of mastery of each measured attribute by the Q-matrix (Yi, 2012). That is, if an item measures more than one attribute, there is no interaction effect among the attribute in the model, and mastering of one attribute is assumed to be independent from mastering other attribute (Henson et al., 2009). Therefore, the C-RUM can be obtained by setting the interaction effects from the LCDM to be 0 (Henson et al., 2009; Rupp et al., 2010).

There are several reasons why the C-RUM is preferable to other models. First, the C-RUM statistical properties resemble the ones of the LCDM, a general model. Yi (2012) fitted the LCDM, C-RUM, DINA, DINO, and NIDO to data collected from a large-scale language assessment and found that the LCDM produced the best fit based on RMSEA and relative fit indices (AIC, BIC, and sample size adjusted BIC) and the differences between the C-RUM and the LCDM in the indices were small. In terms of profile classifications and individual profiles, the C-RUM was more similar than other models with the LCDM. Similarly, Jurish and Bradshaw (2014) found that higher order interaction effects did not produce added benefit based on the information criteria and the likelihood ratio test when they were choosing among a series of LCDM and reduced LCDMs. The ability of producing similar results to the general model is evident that the C-RUM has the potential to be used in practical situations. Second, the C-RUM does not have as high demand on sample size (Kunina-Habenicht et al., 2012). Kunina-Habenicht et al. (2012) found that the intercept and main effect parameters of the LCDM were reasonably on target with sample sizes of 1,000 and 10,000, and that even with the large sample size, some interaction parameter estimates were noticeably imprecise. They further observed that in terms of latent class distributions and identical classification rates, the estimation results of the true model and the correctly specified model with main effects only were almost identical. They then stated that "the correct specification of interaction effect parameters does not seem necessary for many practical situations" (p.77). Following that suggestion, Kunina-Habenicht, Rupp, and Wilhelm (2017) used the LCDM with main effects only (C-RUM) without implementing the full LCDM to investigate the reliability and incremental validity of DCM scores. Having observed empirical DCM studies, I found that the sample size was as low as 138 in Sorrel et al. (2016) in an effort to demonstrate using DCMs to score situational judgement tests and the largest sample size was 10,000 from Ravand (2016) demonstrating using the G-DINA model with the reading comprehension section of the Iranian National University Entrance Examination. In other words, the demand of large sample size for precise estimation of interaction effect parameters is hard to meet in practice. Third, the C-RUM parameters are meaningful, as are other DCM models, at item level among compensatory models. The intercept parameter can be interpreted as the logit of the probability of producing a correct answer to an item if none of the measured attributes is mastered, and each slope parameter can be interpreted as the increase in logit when the corresponding attribute is mastered. Although other DCMs have their own advantages, this study focuses only on the C-RUM.

Although there are DCMs developed for modeling attribute hierarchy (Templin & Bradshaw, 2014) and polytomously scored responses (von Davier, 2005), this article focuses only on modeling unstructured attribute relations and dichotomously scored responses.

Q-matrix Design

Aside from compensatory and noncompensatory contributions of the attributes to the correctness of the item, the attribute-item matching within an assessment is another consideration for DCM applications. This matching is specified in an incidence matrix called a Q-matrix (Tatsuoka, 1990). A Q-matrix is a two-dimensional matrix with 0s and 1s as the cell entries. Its common layout includes a row representing each item, and a column representing each attribute. If an item measures an attribute, the corresponding cell entry is 1, and otherwise 0. The structure of the Q-matrix mainly reflects the number of attributes, the number of items, and combinations of attributes measured by each item. Table 1 is an example of a Q-matrix, outlining a three-item test assessing four attributes: (a) division of integers; (b) addition/subtraction; (c) removing parentheses; and (d) combining like terms. Item1 measures addition/subtraction; item 2 measures division of integers, addition/subtraction, and combining like terms; and item 3 measures all four attributes. In this design, the third attribute, removing parentheses, is only measured once and is only measured in the presence of all other attributes. Ideally, the Q-matrix is developed as part of the test specification prior to item development. The complexity of the Q-matrix is measured by the number of attributes per item (Madison & Bradshaw, 2015). The more complex the Q-matrix is, the higher the number of attributes per item.

Table 1

Q-matrix Example

Item No.	Item	Division of Integer	Addition/Subtraction	Removing Parentheses	Combining Like Terms
1	x-3=5	0	1	0	0
2	4x+1=6x-7	1	1	0	1
3	5(x-4)-x=12	1	1	1	1

Note. This Q-matrix is only for illustration. A three-item test is not long enough to provide diagnostic information.

Several studies have found that the structure of the Q-matrix influences classification accuracy. Under the assumption that the Q-matrix is specified correctly, DeCarlo (2011) stated that the design of the Q-matrix can influence classification accuracy in the context of the DINA model and higher order models. Similarly, in the LCDM context, Kunina-Habenicht et al. (2012) found that the classification accuracy was higher in conditions with three attributes and 50 items than with five attributes and 25 items, indicating that few attributes and more items are associated with higher classification accuracy. In empirical studies for diagnostic assessment development, the number of attributes was around three to four (Harrison et al., 2017; Templin &

Hoffman, 2013), and the number of items had a larger range. For example, the instrument of Jurich and Bradshaw (2014) contained 17 items, while Harrison et al. (2017) had 52 items in their final instrument.

Concerned with achieving classification accuracy, researchers have conducted simulation studies to come up with guidelines for the Q-matrix design. Chiu, Douglas, and Li (2009) suggested that each attribute be measured in isolation by at least one item in DINA and DINO. Madison and Bradshaw (2015) argued, however, that general DCMs do not have such requirement for classification accuracy because each attribute is modeled individually but may have other requirements for model identification. Having conducted a simulation study in the context of the LCDM, Madison and Bradshaw recommended each attribute should be measured at least once in isolation if it could be isolated, and that no two attributes should always be measured together. The way attribute 3 (removing parentheses) is measured in Table 1 violates this guideline because it is measured only once and only measured in the presence of other attributes. Madison and Bradshaw also indicated that several factors have the potential to impact classification accuracy and model interpretation including the discrimination power of the items, sample size, length of the assessment, number of attributes, correlations among attributes, and interaction among attributes. Bradshaw (2017) further recommended a relatively balanced design in which a variety of attribute combinations are specified so that no identical response patterns can be produced by respondents from different classes of classification. Ayers et al. (2009) offered a more explicit definition of a balanced Q-matrix design: “all single skill items occur the same number of times, and each combination of skills occurs the same number of times” (p. 2). The Q-matrix in Table 1 is an unbalanced design because not all combinations of attributes appeared.

Empirical studies so far, however, rarely follow these guidelines to design the Q-matrix and then write items following the Q-matrix specification. For example, Bradshaw, Izsák, Templin, and Jacobson (2014) developed a diagnostic test of assessing teachers' understandings of rational numbers from identifying attributes, writing items, conducting task analysis, to refining the items. The final product contained four attributes and 27 items, including 19 items measuring a single attribute and eight items measuring two of the four attributes. Similarly, in the Autism Stigma and Knowledge Questionnaire developed by Harrison et al. (2017), there were 52 items and four attributes in total, and 46 of the items measured a single attribute and seven items measuring the 4th attribute paired with another attribute. This pattern of single-attribute items combined with pair-attribute items is also found in Kunina-Habenicht, Rupp, and Wilhelm (2009). There were 87 items in the arithmetic skills test measuring four attributes, and only 22 items measured a combination of two attributes in Kunina-Habenicht et al. (2009). Examples of single-attribute-item-instrument can also be found. For example, each of the 17 items of the student learning outcomes assessment in Jurich and Bradshaw (2014) measured one of the four attributes.

These practices not only stifle DCMs' potential of accurate classification with short multidimensional assessments but also greatly increase the possibility of introducing irrelevant attributes when developing test items without the Q-matrix specification. Q-matrices are typically developed after the responses have been collected in empirical studies. To meet the large sample size requirement for model estimation, researchers tend to analyze responses obtained from large-scale assessments which are created for the purpose of placement or selection rather than diagnosis. When these tests are created, what is expected of the respondents' holistic ability is what leads the test writing process. Test writers do not have to consider the strategies or cognitive processes needed to accomplish each task on the test. For DCM purposes, each item is assigned attributes after the test has been developed, leaving the Q-matrix as a description of the test instead of a specification of the test. While the assignment could be accurate, items could measure attributes other than specified in the descriptive Q-matrix, and the structure of the Q-matrix, such as the number of times an attribute is measured is out of the control of the researcher.

While the above approach is less optimal, there are barriers to item writing in accordance to the pre-specified Q-matrix. More specifically, there could be attributes that cannot be measured in isolation and are always measured in conjunction with another attribute. Take the Q-matrix in Table 1 as an example. Attribute 3, removing parentheses, is almost unavoidably measured with attribute 4, combining like terms because after distributing, like terms are likely to appear. If there are no like terms after distributing, such as 4(x+1) = 20, distributing may not even be a step in solving the problem because dividing both sides of the equation by 4 could remove the parentheses. Another example can be found in reading assessments. Inferencing from context as a reading skill may not easily isolated from other reading skills, such as understanding the main idea and identifying logic of arguments, for a specific reading text. As a result, a balanced Q-matrix design may be implausible in practice.

Missing Response Issue

Researchers and practitioners attempting to use DCMs for diagnostic purposes face the difficulties of both designing a practical, statistically-satisfying Q-matrix and meeting the sample size requirement. In regard to retaining a large sample size, an additional, practical decision they have to make regards how to deal with missing responses. Given that a diagnostic test is difficult for respondents and is typically viewed by them as being as a low-stakes testing situation, they are likely omit responding to some questions. Discarding cases with missing responses defeats the point of having a large sample size. Large sample size is needed for DCM model estimation because of the large number of possible response patterns. Take a 20-item test with three attributes for example. If it is a dichotomously scored test, the number of response patterns is 2^{20}(=1048576), so many the response patterns are not observed. Cases with missing responses could provide some information for model estimation. Rupp and Templin (2008) even identified that classification accuracy in the presence of missing responses needed more informative research before DCMs could be applied.

Missing Data Mechanisms

Although the term missing response accurately describes the nonresponding phenomenon in educational settings, in the literature, the term *missing data* is used. Even though missing data have existed as long as the field of measurement, there had not been systematic studies before Rubin (1976) identified three missing data mechanisms. Missing data mechanism or missingness is a widely cited term "to describe the rates and patterns of missing values and to capture roughly possible relationships" (Schafer & Graham, 2002, p.150) between the missing pattern and the missing value themselves. The missing data mechanisms are also referred to as three types of missing data:

1. Missing completely at random (MCAR).
2. Missing at random (MAR).
3. Missing not at random (MNAR).

MCAR indicates that the values of missing data and the patterns of missing data do not consistently relate to the values of other observed or unobserved data. For example, in educational settings, missing data produced by design, in which certain items are administered to a randomly selected sample and other items are administered to another randomly selected sample, is usually considered as MCAR because the could-have-been responses to the unseen items are not associated with other items or other latent traits of the participant.

The second missing data mechanism is MAR, which refers to the phenomenon that the missingness is related to the values of observed data but unrelated to the values of missing data themselves (Rubin, 1976). For example, if a questionnaire contains a question asking for a response conditioned upon the answer to the previous question, the missingness of the response to the second question is related to the response to the first one. Causes of MAR can be attributed to the design of data collection in which certain data are not collected based on the values of observed data, but sometimes MAR is assumed because the values of the missing data are not known (Schafer & Graham, 2002).

If the missingness relates to the values of the missing variables, the type of missingness is MNAR, the third missing data mechanism. This could be the case where missingness is related to unobserved traits. For example, if missing academic records are of students who could have obtained low scores if they had taken the assessment, it will be MNAR. In some cases, plausible MAR is actually MNAR, depending on the relation among variables. Again, these are assumptions and are almost impossible to verify.

All three missing data mechanisms are possible in educational settings (Finch, 2008). For example, if test takers who have a low score on a certain item also tend to miss another item, it could be that the test takers skip the second item because it is as difficult as the first item, but it could also be because the test takers arrive at an impossible answer, and they choose to skip it. In the former scenario, the test takers could arrive at the correct answer, so the missingness would be classified as MAR. The latter scenario would be classified as MNAR in which the missingness relates to the values of missing data. If the test taker skips items inadvertently (Finch, 2008), the missingness would be classified as MCAR.

Common Missing Data Treatments

Common missing data treatments (MDTs) can be classified into three categories: (a) deletion methods, (b) substitution or imputation methods, and (c) direct parameter estimation without imputation (maximum likelihood estimation). In this section, I will give conceptual explanations of these MDTs and present an overview of MDTs in educational research. Table 2 lists some deletion methods and imputation methods that commonly appear in the literature. Although each might have demonstrated its use, those methods do not adequately account for the noise or variability of missing values, and for this reason. Rubin developed multiple imputation (MI) in 1978 (Sinharay, Stern, & Russell, 2001) to solve this problem. The advantage of MI is that "the inferences—standard errors, p-values, etc.—obtained from MI are generally valid because they incorporate uncertainty due to missing data" (Schafer & Olsen, 1998, p. 548). The assumptions of MI are that (a) the imputation model is "compatible with the analyses to be performed on the imputed datasets" (Schafer & Olsen, 1998, p. 550) and reflects variable associations in the subsequent analyses, that (b) the prior distribution is reasonable, and that (c) the missingness is MCAR. In the MI process, each missing value is imputed with a plausible value based on the observed data. Once all the missing values are imputed, an imputed data set with no missing values is produced. This procedure is conducted m ($m > 1$) times with the plausible values having a residual distribution to account for inherent uncertainty in sampling distribution. MI produces m datasets, and then each data set is analyzed individually to produce m results. The results are then combined according to Rubin's rules to arrive at the final result. Although previous guideline suggested that $m = 10$ was sufficient, further study by Bodner

(2008) suggested that more was needed. Several algorithms other than using the residual distribution have been used to produce each individual imputed data set in MI, including the expectation-maximization (EM) algorithm and the data augmentation (DA) algorithm, which will be introduced below.

Maximum likelihood estimation. When missing data are present, maximum likelihood (ML) estimation can be used for parameter estimation so that population parameter estimates are the most probable values to produce the sample data (Baraldi & Enders, 2010). The underlying assumptions are multivariate normality (Enders, 2001) and MAR or MCAR. Nevertheless, Schafer and Olsen (1998) pointed out that dichotomous variables could also be imputed with ML and that rounding of the results to zero or one was common. However, when missing data is present, the robustness of ML chi-square no longer holds, even though it may be robust enough when the multivariate normality assumption is violated for complete data (Savalei, 2008).

Enders (2001) introduces full information maximum likelihood (FIML) as one of the maximum likelihood algorithms. FIML directly estimates parameters without imputing missing values (Enders, 2001). As an ML algorithm, the assumptions of FIML are also that the values of variables involved achieve multivariate normality and that missing values have the probability of being inferred from the available data (the missing data mechanism is MAR or MCAR). In other words, the second assumption is met for FIML only "if the 'cause' of missingness is included in the analysis model" (Enders, 2008, p. 436). It is conceptually similar to pairwise deletion (PD) in that it utilizes information from all the observed data including cases where missing values exist, but mathematically it is unrelated to PD (Enders, 2001). Mplus uses FIML estimation to deal with missing values by default (Bowen, 2015).

Table 2

Summary of Missing Data Treatments

MDT	Meaning
	Deletion method
Listwise deletion (LD)	Only retaining complete cases for analysis
Pairwise deletion (PD)	Disregarding missing values "based on an analysis-by-analysis basis" (Baraldi & Enders, 2010, p. 10)
	Imputation method
Zero imputation (IN)	Treating missing values as incorrect
Mean imputation	Filling in missing values with the mean of the person, of the variable, or of the data set
Corrected item mean imputation (CM)	Filling in missing values with weighted variable mean (Huisman & Molenaar, 2001)
Two-way imputation (TW)	Filling in missing values with the sum of observed case mean and observed variable mean subtracted by the overall mean of all the observed values in the data set (van Ginkel, van der Ark, & Sijtsma, 2007)
Response-function imputation (RF)	For dichotomously scored test, the first step is to estimate a person's score without missing items by multiplying the mean score with the number of observed items minus one. Then, estimate the probability of having the estimated score based on the test. The last step is to draw a number from the Bernoulli distribution with that probability, and this value is used as imputed value. (Sijtsma & van der Ark, 2003)
Regression-based imputation	Filling in missing values with predicted values from observed values
Stochastic regression imputation	Incorporating an error term in the regression equation in the imputation process
Hotdeck imputation	Filling in missing values with the values of similar cases based on observed variables

The Expectation-Maximization (EM) algorithm. Enders (2001) also introduced the EM algorithm as one of the ML algorithms. The EM algorithm is used to produce imputed data sets or estimates of a mean vector and covariance matrix (Enders, 2001). It also assumes that multivariate normality exists and that the observed values contain information of the missing values (Enders, 2001). Enders conceptually introduced the procedure: The EM algorithm is an iterative process cycling through two steps—the E step and the M step. In the E step, the missing values are filled in with expected values based on the observed data and the estimated covariance matrix; in the M step, the mean vector and covariance matrix are estimated based on the imputed complete data. The two steps are repeated until the difference of the result from the current iteration and the previous one meets certain convergence criterion. One criticism of the EM algorithm approach is that it lacks residual variability, but the bootstrap technique and adding a correction factor can be used to mitigate the drawback (Enders, 2001). The EM algorithm is also recommended as a tool for MI because of excellent starting values of parameter estimation and of the predicting ability of the convergence behavior (Schafer & Olsen, 1998).

Data augmentation (DA). Schafer and Olsen (1998) introduced DA as a tool for MI. Similar to EM, DA is an iterative process, but instead of updating the parameter estimates based on tentative estimates from the previous cycle, the parameter estimates are drawn from a Bayesian posterior distribution based on the imputed data set (Little & Rubin, 2002). Convergence is reached when the parameter distribution is stable. Running the DA algorithm *m* times will produce *m* imputed data sets composing of the result of MI.

The above MDTs are by no means comprehensive. There are specific MDTs created for specific use, such as the technique developed by Lang and Little (2014) for hypothesis testing with incomplete data, and the method developed by Song and Lee (2006) for analyzing a multisample nonlinear structural equation model. Because this section only serves as a brief introduction to common MDTs, and the MDT research in the DCM context is at the initial stage, other more specific MDTs are not included here.

Overview. Educational researchers rarely report the percentage of missing responses and how they were handled. Peugh and Enders (2004) conducted a methodological review of reporting MDTs in 16 educational research journals published in the year 1999 and 23 published in 2003 respectively. They found that among the studies published in 1999, at least 16% of the studies had missing data and the percentage ranged from less than 1% to 67%. All these studies used traditional missing-data techniques, including LD, PD, mean imputation, and regression imputation, with the majority of the studies adopting LD, PD, or a combination of the two. On the other hand, the percentage of studies in 2003 had detectable missing data increased to 42%, but the MDT usage pattern did not change much. More recently, Pichette, Béland, Jolani, and Leśniewska (2015) surveyed language researchers and found that the most common methods to treat missing data of binary outcome were LD, leaving them to the software, PM, and IN. Vague reporting of the presence of missing data and MDTs and using traditional methods also prevail in the more specific DCM context.

Missing Data Treatments in Diagnostic Classification Model Context

Similar to the findings of Peugh and Enders (2004) and of Pichette et al. (2015), a thorough search in the literature indicated that researchers in DCM contexts rarely report the percentage of missing data or how they were handled in their research. Eight studies acknowledged the existence of missing responses, and some of them reported the missing data treatments (MDTs) (Table 3). As shown in Table 3, the most common approach for treating missing responses was to treat missing responses as incorrect responses (Hansen, 2013; Harrison et al., 2017; Lee et al., 2011). The other approach was to leave the missing responses to the estimating software. For example, Templin and Hoffman (2013) mentioned that Mplus could handle missing responses, so they did not take additional action to handle them. Even though Gu (2012) implemented a unique treatment—assigning the item option least selected to the missing value, the approach is only feasible for polytomously coded multiple-choice items and is beyond the scope of this study.

Two of the remaining studies (Sheehan et al., 1993; Skaggs et al., 2016) dealt with missing responses from unadministered items, and neither of them included these items in the analysis, which is a different issue from respondents not providing a response to a seen item.

The last two studies from the search indicated that missing responses in DCM should not be ignored. Xin and Zhang (2015) recognized that the assumptions made about the missing data were required in equating, but they did not address this issue in their proposed method using attribute mastery profile. Similarly, considering the mastery profile, Ayers et al. (2009) conducted simulation studies and found that the classification accuracy decreased as the percentage of missing responses increases (under MCAR) with the Bayesian estimation procedure, based on the agreement between the true skill profiles and the clustering results. Although these studies did not focus on MDT, the results showed that the missing response issue in DCM should be further studied.

In summary, educational researchers in the DCM context in general use LD, PM, or IN, or leave missing responses to the software. Although advanced MDTs have been developed in contexts such as SEM and IRT, it is unclear how well DCMs are estimated in the presence of missing data.

Table 3

Diagnostic Classification Model Studies Mentioned Missing Data

Study	Purpose	Data source	DCM	Treatment of missing data
Ayers, Nugent, & Dean (2009)	To compare three estimate methods of student attribute knowledge.	Simulated data	DINA	NM for DINA; zero imputation for sum scores; ignore the items for capacity matrix
Gu (2012)	To explore additional information gained through distractors from the multiple-choice test items.	Simulated data and real data	DINA	Missing responses were assigned the item option least selected by the sample
Hansen (2013)	To propose a DCM that accounts for item local dependence.	Simulated data and real data	LCDM	Zero imputation
Harrison, Bradshaw, Naqvi, Paff, & Campbell (2017)	To examine the psychometric properties of an autism spectrum disorder knowledge measure	Real data	---	“Don’t know” = IN
Lee, Park, & Taylan (2011)	To obtain item information and attribute mastery information	Real data	DINA	IN
Sheehan, Tatsuoka, & Lewis (1993)	To introduce a modification to the rule space model for processing response vectors containing missing data.	Real data	Rule space model	Only responses to administered items were compared to the ideal response vectors of those items.
Skaggs, Wilkins, & Hein (2016)	To explore the grain size and sample size on parameter recovery	Real data	GDM	Mdltm program marked not administered booklets as not reached
Templin & Hoffman (2013)	To demonstrate using Mplus.	Real data	LCDM	FIML (Handled by Mplus)
Xin & Zhang (2015)	To propose a local equating method without an anchor test.	Simulated data	DINA	NA

Note. DCM = Diagnostic classification model; DINA = Deterministic input noisy *and* gate model; NA = Not apply; NM = No mention; LCDM = Log-linear cognitive diagnostic model

Review of DCM Missing Data Studies

There have been four studies regarding missing data in the DCM context. Zhang (2014) studied the missing patterns, and three studies (Dai, 2017; Dai, Svetina, & Chen, 2018; Sünbül, 2017) directly focused on MDTs.

Zhang (2014) investigated the relationships between missingness and students' characteristics and between missingness and the skill mastery profile. The data were from the 2006 Ontario Secondary School Literacy Test in which the students were instructed to answer all items. In his preliminary study, Zhang compared IN and leaving missing responses blank when estimating the R-RUM to select a more ideal method to handle missing data. For each of the skills, he compared the probability of skill mastery estimated by the two methods for the group of respondents who did not have missing responses and for the group who had missing responses. He found that while the probability estimates for students who did not have missing responses did not differ, treating missed items as incorrect lowered the probability estimates. He continued his research leaving missing responses blank using the software Arpeggio (DiBello & Stout, 2008). He found that (a) the majority examinees did not have any missing responses; (b) items measuring more difficult attributes and items that were more difficult tended to have more missing responses; (c) when students had high numbers of missing responses, they tended to be not-reached items; and (d) missing responses were related to examinee characteristics. He also stated that "if nonresponse is concentrated in items that require particular skills, the accuracy of the estimates for those skills will be lower than for other skills" (p. 79), which could be addressed by Q-matrix design. Since diagnostic tests are typically low-stakes tests and since respondents are usually allowed enough time to complete the entire test, not-reached items should rarely occur. In the current study, I did not treat unreached items differently from missing responses in general. One salient difference between high-stakes tests and diagnostic tests regarding unanswered items lies in respondents' obligation to provide all answers. Under high-stakes settings, examinees may be more willing to guess and have fewer nonresponses, while respondents to diagnostic tests could feel not as pressured to provide answers to all questions because of their characteristics, because of their lack of mastery or test-taking skills, or because of unwillingness. However, it is reasonable to assume that the missingness is related to attribute difficulty, item difficulty, and examinee characteristics, that is, MNAR or MAR; on the other hand, as Finch (2008) stated, the missingness is produced by respondents inadvertently skipping items, which indicates MCAR.

The three MDT studies were conducted in the DINA model context. Some of the simulation conditions from these studies are shown in Table 4.

Table 4

Simulation Conditions of Previous Three Studies

Conditions	Dai (2017)	Dai, Svetina, & Chen (2018)	Sünbül (2017)
Missing data mechanisms	MAR, MNAR, MIXED	MAR, MNAR	MAR, MCAR
Missing rates	0%, 5%, 10%, 30%	0%, 10%, 20%, 30%	5%, 10%, 15%
Sample sizes	1,000	500; 1,000; 2,000	1,000; 2,000; 3,000
Number of attributes	3, 5, 8	3, 5	4
Number of items	35	20, 40	10, 30
Item discrimination [a]	High U (.05, .25); Low U (.25, .45)	U (0, .2)	—-

Note. [a]Item discrimination is based on a uniform distribution of slipping and guessing parameters.

Dai (2017) explored the effect of MDTs (IN, person mean imputation [PM], TW, RF, and EM) on the recovery of item parameters and of the attribute profile. Three Q-matrix designs were used, and the average numbers of attributes measured per item were 1.77, 2.57, and 3.71 in the three-attribute, five-attribute, and eight-attribute Q-matrices respectively. He generated missing responses of MAR by referring to a hypothetical continuous variable as the proxy inversely related with the probability of missing data. MNAR missing responses were generated by assigning a higher probability of missing to incorrect responses and lower probability of missing to correct responses in the complete data set. The MIXED missing responses were generated through the approach of De Ayala, Plake, and Impara (2001): the probability of omitting an item was stochastically related to the correctness of the response and to the relative frequency of omission of all respondents in the specific total score fractile. Mean bias and root mean squared error (RMSE) of the slipping and guessing parameters, attribute-wise classification accuracy, and pattern-wise classification accuracy were used as criteria to evaluate the performance of each of the MDTs. Dai reached the conclusion that (a) as the missing rate increased, the differences among MDT performances became more prominent, especially under MAR and MNAR conditions, that (b) although all MDTs' performance in classification accuracy decreased with an increase in missing rate, EM performed relatively better in all MAR, MNAR, and MIXED conditions, and PM performed relatively worse, and that (c) IN performed consistently below satisfying in all three missing data mechanism conditions, while PM was only acceptable under MAR with high item discrimination.

The superiority of EM is also supported Dai et al. (2018) in a different context. They investigated the effect of four MDTs (IN, logistic regression imputation, LD, and EM) on two Q-matrix refining methods, and they found that the patterns in both MAR and MNAR data were similar–EM and logistic regression imputation performed superior with EM slightly better.

The third study (Sünbül, 2017) focusing on MDTs also involved IN and EM as in Dai (2017) and Dai et al. (2018), together with PM and TW. Similarly, Sünbül (2017) found that IN had higher average RMSEA and lower pattern-wise classification accuracy with MAR and MCAR data, compared with PM, TW, and EM. In Sünbül's study, missing data under MAR were created based on the total score, higher total scores associating with lower missing rate. Although she neither introduced the structure of the Q-matrix nor specified how the mean RMSEA and pattern-wise classification accuracy were computed, her conclusions echoed the results of Dai (2017) indicating: IN, as a common MDT, is not satisfying.

Although the above studies contribute to the literature, some of the decisions made regarding the simulation process could be improved to better assist practitioners in practical settings. From a theoretical perspective, using PM, TW, and RF as MDTs in the context of DCM ignores the multidimensional nature of the data because these methods work best when the data are unidimensional. Additionally, the way MAR data were generated in all three studies was related either to a hypothetical normally distributed continuous variable (as in Dai, 2017; Dai et al., 2018) or to the sum score (as in Sünbül, 2017), assuming missingness related to general ability instead of attribute difficulty. From a practical perspective, the favored MDT, the EM algorithm, requires deep statistical knowledge, which hinders its use by practitioners. In the previous section, I found that practitioners rarely employ such advanced methods of treating missing data. Therefore, easy, accessible MDT should be studied, and FIML was chosen in this study.

The authors of all the above studies recognized that the DINA model had stringent assumptions and recommended further research on other models. As mentioned above the LCDM is a general model that is also widely used in the field. Additionally, the use of this model has been demonstrated in several studies (Kunina-Habenicht et al., 2017). Especially, Templin and Hoffman (2013) have provided easily accessible guidance on using SAS to generate macros to produce syntax that can be used in Mplus, and Templin (2016) has published the R functions for the same purpose. Considering the easy access of R (R Core Team, 2019), the current study used the Mplus Automation package (Hallquist & Wiley, 2018) in R and Mplus to handle missing data.

I investigated three measures as indicators for the estimation: (a) attribute classification accuracy, (b) profile classification accuracy, and (c) attribute reliability. Attribute classification accuracy was defined as the percentage of accurate mastery classification at the attribute level, that is, the percentage of respondents where the estimated attribute mastery from the estimation and the known attribute mastery were the same. Profile classification accuracy was also defined as the percentage of accurate mastery classification, but at the profile level. The attribute classification accuracy and the profile classification accuracy were values within the range of 0 and 1, and since profile classification accuracy encompasses all measured attributes, it was expected to be lower than attribute classification accuracy. Attribute reliability was defined to measure attribute classification consistency, calculated according to Templin and Bradshaw (2013).

CHAPTER 3

Method

The purpose of this study was twofold: (a) to investigate how well the MLR estimator which uses the full information maximum likelihood technique (FIML) to handle missing data in Mplus estimating the main effect LCDM, or C-RUM, in the presence of missing responses, and (2) to make recommendations on choosing between two alternative, unbalanced Q-matrix designs. In this study, I conducted simulation studies using conditions reflecting scenarios practitioners would likely be in.

Simulation Study

In this section, I describe the simulation conditions in detail.

Q-matrix. As the literature review indicates, the majority of diagnostic instruments consist of three to four attributes. The current study included three-attribute Q-matrices. Considering the practical challenges in writing items prescribed by the Q-matrix, researchers could encounter situations where certain attributes cannot be measured in isolation. For example, to assess skills to solve linear algebra equations, there is the skill involving how to and when to simplify an equation by removing parentheses. Removing parentheses and distributing terms is almost always followed by combining like terms. Hence, it would be unrealistic to isolate this skill alone. Therefore, the Q-matrix design (Appendix A, Appendix B, and Appendix C) in the current study reflected both the balanced design and the scenario where an attribute cannot be isolated (Table 5).

For the ideal three-attribute Q-matrix, BAL-3, I used the one used in Dai et al. (2018) with 20 items. The numbers of items ensured that all combinations could be measured and that there were not too many items to burden the respondents. DCM studies have generally used instruments containing 20 to 30 items. For example, reading comprehension assessments of Chen and Chen (2016), Kim (2015), Li et al. (2016), and Ravand (2016) contained 26, 30, 20, and 20 items respectively; situational judgement tests of García, Olea, and de la Torre (2014) and Sorrel et al. (2016) contained 26 and 23 items respectively.

Table 5

Q-matrix Description

Q-matrix	Number of attributes per item	Description
BAL-3	1.65	Single attributes and paired attributes were measured three times; all three attributes combined are measured twice
PAIR-3	1.85	All combinations were measured three times, except for the single focal attributes; the last two items measure two different combinations of the focal attribute and one other attribute
ALL-3	1.95	All combinations were measured three times, except for the single focal attributes; the last two items measure all three attributes

Note. The focal attribute is assumed not to be isolated in PAIR-3 and ALL-3.

Following the example of Madison and Bradshaw (2015), I focused on one focal attribute (FA) in the analysis of attribute classification and attribute reliability. Because more than one attribute is specified in the test, the FA was selected to be the attribute of focus for simplicity when measuring attribute classification accuracy and attribute reliability. In order to reflect more authentic situations, the FA in each of the unbalanced Q-matrices were assumed to be unisolatable. It is worth pointing out that the FA was assumed to be measured in conjunction with any of the other attributes, which is different from Madison and Bradshaw. In some of their Q-matrices, the FA could only be combined with another specific attribute. This design could be appropriate to show that the mastery of FA relies on the mastery of another attribute, but because attribute hierarchy is not the focus of this study, this design is not used in the current study.

Two versions of an unbalanced Q-matrix were created. Because FA cannot be measured by itself, the unbalanced Q-matrices shared the same 12 items that did not measure FA in isolation. With a fixed test length, that left eight items to be matched to attributes. To follow the recommendation of Madison and Bradshaw (2015), six of the eight items replicated each attribute combination, and two items were left. One version of the unbalanced Q-matrixes (PAIR-3) had those two items measuring pairs of FA and another attribute. As in the example of linear algebra, distributing is measured either together with combining like terms or together with operating across the equal sign. The second version of the unbalanced Q-matrix (ALL-3) had the remaining two items measuring all three attributes. That is, distributing is measured together with combining like terms and operating across the equal sign in all the rest of the items.

Sample sizes. The sample sizes in DCM studies range from 138 in Sorrel et al. (2016) to 10,000 in Ravand (2016), while some were slightly above 2,000 (Chen & Chen, 2016; Li & Suen, 2013) and some were slightly less than 2,000 (Kim, 2015). Because most simulation studies had sample size of 1,000, the current study uses sample sizes of 500; 1,000; and 2,000.

Item and attribute conditions. In this study, I adopted the simulation conditions of Madison and Bradshaw (2015) in terms of items and attributes: (a) the probability of correct response interval for complete nonmasters was between .10 and .30; (b) the probability of correct response interval for masters of one attribute was between .35 and .45, (c) the probability of correct response interval for masters of two attributes was between .46 and .70, (d) the probability of correct response interval for complete masters as between .75 and .90; (e) tetrachoric correlation was fixed at .70; and (f) attribute mastery base rate as .5.

The intercepts were drawn from U (-2.197, -0.847) on the logit scale corresponding to the probability interval of correct response for complete nonmasters of .10 and .30 as mentioned above. All interaction terms were set to 0, which means the data generating model was the main effect LCDM, or the C-RUM. When it came to generating the main effects, the procedure depended on the number of attributes assessed by the item. If an item measured one attribute, the main effect was drawn from a uniform distribution with a minimum of the difference of 1.099 and the intercept and a maximum of the difference of 2.197 and the intercept so that the probability of complete masters were within the interval of .75 and .90. If an item measured two attributes, main effects were drawn so that the sum of either main effect and the intercept was between -0.619 and -0.200 (corresponding to the probabilities of correct response from masters of one attribute which were .35 and .45) and that the sum of the intercept and two main effects were between 1.099 and 2.197 (corresponding to the probabilities of correct response from masters of all attributes which were .75 and .90). If an item measured three attributes, main effects were drawn so that the sum of either main effect and the intercept was between -0.619 and -0.200 (corresponding to the probabilities of correct response from masters of one attribute which were .35 and .45), the sum of any two main effects and the intercept was between -0.160 and 0.847 (corresponding to the probabilities of correct response from masters of two attributes which were .46 and .70), and that the sum of the intercept and all three main effects was between 1.099 and 2.197 (corresponding to the probabilities of correct response from complete masters which were .75 and .90).

Missing percentage. It is reasonable to assume that the test is beyond the respondents' current ability and a higher missing rate can be expected, but it is also reasonable to assume that the respondents try hard to guess especially in multiple-choice tests, so a lower missing rate can also be expected. As Peugh and Enders (2004) found missing rates could range from less than 1% to 67% in educational research, it is hard to speculate a reasonable range for simulation. Therefore, this study will use missing rates from the previous three studies, 0%, 5%, 10%, 15%, 20%, and 30%. In total, the study contains 144 simulation conditions with missing values (3 Q-matrices * 3 sample sizes and 3 Q-matrices * 3 sample sizes * 3 missing data mechanisms * 5 missing rates), which are summarized in Table 5 and Table 6. Each simulation condition will be repeated 100 times.

Table 6

Simulation Conditions

Characteristic	Value/ Interval
Number of attributes and items	3 attributes with 20 items
Sample sizes	500; 1,000; 2,000
Tetrachoric correlations among attributes	.70
Probability of correct response interval for complete nonmasters	(.10, .30)
Probability of correct response interval for complete masters	(.75, .90)
Missing data mechanism	MCAR, MAR, MNAR
Missing rates	0%, 5%, 10%, 15%, 20%, and 30%

Note. Complete nonmasters do not master any attributes; complete masters have mastered all attributes.

Data sets containing missing responses. Data sets containing missing responses were generated from the corresponding complete data set. For each of the 100 complete data sets, I generated a separate data set under MCAR, MAR, and MNAR missing data mechanisms. For each simulated data set, there were two conditions needed to be met: (a) the missing rate specified by the study, and (b) the propensity of missing for each cell dictated by the missing data mechanism. In order for both conditions to be met at the same time, a transformation was made of the missing propensity matrix so that the mean of all the cells was the specified missing rate.

The missing propensity matrices for the three missing data mechanisms were specified differently. For the MCAR data set, the missingness cannot be predicted. Therefore, the missing propensity matrix of the size of the complete data was created with the cells being random draws from a uniform U (0,1). The MAR data set generation process is detailed in Appendix D. The missingness was inversely related to attribute difficulties defined as the difference of 1 and the marginal probability of mastering the attribute, or the probability of not mastering the attribute. For the MNAR data set, the missingness was specified to be related to the incorrectness of each item. If the response in the complete data was wrong, the corresponding cell had a higher propensity of being missing. To differentiate MNAR from MAR, the missing propensity matrix for MNAR was specified such that the corresponding correct response cells were random draws from U (0.5, 1) and the corresponding incorrect response cells were random draws from U (0, 0.5).

Analysis

Each simulated data set was analyzed using FIML in Mplus with the C-RUM model and corresponding Q-matrix. For each analysis, EAP attribute estimates, MAP attribute estimates, and MAP profile estimates were saved using R.

The attribute classification accuracy of FA and profile classification were calculated as previously described, holding the known attribute mastery or the known latent class membership as the standard. Due to maximum likelihood estimation, some values of marginal attribute estimates were greater than one, and the value of 1 was used to replace those values in the process of calculating attribute reliability of FA.

Prior to answering the research questions, the convergence rates for each of the simulation conditions were calculated and were reported in the next section. Although the convergence rates do not answer research questions directly, they provide the sample size context for the study and for making recommendations.

Research question 1. Under the condition of a balanced Q-matrix design, what is the effect of different percentages of missing responses (5%, 10%, 15%, 20%, and 30%) on attribute classification accuracy, profile classification accuracy, and attribute reliability across sample sizes (500; 1,000; and 2,000) in all missing data mechanism conditions (MCAR, MAR, and MNAR), compared with complete data?

To answer research question 1, only data from BAL-3 were used for the analysis. Factorial ANOVA analyses were conducted with each of the three measures as a separate dependent variable and the following independent variables: missing percentage, sample size, missing data mechanism. Because of the large sample size, $\eta^2$was examined to isolate the important effects rather than simply relying on the p-values. Additionally, line graphs were graphed with y-axis being the measure (attribute classification accuracy, profile classification accuracy, or reliability), x-axis being the percentage of missing responses to assist understanding of the results of factorial ANOVA results.

Research question 2. When an attribute cannot be measured in isolation (unbalanced Q-matrix), what is the effect of different percentages of missing responses (5%, 10%, 15%, 20%, and 30%) on attribute classification accuracy, profile classification accuracy, and attribute reliability across sample sizes (500; 1,000; and 2,000) in all missing data mechanism conditions (MCAR, MAR, and MNAR), compared with complete data?

To answer research question 2, the same analyses for question 1 were conducted with Q-matrices PAIR-3 and ALL-3 instead of BAL-3. The results would be presented in a similar way, separating the results for each Q-matrix.

Research question 3. Between the two alternative Q-matrix designs, which one obtains higher attribute classification accuracy, profile classification accuracy, and attribute reliability?

Attribute classification accuracy, profile classification accuracy, and attribute reliability were compared between PAIR-3 and ALL-3 matrices, and Cohen's *d* would be reported as effect sizes. Boxplots of both Q-matrices across missing rates for each of the three measures would be presented in the next section.

CHAPTER 4

Results

As anticipated, not all analyses converged. Table 7 reports the convergence rate for each simulation condition as described in Table 5 and Table 6. Each simulation condition was repeated 100 times, so the number of converged analyses for each condition was the product of 100 and the corresponding convergence rate. Model convergence criteria were default settings with the maximum number of iterations of the EM algorithm set at 500 across all conditions.

Research Question 1

Under the condition of a balanced Q-matrix design, what is the effect of different percentages of missing responses (5%, 10%, 15%, 20%, and 30%) on attribute classification accuracy, profile classification accuracy, and the attribute reliability across sample sizes (500; 1,000; and 2,000) in all missing data mechanism conditions (MCAR, MAR, and MNAR), compared with complete data?

Attribute classification accuracy. Factorial ANOVA results showed that the missing percentage [F (5, 4745) = 589.555, p < .001] and the sample size [F (2, 4745) = 7.001, p < .001] were statistically significant effects, but other than the missing percentage (η^2=0.382), all other effects (sample size, missing data mechanism, and all possible interaction effects among these factors) had negligible effect sizes. Figure 1 and Figure 2 illustrate that attribute classification accuracy did not differ much across missing data mechanism or sample size as the colored dots and lines are close together.

Table 7

Convergence Rate Based on Simulation Condition

	Percent of Missing Data					
Condition	0%	5%	10%	15%	20%	30%
BAL-3						
500						
MCAR	1.00	1.00	1.00	1.00	1.00	.97
MAR	1.00	1.00	1.00	1.00	.98	.99
MNAR	1.00	1.00	1.00	1.00	1.00	1.00
1,000						
MCAR	1.00	1.00	1.00	1.00	1.00	1.00
MAR	1.00	1.00	1.00	1.00	1.00	1.00
MNAR	1.00	1.00	1.00	1.00	1.00	.99
2,000						
MCAR	1.00	1.00	1.00	1.00	1.00	1.00
MAR	1.00	1.00	1.00	1.00	1.00	1.00
MNAR	1.00	1.00	1.00	1.00	1.00	1.00
PAIR-3						
500						
MCAR	.98	.99	.95	.95	.89	.87
MAR	.98	.95	.93	.91	.86	.86
MNAR	.98	.96	.96	.90	.87	.85
1,000						
MCAR	1.00	1.00	1.00	1.00	.99	.98
MAR	1.00	1.00	.99	.99	1.00	.96
MNAR	1.00	1.00	1.00	1.00	1.00	.97
2,000						
MCAR	1.00	1.00	1.00	1.00	1.00	1.00
MAR	1.00	1.00	1.00	1.00	1.00	1.00
MNAR	1.00	1.00	1.00	1.00	1.00	1.00
ALL-3						
500						
MCAR	.91	.92	.89	.90	.82	.71
MAR	.91	.88	.80	.72	.64	.52
MNAR	.91	.89	.89	.85	.84	.73
1,000						
MCAR	.97	.99	.96	.94	.92	.90
MAR	.97	.97	.96	.97	.99	.88
MNAR	.97	.98	.99	.95	.95	.95
2,000						
MCAR	1.00	1.00	1.00	1.00	1.00	.99
MAR	1.00	1.00	1.00	.99	1.00	1.00
MNAR	1.00	1.00	1.00	1.00	1.00	1.00

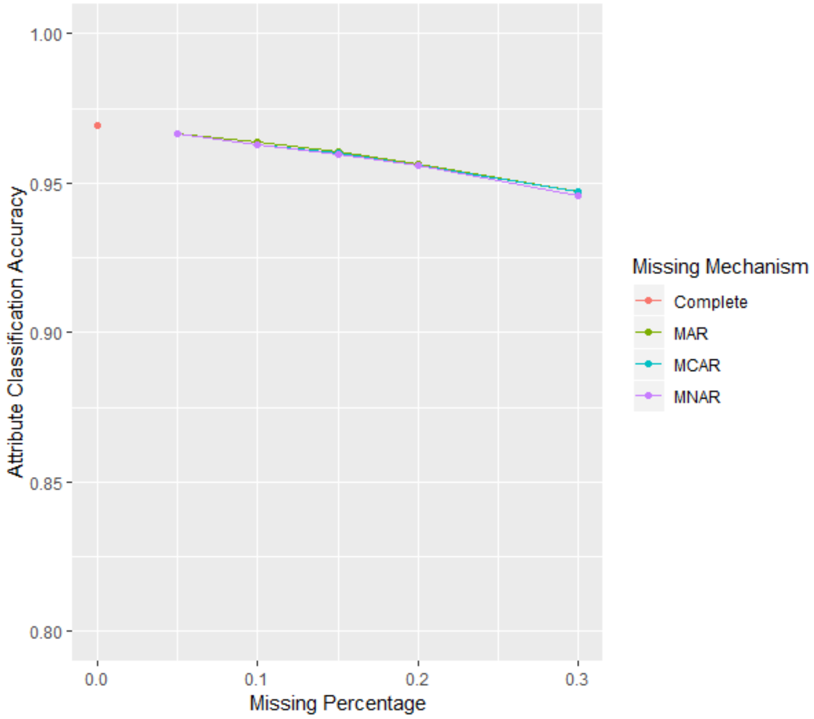

Figure 1. Attribute classification accuracy across missing data mechanisms under BAL-3 Q-matrix design.

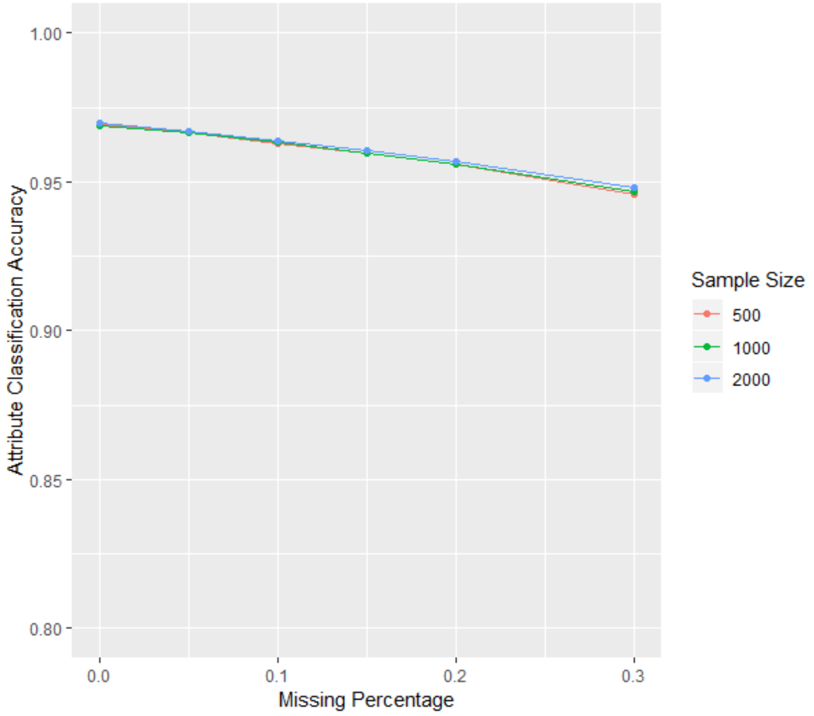

Figure 2. Attribute classification accuracy across sample sizes under BAL-3 Q-matrix design.

Profile classification accuracy. Factorial ANOVA results showed that the missing percentage [F (5, 4745) = 1278.314, p < .001], the sample size [F (2, 4745) = 19.870, p < .001], and the missing data mechanism [F (2, 4745) = 5.376, p < .01] were statistically significant effects, but other than the missing percentage (η^2=0.571), all other effects had negligible effect sizes. Figure 3 and Figure 4 illustrate that profile classification accuracy did not differ much across missing data mechanisms or sample sizes. Figure 3 and Figure 4 illustrate that profile classification accuracy did not differ much across missing data mechanism or sample size as the colored dots and lines are close together.

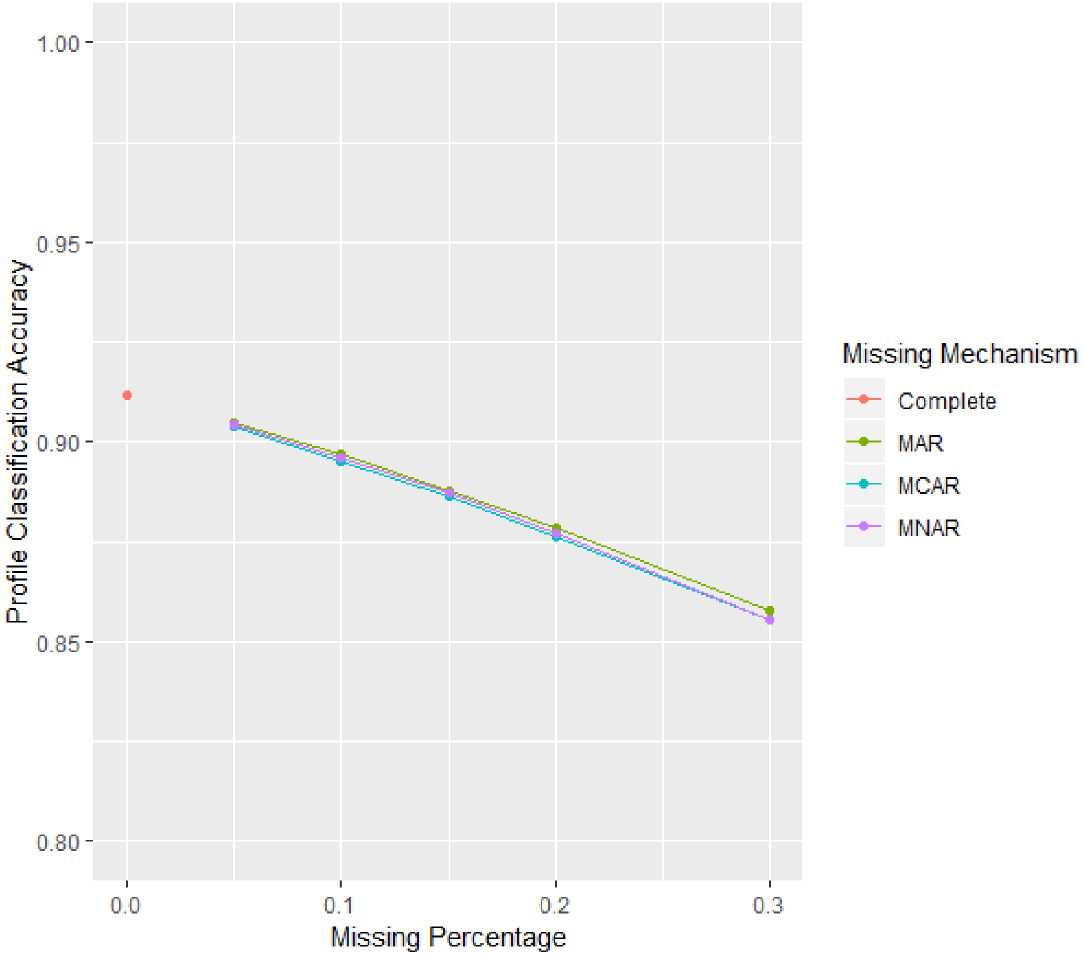

Figure 3. Profile classification accuracy across missing data mechanisms under BAL-3 Q-matrix design.

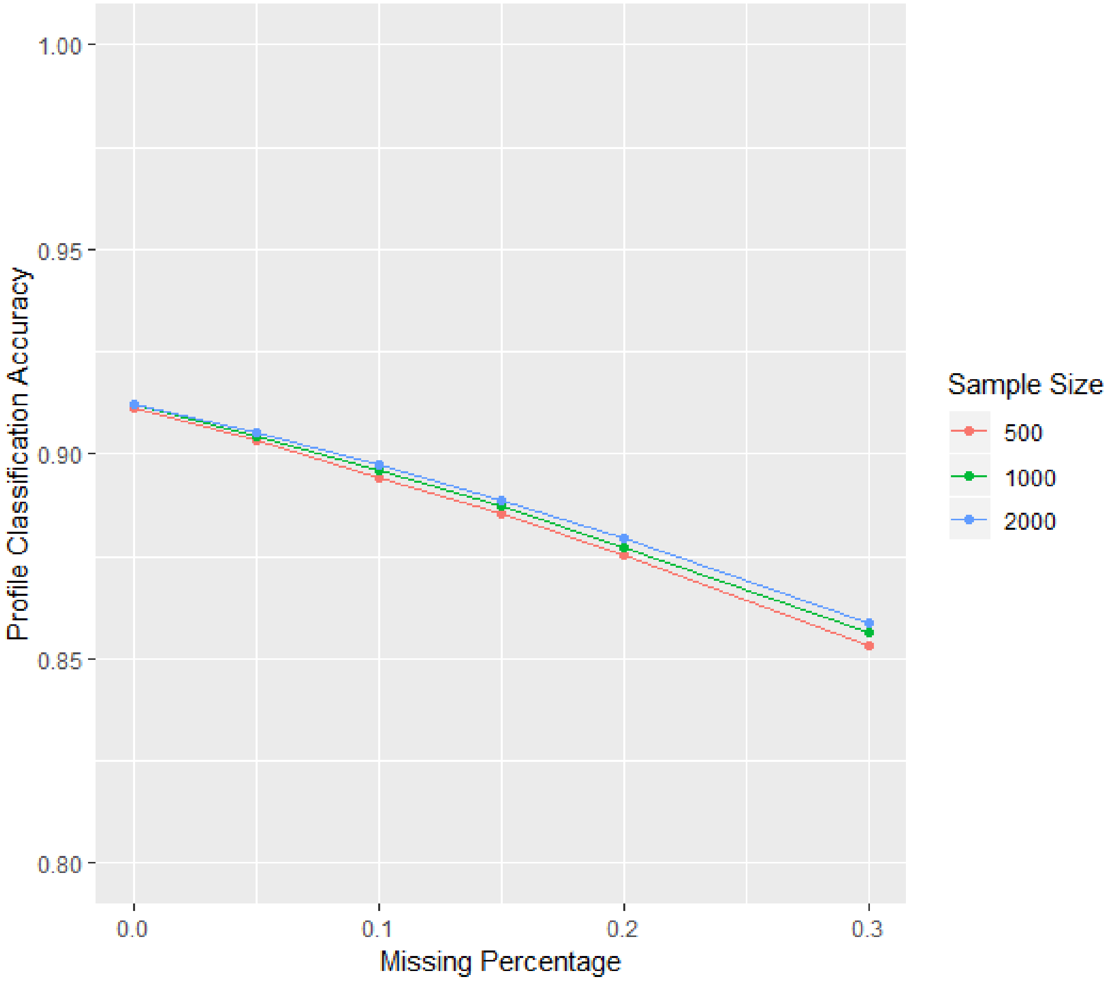

Figure 4. Profile classification accuracy across sample sizes under BAL-3 Q-matrix design.

Attribute reliability. Factorial ANOVA results showed that the missing percentage [F (5, 4745) = 782.535, p < .001], the sample size [F (2, 4745) = 41.625, p < .001], the missing data mechanism [F (2, 4745) = 3.040, p = .048], and the two-way interaction between missing percentage and sample size [F (10,4735) = 2.881, p<.01] were statistically significant effects, but other than the missing percentage (η^2=0.445) and the sample size (η^2=0.009), all other effects had negligible effect sizes. Figure 5 and Figure 6 illustrate that attribute reliability did not differ much across missing data mechanism or sample size as the colored dots and lines are close together.

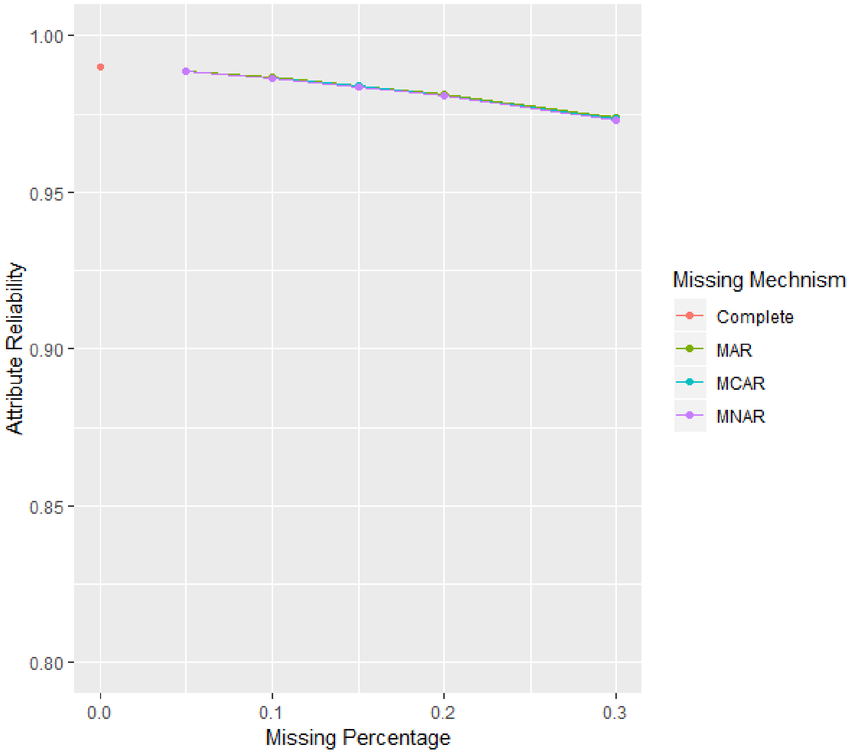

Figure 5. Attribute reliability across missing data mechanisms under BAL-3 Q-matrix design.

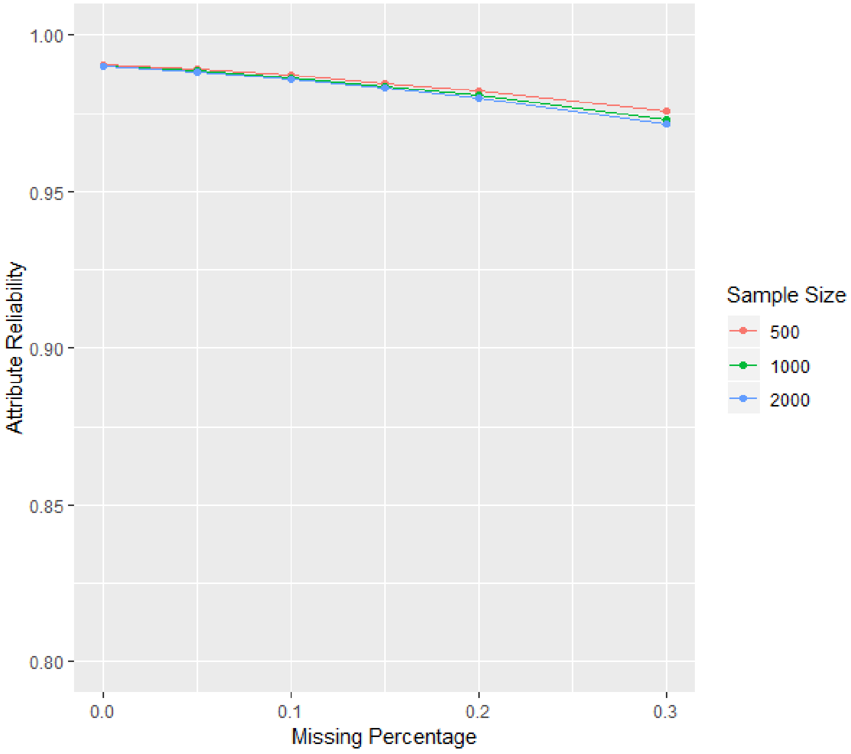

Figure 6. Attribute reliability across sample sizes under BAL-3 Q-matrix design.

Summary. Although the effects that were statistically significant from the factorial ANOVA were not consistent across measures, from the perspective of effect size, only missing percentage had an effect on all dependent measures. Each measure has a tendency of decreasing with the increase of missing percentage regardless of the missing data mechanism. An unexpected phenomenon of higher attribute reliability associated with small sample size is noticed in Figure 6, and this issue is discussed in the later chapter.

Research Question 2

When an attribute cannot be measured in isolation (unbalanced Q-matrix), what is the effect of different percentages of missing responses (5%, 10%, 15%, 20%, and 30%) on attribute classification accuracy, profile classification accuracy, and the attribute reliability across sample sizes (500; 1,000; and 2,000) in all missing data mechanism conditions (MCAR, MAR, and MNAR), compared with complete data?

Attribute classification accuracy with PAIR-3. Factorial ANOVA results showed that the missing percentage [F (5, 4608) = 53.282, p < .001] and the sample size [F (2, 4608) = 153.744, p < .001] were statistically significant effects, and only these two effects had nonnegligible effect sizes (missing percentage η^2=0.445 and sample size η^2=0.059). Figure 7 illustrates that attribute classification accuracy did not differ much across missing data mechanism, and Figure 8 illustrates that attribute differed among three sample sizes.

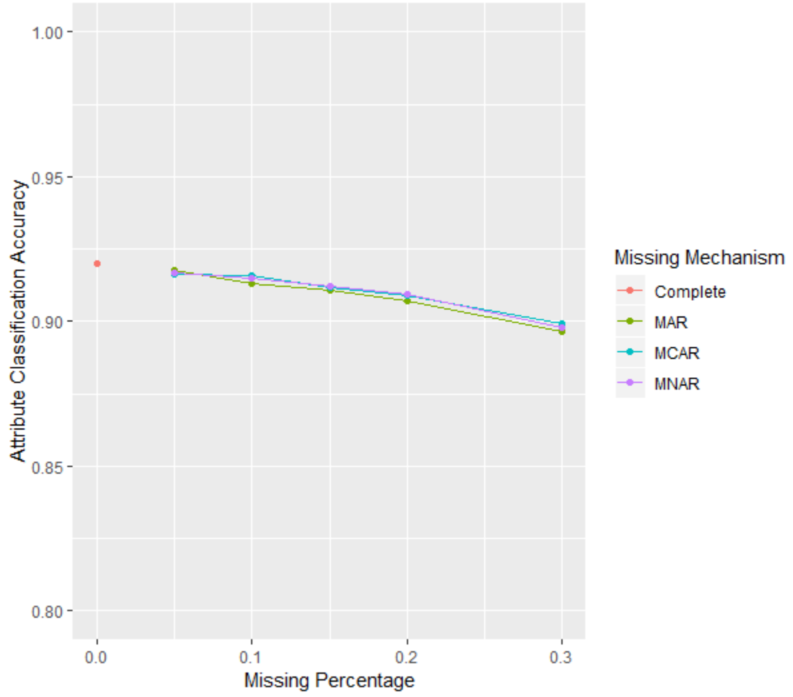

Figure 7. Attribute classification accuracy across missing data mechanisms under PAIR-3 Q-matrix design.

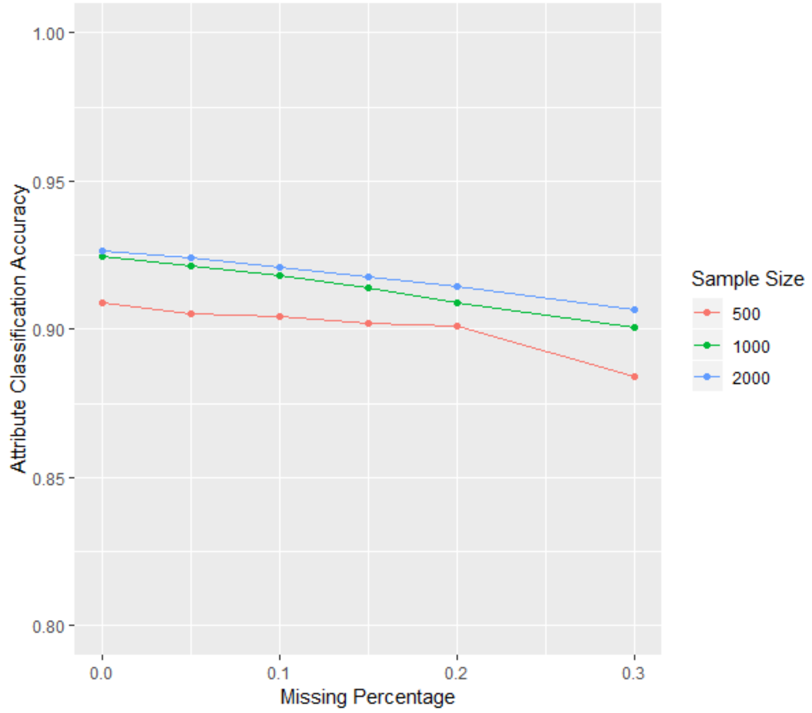

Figure 8. Attribute classification accuracy across sample sizes under PAIR-3 Q-matrix design.

Profile classification accuracy with PAIR-3. Factorial ANOVA results showed that the missing percentage [F (5, 4608) = 253.243, p < .001] and the sample size [F (2, 4608) = 200.614, p < .001] were statistically significant effects, and only these two effects had nonnegligible effect sizes (missing percentage η^2=0.201 and sample size η^2=0.064). Figure 9 illustrates that profile classification accuracy did not differ much across missing data mechanism, and Figure 10 illustrates that profile classification accuracy did differ across three sample sizes.

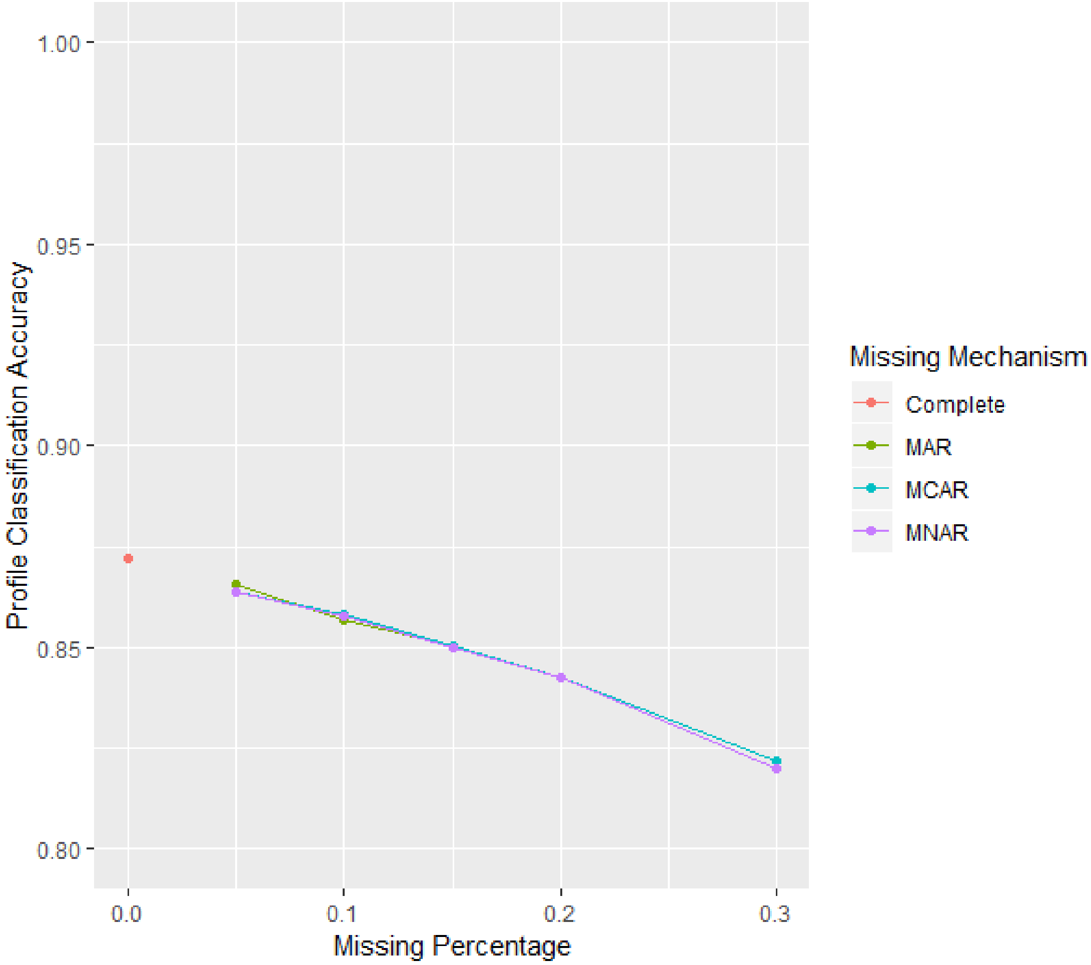

Figure 9. Profile classification accuracy across missing data mechanisms under PAIR-3 Q-matrix design.

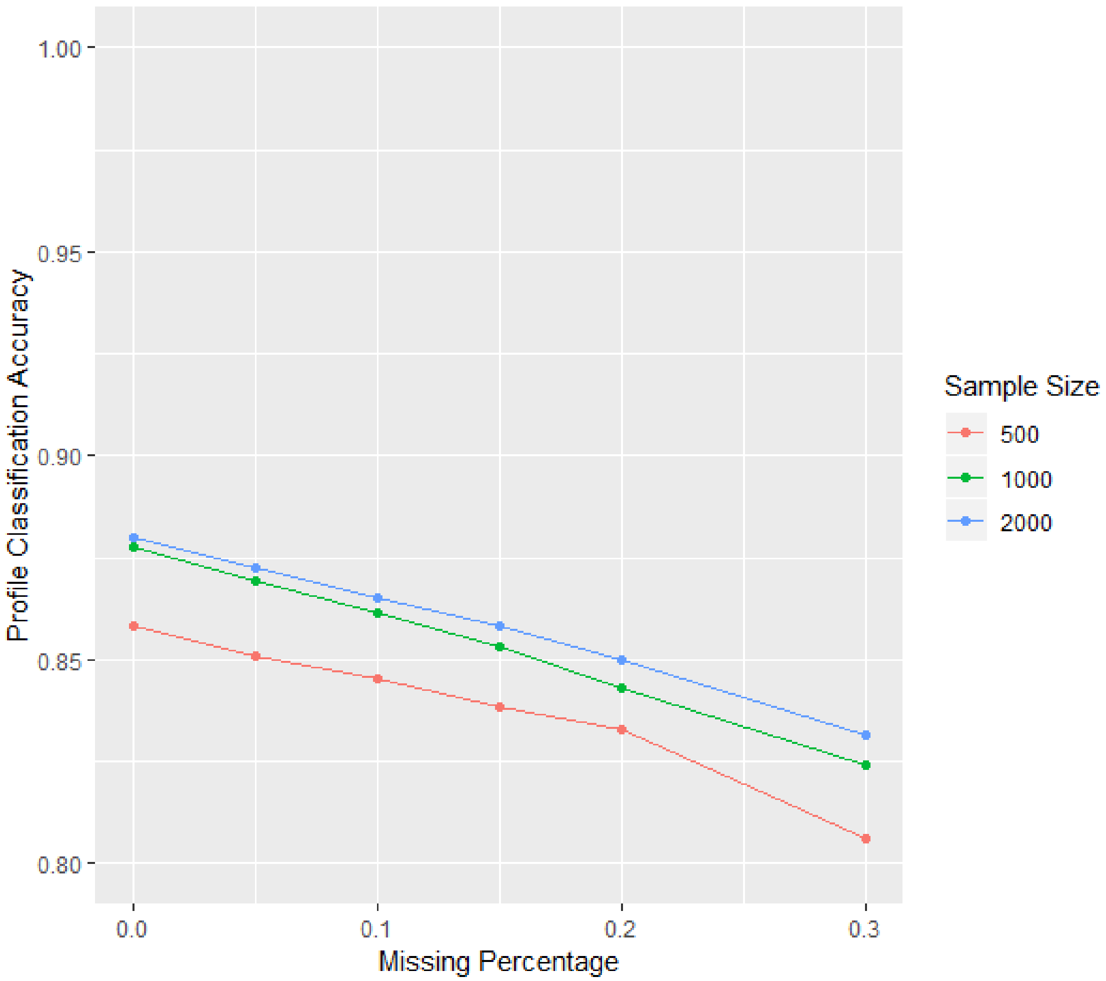

Figure 10. Profile classification accuracy across sample sizes under PAIR-3 Q-matrix design.

Attribute reliability with PAIR-3. Factorial ANOVA results showed that only the missing percentage [F (5, 4608) = 47.418, p < .001] was statistically significant, and only this effect had nonnegligible effect size (η^2=0.049). Figure 11 and Figure 12 illustrate that attribute reliability did not differ much across missing data mechanism or sample size as the colored dots and lines are close together.

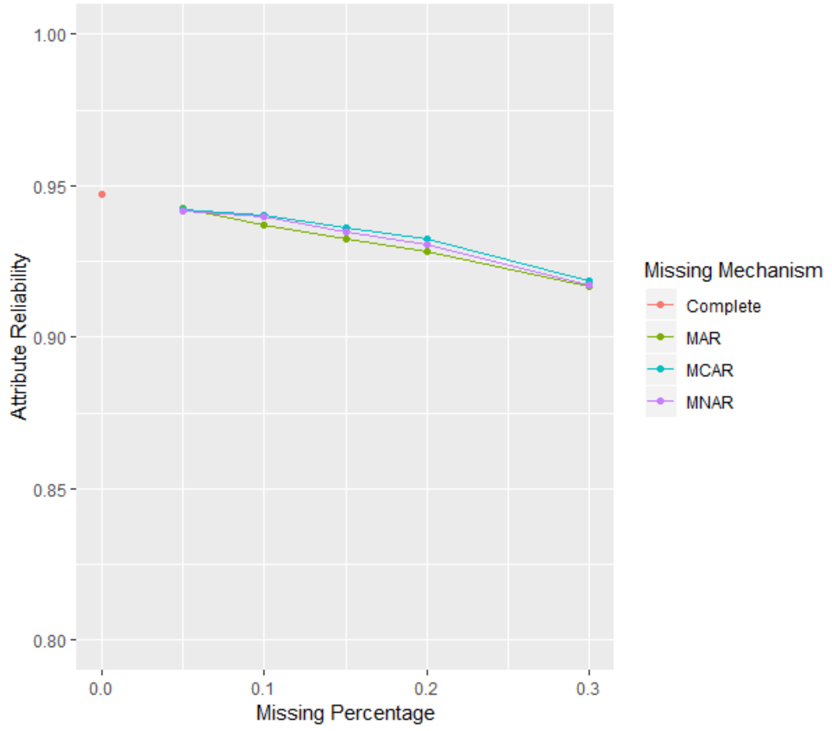

Figure 11. Attribute reliability across missing data mechanisms under PAIR-3 Q-matrix design.

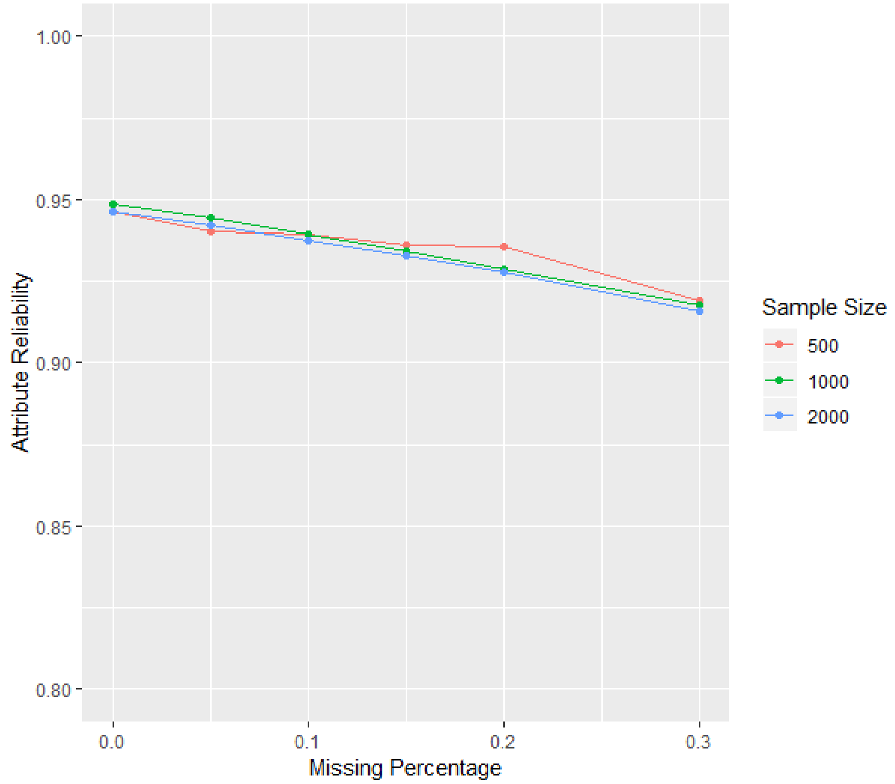

Figure 12. Attribute reliability across sample sizes under PAIR-3 Q-matrix design.

Attribute classification accuracy with ALL-3. Factorial ANOVA results showed that several effects were statistically significant: the missing percentage [F (5, 4368) = 10.373, p < .001], the sample size [F (2, 4368) = 113.540, p < .001], the missing data mechanism [F (2, 4368) = 9.668, p < .001], the two-way interaction between sample size and the missing data mechanism [F (4, 4368) = 3.149, p=.0135], and the three-way interaction [F (16, 4368) = 3.667, p < .001]. Three effects had nonnegligible effect sizes: missing percentage (η^2=0.011), sample size (η^2=0.048), and the three-way interaction among missing percentage, sample size, and missing data mechanism (η^2=0.012). Figure 13 illustrates that attribute classification accuracy did not differ more than .10 across missing data mechanisms, and Figure 14 illustrates that attribute classification accuracy differ across sample sizes.

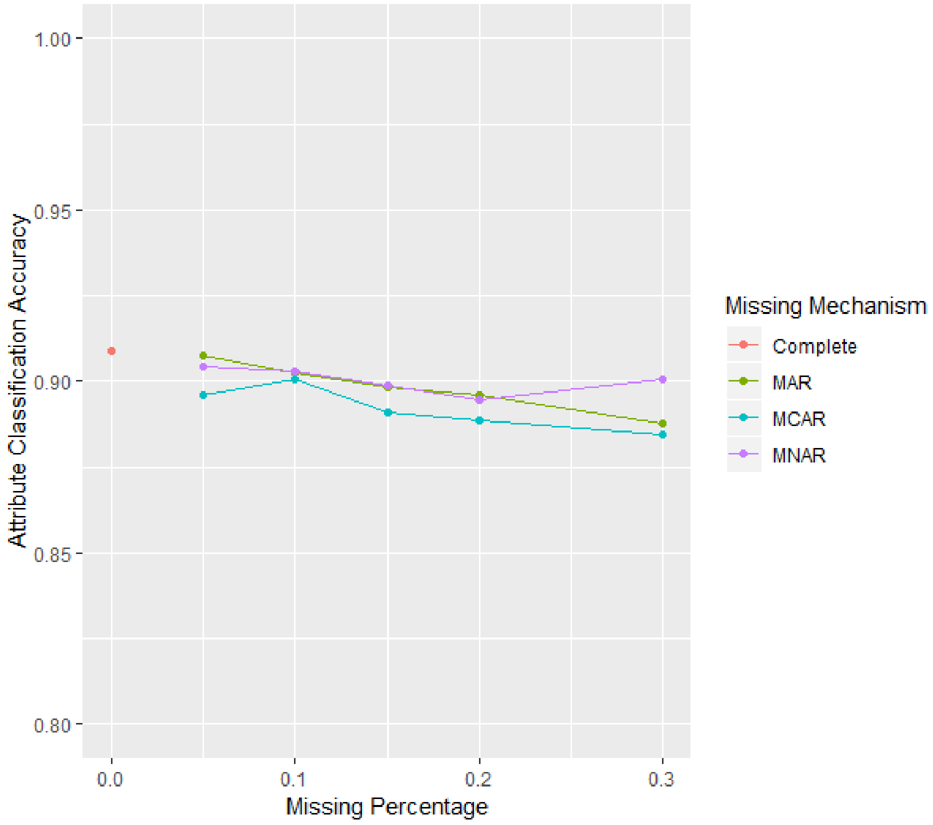

Figure 13. Attribute classification accuracy across missing data mechanisms under ALL-3 Q-matrix design.

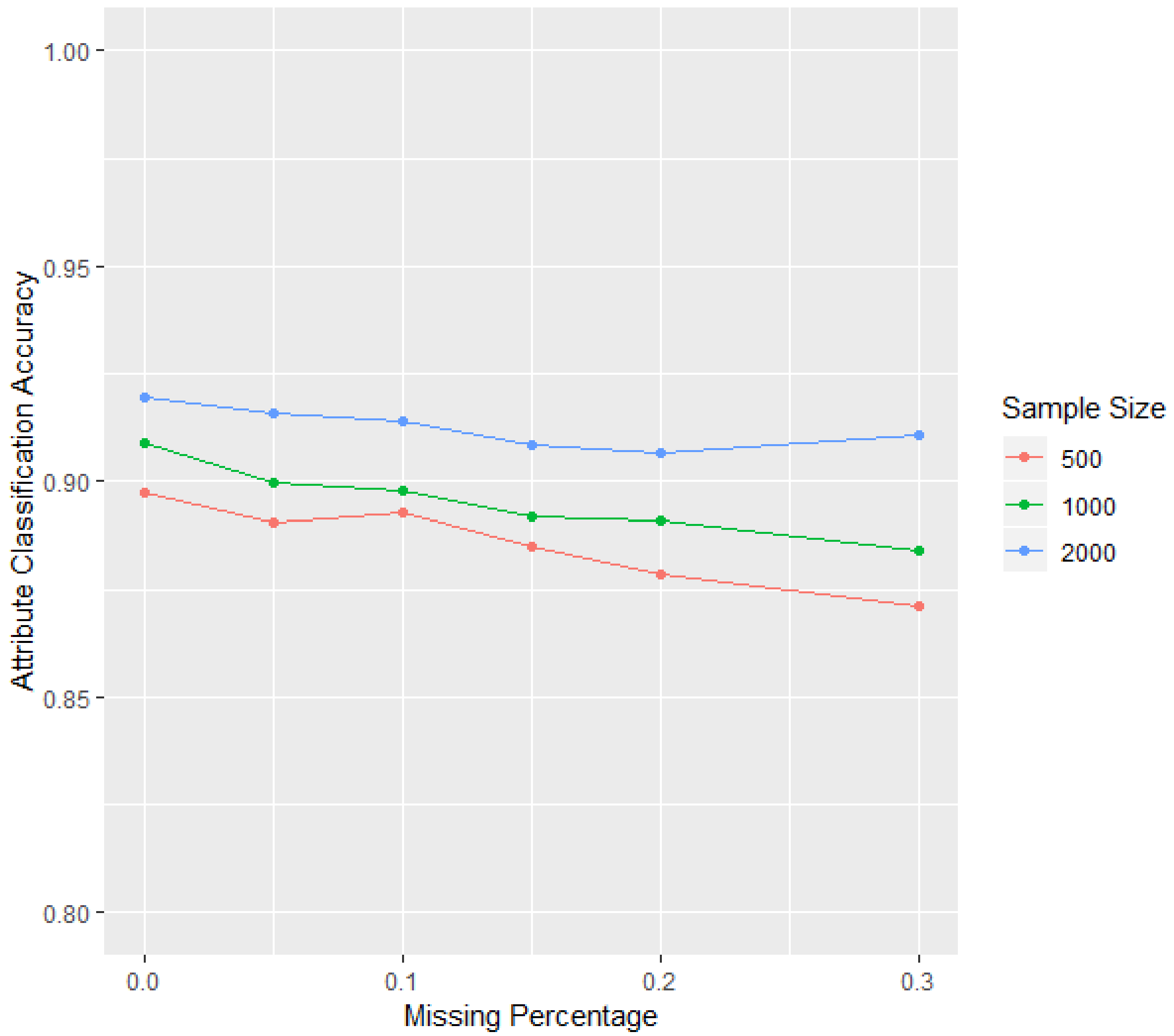

Figure 14. Attribute classification accuracy across sample sizes under ALL-3 Q-matrix design.

Profile classification accuracy with ALL-3. Factorial ANOVA results showed that all effects were statistically significant: the missing percentage [F (5, 4368) = 62.504, p < .001], the sample size [F (2, 4368) = 137.047, p < .001], the missing data mechanism [F (2, 4368) = 12.424, p < .001], the two-way interaction between missing percentage and sample size [F (10, 4368) = 2.692, p < .01], the two-way interaction between missing percentage and missing data mechanism [F (8, 4368) = 2.796, p < .01], and the two-way interaction between sample size and missing data mechanism [F (4, 4368) = 5.403, p< .001], and the three-way interaction [F (16,

4368) = 5.697, p < .001]. Three effects had nonnegligible effect sizes: missing percentage (η^2=0.061), sample size (η^2=0.053), and the three-way interaction among missing percentage, sample size, and missing data mechanism (η^2=0.018). Figure 15 and Figure 16 illustrate that profile classification accuracy differ across missing data mechanisms and sample sizes.

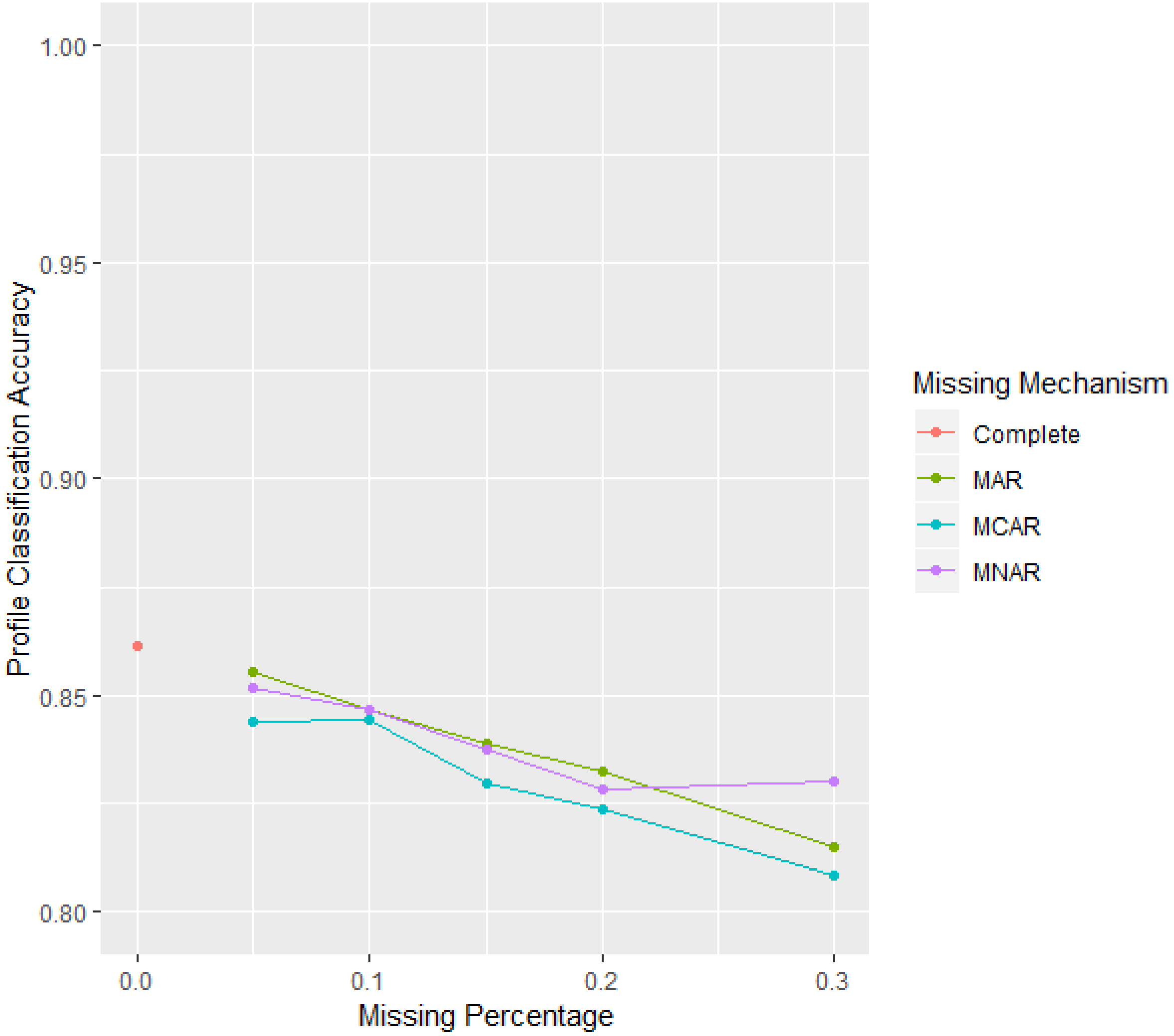

Figure 15. Profile classification accuracy across missing data mechanisms under ALL-3 Q-matrix design.

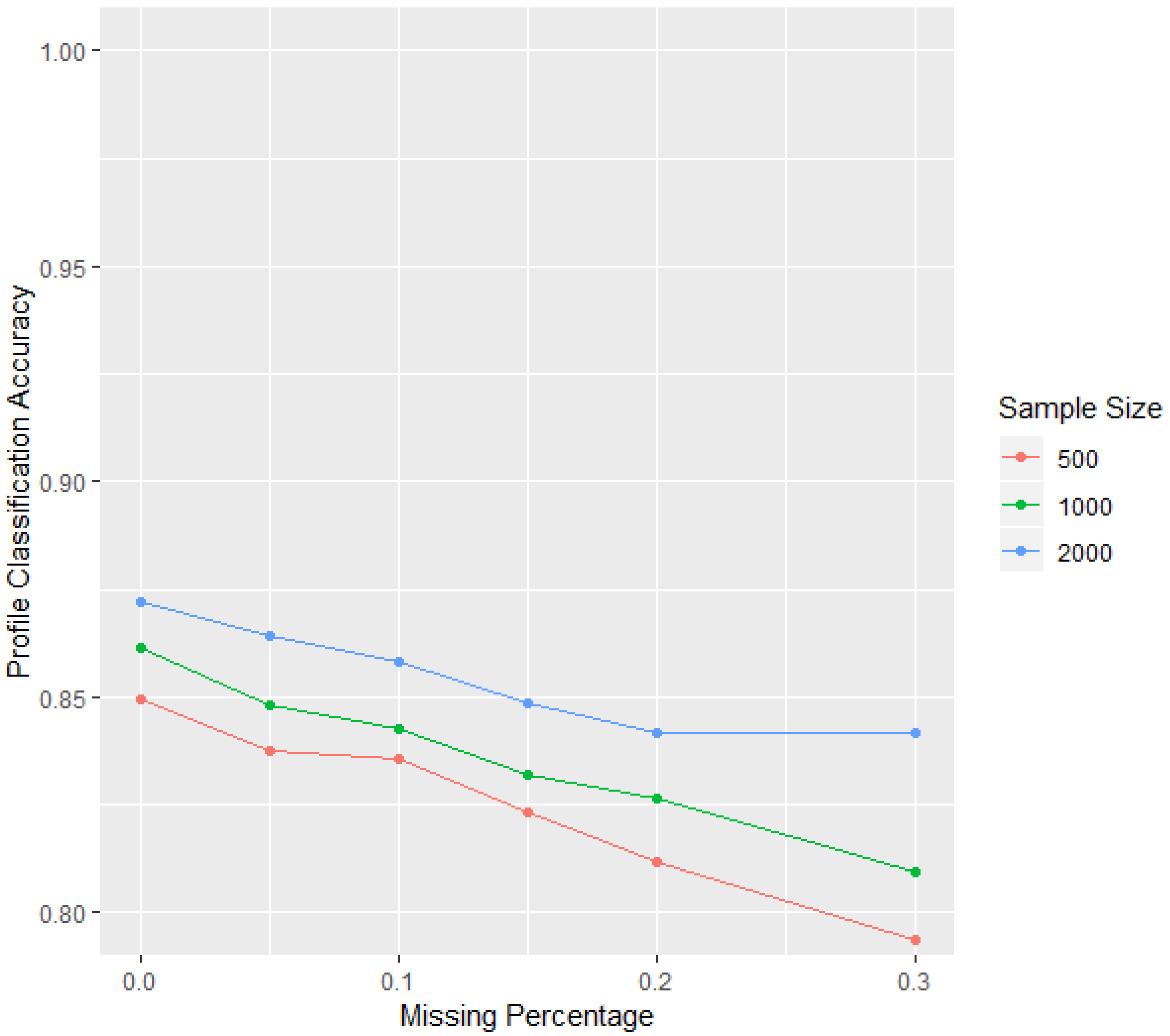

Figure 16. Profile classification accuracy across sample sizes under ALL-3 Q-matrix design.

Attribute reliability with ALL-3. Factorial ANOVA results showed that the missing percentage [F (5, 4368) = 26.535, p < .001] and the missing data mechanism [F (2, 4368) = 7.632, p < .001] are statistically significant, while only the missing percentage (η^2=0.029) was an influential factor. Figure 17 and Figure 18 illustrate that attribute reliability did not differ much across missing data mechanisms and sample sizes.

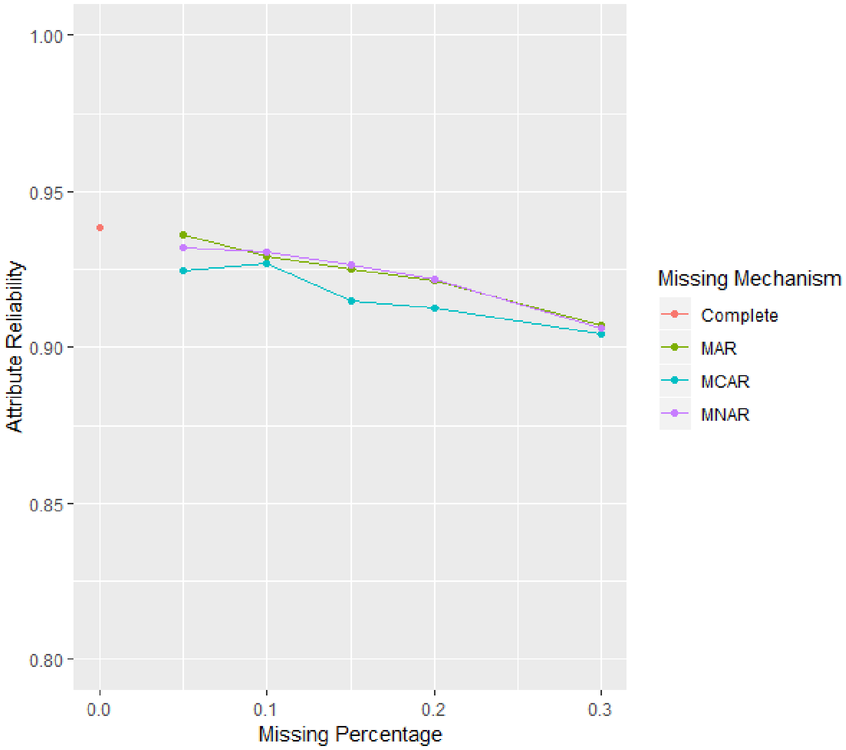

Figure 17. Attribute reliability across missing data mechanisms under ALL-3 Q-matrix design.

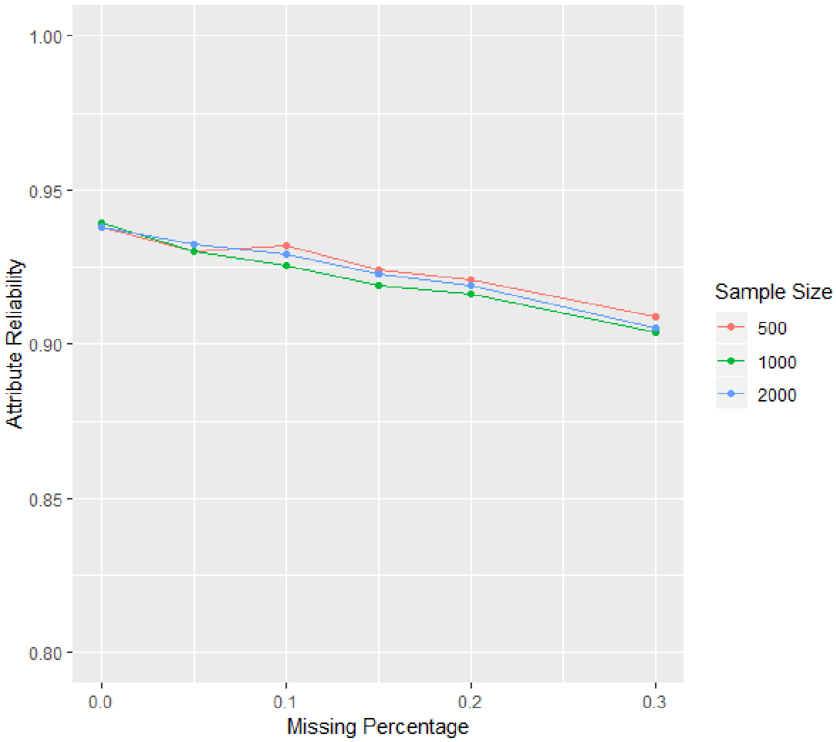

Figure 18. Attribute reliability across sample sizes under ALL-3 Q-matrix design.

Summary. For both PAIR-3 and ALL-3 Q-matrix designs, missing percentage was influential on all three measures, sample size was influential on attribute classification accuracy and profile classification accuracy. A general trend was that as the missing percentage increased and the sample size decreased, the accuracy measures decreased. Similar phenomena with BAL-3 on attribute reliability were observed and are discussed in the later chapter.

Research Question 3

Between the two alternative Q-matrix designs, which one obtains higher attribute classification accuracy, profile classification accuracy, and attribute reliability?

PAIR-3 consistently obtained higher measurements than ALL-3 did with Cohen's *d* for attribute classification accuracy of 0.30, Cohen's *d* for profile classification accuracy of 0.25, and Cohen's *d* for attribute reliability of 0.23. Boxplots in Figure 19-21 facilitate visualizing the differences between PAIR-3 and ALL-3 for all three measures across sample size. As shown in the figure, there are long trailing tails toward the low values on all three measures, and further investigation indicates that low classification accuracy measures are associated with small sample size and that low reliability is associated with both high missing percentage and small sample size.

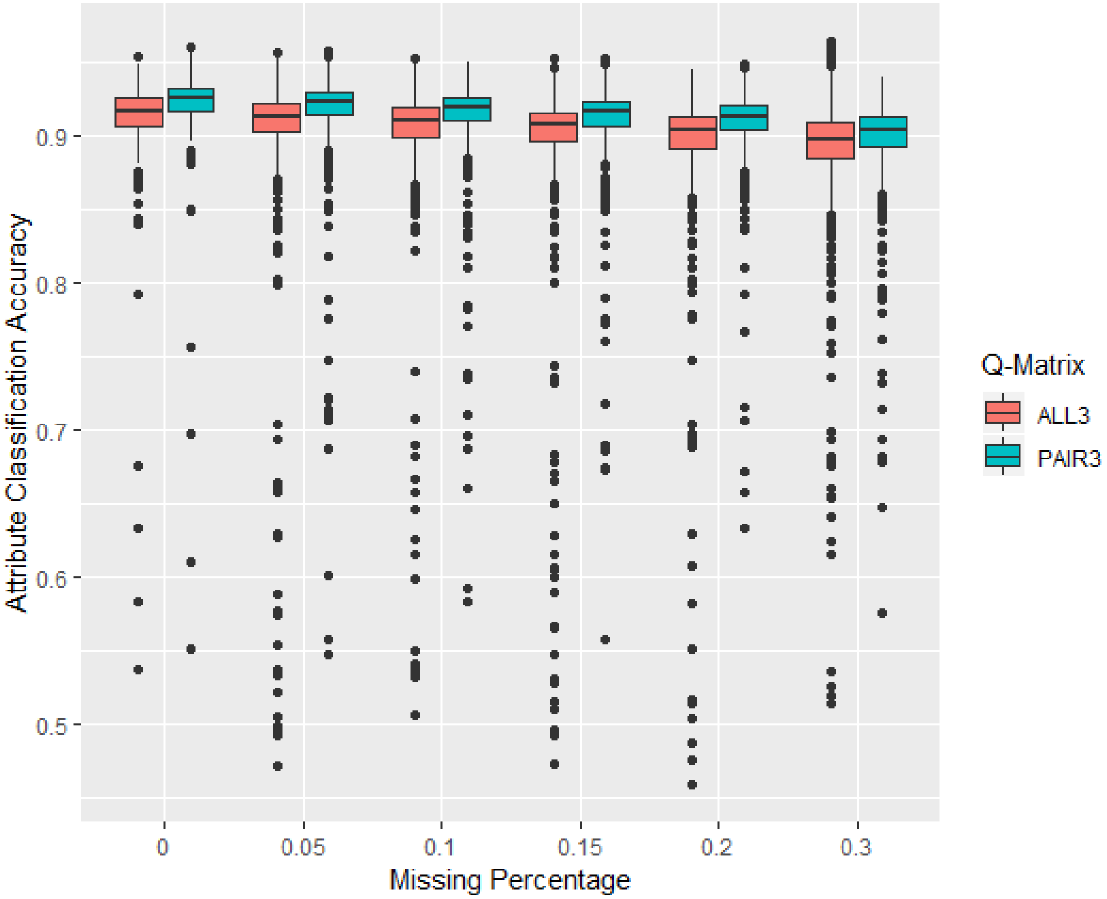

Figure 19. Boxplot of attribute classification accuracy by Q-matrix design.

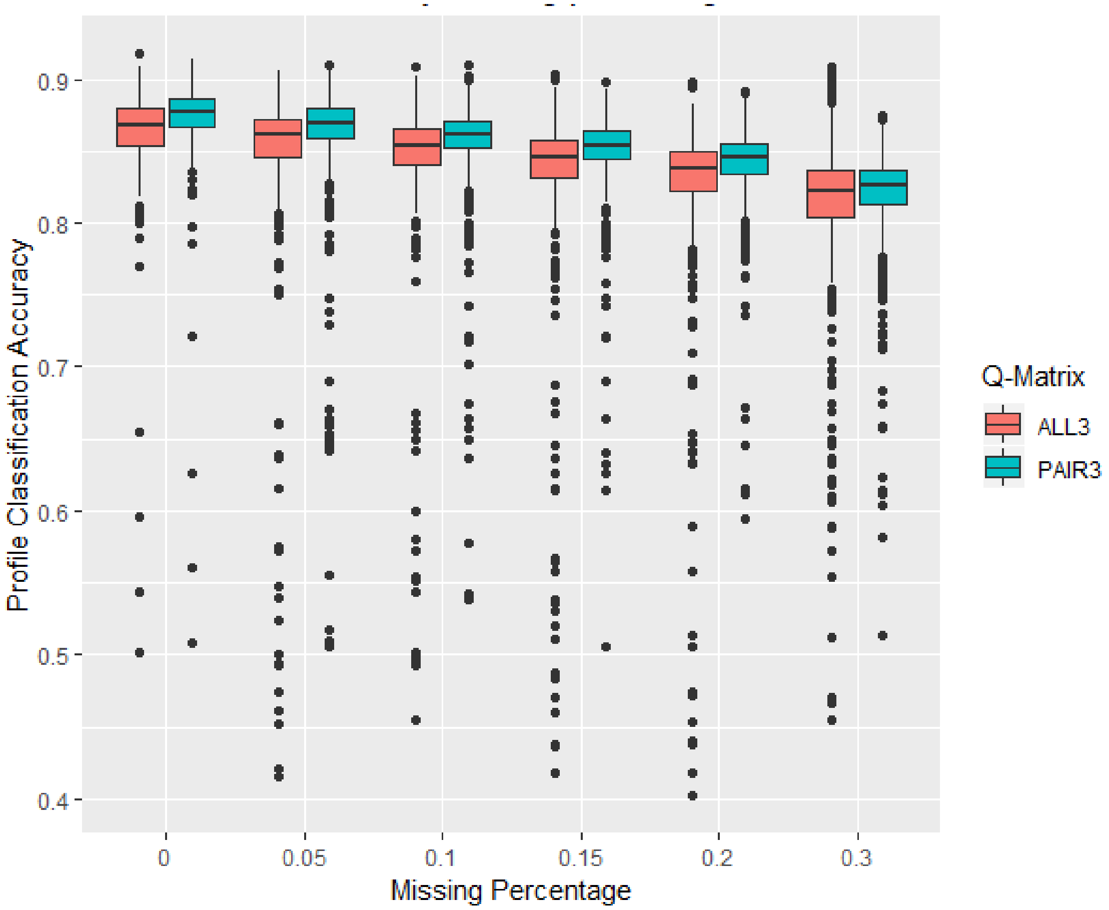

Figure 20. Boxplot of profile classification accuracy by Q-matrix design.

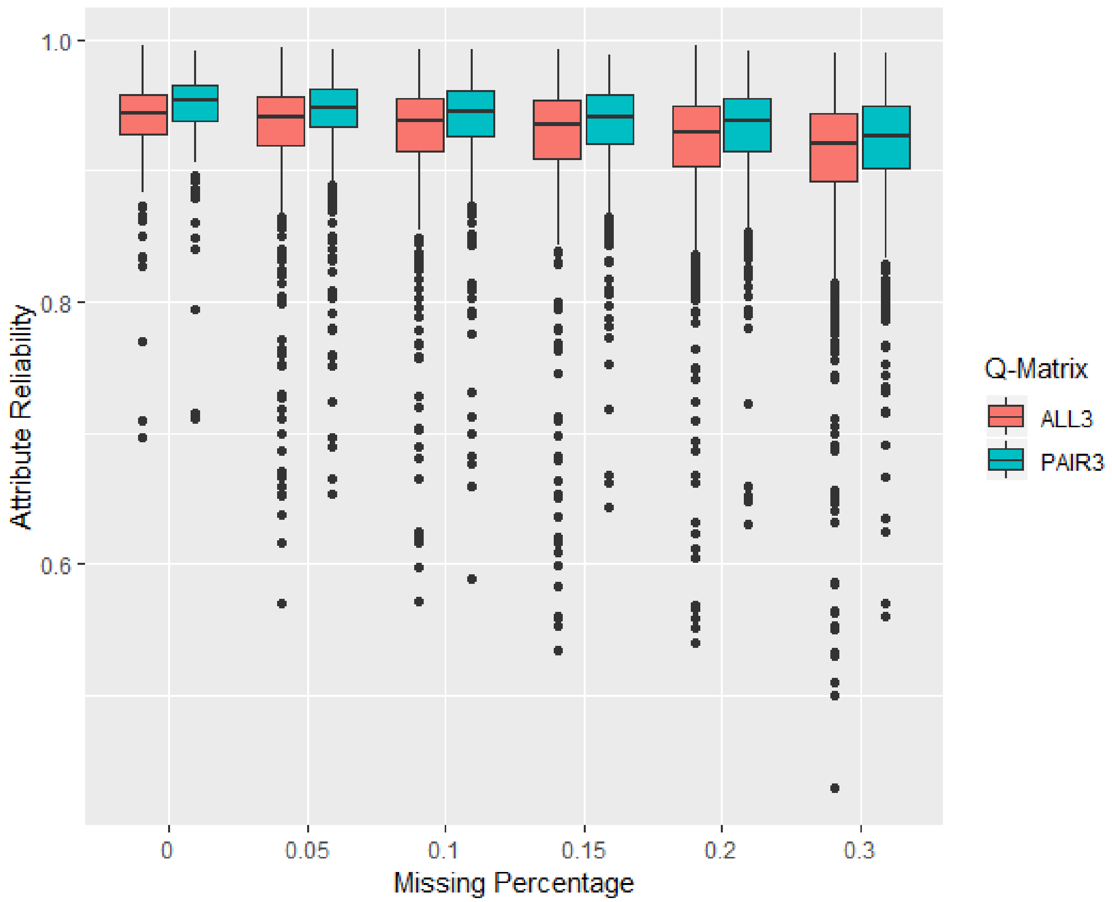

Figure 21. Boxplot of attribute reliability by Q-matrix design.

CHAPTER 5

Discussion

Despite the great potential of DCMs to give detailed feedback to individual examinees, a variety of methodological issues and practical issues stand in the way of realizing this potential. This study sought to inform practitioners regarding the effects of missing responses on classification accuracy and attribute reliability. Depending on how the diagnostic information is used, inaccurate mastery classification likely leads to wrong diagnoses and treatment prescriptions in psychological settings and wasted time and effort in studying already mastered attributes in educational settings. For example, if a questionnaire measuring three symptoms is used to diagnose a certain mental illness, which is analogous to an assessment with three attributes, accurately classifying whether each symptom exists will lead to a decision as to whether someone has the mental illness. A false negative classification error may cause a person who needs to be treated be ignored, and a false positive classification error may cause a person who does not need treatment be given treatments. This section discusses possible reasons for the observed phenomenon with attribute reliability and provides recommendations for handling missing responses, Q-matrix design, and future research.

Attribute Reliability

The results of the analyses showed a consistent pattern of attribute reliability being associated with small sample size, which is contrary to common belief that large sample size provides more accurate estimates and leads to higher reliability. In this context reliability is a measure of the consistency of the classification decisions, and it is related to test length, base rates, attribute correlation, and sample size (J. Templin, personal communication, Sep. 23, 2019).

To test if bias could have potentially caused the observed phenomenon, attribute reliability with known item parameters for complete data sets were calculated (Table 8).

Table 8

Comparison of Average Reliability with Known and Estimated Item Parameters

Sample Size	Average Theoretical Reliability Based on Specified Parameters	Average Reliability (Converged Data Sets)
500	.827	.830
1,000	.827	.828
2,000	.828	.828

Note. All analyses with known item parameters converged, and the third column only included analyses results of which the analyses with unknown item parameters reached convergence.

From Table 8, although the smallest sample size is associated with high reliability in the third column, it is the reverse in the second column, indicating that the nonconvergence could cause the observed phenomenon that small sample sizes are associated with high reliability estimates. Because the convergence rate is lower with smaller sample sizes, we may expect inflated reliability for small sample sizes. The nonconvergence could also explain the nonlinear trend we observe in for attribute reliability for PAIR-3 and ALL-3 matrices (see Figure 12 and Figure 18).

Another reason that could explain the large inflation with small sample size is bias of the attribute reliability. Johnson and Sinharay (2019) indicated that larger positive bias of attribute reliability was associated with smaller sample size, which means attribute reliability gets inflated to a greater degree for small sample size than for large sample size. Comparing the values in Table 8 with the average attribute reliability obtained from the simulation, which is .950 for sample size of 500, .946 for sample size of 1,000, and .946 for sample size of 2,000, we could observe large inflation with small sample sizes. Therefore, attribute reliability bias could be another reason for higher attribute reliability with small sample size.

Research has not ceased to define reliability for DCMs and explore the properties of each of the definitions (Johnson & Sinharay, 2018, 2019). Future research could be conducted with other reliability measures in the context of missing responses.

Recommendations for Handling Missing Responses

The results for balanced Q-matrix show that the missing percentage was the most influential factor on attribute classification accuracy, profile classification accuracy, and reliability, and that both missing percentage and sample size are influential on the three measures for alternative Q-matrix designs. Therefore, the percentage of missing responses should be taken into consideration for the estimation process. It is observed that even with complete data, the measures decreased from BAL-3 to PAIR-3 to ALL-3, and that profile classification accuracy was lower than attribute classification accuracy across all conditions. Despite these trends across Q-matrices, sample sizes, and percentages of missing responses, it was noticeable that with 5% missing responses, the measures were close to the complete data analyses. Nevertheless, if the percentage of missing is greater than 5%, classification accuracy measures are noticeably divergent from the accurate classification. This diagnostic information could lead to inaccurate feedback to more examinees. Therefore, I recommend looking into other missing data treatments such as imputation to improve classification accuracy and attribute reliability with the percent of missing greater than 5%, rather than relying on maximum likelihood estimation.

Recommendations for Q-Matrix Design

Madison and Bradshaw (2015) have already recommended a balanced Q-matrix design, and this study is seeking to help researchers make informative decisions on what to do if an attribute can never be measured in isolation with a fixed test length. From research question 3, I recommend PAIR-3 over ALL-3 design for attributes that cannot be measured in isolation because this design obtains higher classification accuracy measures and attribute reliability, and the convergence rate is higher.

It is important to reiterate the assumptions about the attribute and the design of the Q-matrix to provide context for the recommendation so that researchers and practitioners may make their own judgments. It was assumed that one of the three measured attributes could not be measured alone in a test item and that this attribute could be measured in combination with any of the other attributes. Under these assumptions, and after following the guidelines of Madison and Bradshaw (2015) to fill a 20-item test with all combinations three times, there were two spare items. PAIR-3 and ALL-3 differed in those two items: the two items in PAIR-3 measured the FA and one other attribute, while the two items in ALL-3 measured all three attributes. The study focused on manipulating the remaining items for Q-matrix design after populating the matrix with all attribute combinations, and PAIR-3 was superior.

While there was a FA in this study, attribute classification accuracy and attribute reliability are higher for PIAR-3 than for ALL-3. The recommendation could be different if another attribute that could be measured in isolation were the focal attribute, if the last two items in PAIR-3 measured attributes other than the FA, or if a different DCM were used. These questions need to be answered by future research, and researchers and practitioners should make their decision based on their purpose.

Limitations

The way the MAR condition was simulated in this study was related to the estimated attribute mastery which reflected how the correctness of the responses was generated. Therefore, this way of simulating the MAR condition might have been confounded with the MNAR condition, which was simulated based on the correctness of the responses. This way of simulating the MAR condition was not employed by previous researchers. Dai (2017) and Dai et al. (2018) used a hypothetical continuous variable to create fractiles to determine the probability of missingness for each respondent, and Sünbül (2017) used the total score. It follows that the influence of the particular way of simulating the MAR condition on the results is hard to pinpoint.

As with all simulation studies, although the purpose was to generalize, there were limitations to how far the generalization would go. In this study, the specific simulation conditions, including model selection, Q-matrix design, base rates, attribute correlations, master's probability for correct responses, and sample size, led to the observed results and consequently the recommendations. Although the simulation conditions were intended to imitate practical scenarios, these conditions may confine how good the recommendations would be in specific situations.

Implications for Test Developers

Although diagnostic assessments have the advantage of producing accurate classification information with short test, test developers are encouraged to set reasonable expectations of how long the test should be and how many times each attribute is measured. From Figure 1 to Figure 18, it is observed that with the increase of the missing percentage, all three dependent variables (attribute classification accuracy, profile classification accuracy, and attribute reliability) decreased. However, all the aggregated values in those figures were no less than .75. The positive results were obtained probably due to the length of the test. In the 20-item test, the FA was measured 11 times in all matrices. Allowing the attribute to be measured sufficient times could be a vital factor in diagnostic assessment design to achieve optimal person parameters. High missing percentage (30%) could play a more detrimental role in a shorter test in terms of classification accuracy and attribute reliability. Test developers should do their best to adhere to the suggestions from Madison and Bradshaw (2015) to ensure that each attribute is measured sufficient times.

Implications for Future Research

I suggest future researchers use other ways to simulate the MAR condition so that the effect of missing data mechanism could be further explored. Although the current study showed that missing data mechanisms (MCAR, MAR, and MNAR) were not influential in (a) attribute classification accuracy, (b) profile classification accuracy, and (c) attribute reliability, the results might not hold using different ways to simulate the MAR condition. In the current study, the MAR condition was related to the estimated attribute mastery from the complete data, and the MNAR condition was related to the correctness of the response. Although the MAR condition did not directly relate to the correctness, attribute mastery was estimated from the data, and therefore, the MAR condition was indirectly related to the MNAR condition. Given this logic, the differences among the three missing data mechanisms could have been disguised. As Zhang (2014) identified that the missingness related to the individual respondent's characteristics, future research could employ how Dai (2017) simulated the MAR condition and generate a hypothetical variable to represent the characteristics of individual respondents, and have that predict the missingness.

Additionally, future researchers could incorporate levels of association of the missing data in MAR and MNAR conditions and explore how the strength of the association plays a role in the estimation process. For example, for the MNAR condition, the small association could be that the incorrect responses have a low probability of being missing such as drawing from U (0,0.2), and the large association could be that the incorrect responses have a high probability of being missing such as drawing from U (0.6, 0.9).

As previous section mentioned, the interpretations of the results and the recommendations were limited to the simulation conditions in this study. To explore how well these recommendation holds, these conditions could be altered to further study the research questions.

Finally, this study only investigated person parameters, and I suggest future researchers explore the effects of Q-matrix design, sample size, missing data mechanism, and missing percentage on the bias of item parameters. While person parameters are important to investigate because these parameters are the bases of the diagnostic results, item parameters are important for calibration purposes so that these estimates would be used as known parameters for future assessments. Future researchers are suggested to bring this research topic to the realm of item parameters.

Conclusion

This study explored Q-matrix design and maximum likelihood estimation for the C-RUM using simulations. A total of 144 simulation conditions were executed, including three Q-matrices (BAL-3, PAIR-3, and ALL-3), three sample sizes (500; 1,000; and 2,000), six missing percentages (0%, 5%, 10%, 15% 20%, 30%), and three missing data mechanisms (MCAR, MAR, and MNAR). The results showed that PAIR-3 was superior than ALL-3 in that it obtained higher attribute classification accuracy, profile classification accuracy, and attribute reliability, and the results also indicated that maximum likelihood estimation could handle the percentage of missing less than 5%. It is recommended that with a fixed length test, practitioners could fill up the matrix with items measuring paired attributes instead of items measuring all attributes, and that more advanced missing data treatments could be sought if the percentage of missing is larger than 5%.

REFERENCES

Ayers, E., Nugent, R., & Dean, N. (2009). A comparison of student skill knowledge estimates. In T. Barnes, M. Desmarais, C. Romero, & S. Ventura (Eds.), *Educational data mining 2009: 2nd international conference on educational data mining* (pp.1-10). Retrieved from https://pdfs.semanticscholar.org/d1e0/d84dc6319175cf1751d5d73fc4b0931063c9.pdf

Baraldi, A., & Enders, C. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*, 5-37. doi: 10.1016/j.jsp.2009.10.001

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling, 15*, 651-675. doi: 10.1080/10705510802339072

Bowen, N. K. (2015). *Quick guide for using Mplus: Essentials for getting started with Mplus*. Retrieved from http://global.oup.com/us/companion.websites/9780195367621/pdf/MplusQuickGuide2015.pdf

Bradshaw, L. (2017). Diagnostic classification models. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 297-327). Hoboken, NJ: Wiley-Blackwell.

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice, 33*(1), 2-14. doi: 10.1111/emip.12020

Chen, H., & Chen, J. (2016) Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly, 13*, 218-230, doi: 10.1080/15434303.2016.1210610

Chiu, C. Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*, 633-665. doi: 10.1007/s11336-009-9125-0

Chiu, C. Y. & Köhn, H. F. (2016). The reduced RUM as a logit model: Parameterization and constraints. *Psychometrika, 81*, 350-370. doi: 10.1007/s11336-015-9460-2

Dai, S. (2017). *Investigation of missing responses in implementation of cognitive diagnostic models* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global (Order No. 10606810).

Dai, S., Svetina, D., & Chen, C. (2018). Investigation of missing responses in Q-matrix validation. *Applied Psychological Measurement, 42*, 660-676. doi: 10.1177/0146621618762742

De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement, 38*, 213–234. doi: j.1745-3984.2001.tb01124.x

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179-199. doi: 10.1007/S11336-011-9207-7

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35*, 8-26. doi: 10.1177/0146621610377081

DiBello, L. & Stout, W. (2008). Arpeggio Documentation and Analyst Manual (Ver.3.1.001) [Computer software]. Applied Informative Assessment Research Enterprises (AIARE)—LLC. St. Paul: MN: Assessment Systems Corporation.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling, 8*, 128-141. doi: 10.1207/S15328007SEM0801_7

Enders, C. K. (2008). A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. *Structural Equation Modeling, 15*, 434-448. doi: 10.1080/10705510802154307

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*, 225-245. doi: 10.1111/j.1745-3984.2008.00062.x

García, P. E., Olea, J., & de la Torre, J. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psicothema, 26*, 372-377. doi: 10.7334/psicothema2013.322

Gu, Z. (2012). *Maximizing the potential of multiple-choice items for cognitive diagnostic assessment* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global; Social Science Premium Collection. (Order No. NR78204)

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

Hallquist, M. N. & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling, 25*, 621-638. doi: 10.1080/10705511.2017.1402334.

Hansen, M. P. (2013). *Hierarchical item response models for cognitive diagnosis* (Doctoral dissertation). Retrieved from ProQuest. (UMI 3567806)

Harrison, A. J., Bradshaw, L. P., Naqvi, N. C., Paff, M. L., & Campbell, J. M. (2017). Development and psychometric evaluation of the autism stigma and knowledge questionnaire (ASK-Q). *Journal of Autism and Developmental Disorders, 47*, 3281-3295. doi: 10.1007/s10803-017-3242-x

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

Hartz, S., & Roussos, L. (2008). *The fusion model for skills diagnosis: Blending theory with practicality* (Report No. RR-08-17). Princeton, NJ: Educational Testing Service. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.2008.tb02157.x

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*, 262-277. doi: 10.1177/0146621604272623

Henson, R., Templin, J. L., & Douglas, J. (2007). Using efficient model based sum-scores for conducting skills diagnoses. *Journal of Educational Measurement, 44*, 361–376. doi: 10.1111/j.1745-3984.2007.00044.x

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. doi: 10.1007/S11336-008-9089-5

Huisman, M., & Molenaar, I. W. (2001). Imputation of missing scale data with item response models. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 221–244). New York, NY: Springer.

Jang, E. E. (2009a). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to LanguEdge assessment. *Language Testing, 26*, 31-73. doi: 10.1177/0265532208097336

Jang, E. E. (2009b). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly, 6*, 210-238. doi: 10.1080/15434300903071817

Jang, E. E., Dunlop, M., Wagner, M., Kim, Y. H., & Gu, Z. (2013). Elementary school ELLs' reading skill profiles using cognitive diagnosis modeling: Roles of length of residence and home language environment. *Language Learning, 63*, 400-436. doi: 10.1111/lang.12016

Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement, 55*, 635-664. doi: 10.1111/jedm.12196

Johnson, M. S., & Sinharay, S. (2019). The reliability of the posterior probability of skill attainment in diagnostic classification models. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/1076998619864550

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. doi: 10.1177/01466210122032064

Jurich, D. P., & Bradshaw, L. P. (2014). An illustration of diagnostic classification modeling in student learning outcomes assessment. *International Journal of Testing, 14*, 49-72. doi: 10.1080/15305058.2013.835728

Kim, A. Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*, 227-258. doi: 10.1177/0265532214558457

Köhn, H. F. & Chiu, C. Y. (2016). A proof of the duality of the DINA model and the DINO model. *Journal of Classification, 33*, 171-184. doi: 10.1007/s00357-016-9202-x

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation, 35*, 64-70. doi: 10.1016/j.stueduc.2009.10.003

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. doi: 10.1111/j.1745-3984.2011.00160.x

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2017). Incremental validity of multidimensional proficiency scores from diagnostic classification models: An illustration for elementary school mathematics. *International Journal of Testing, 17*, 277-301. doi: 10.1080/15305058.2017.1291517

Lang, K. M., & Little. T. D. (2014) The supermatrix technique: A simple framework for hypothesis testing with missing data. *International Journal of Behavioral Development, 38*, 461-471. doi: 10.1177/0165025413514326

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

Li, H., & Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment, 18*, 1-25. doi: 10.1080/10627197.2013.761522

Li, H., Hunter, C. V., & Lei, P. W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing, 33*, 391-409. doi: 10.1177/0265532215590848

Little, R. J., & Rubin, D. B. (2002). Bayes and multiple imputation. *Statistical analysis with missing data* (2nd ed., pp. 200-220). Hoboken, NJ: John Wiley & Sons, Inc.

Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement, 75*, 491-511. doi: 10.1177/0013164414539162

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74,* 525-556. doi: 10.3102/00346543074004525

Pichette, F., Béland, S., Jolani, S., & Leśniewska, J. (2015). The handling of missing binary data in language research. *Studies in Second Language Learning and Teaching, 5,* 153-169. doi: 10.14746/ssllt.2015.5.1.8

R Core Team (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment, 34*, 782-799. doi: 10.1177/0734282915623053

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement, 44*, 293-311. doi: 10.1111/j.1745-3984.2007.00040.x

Rubin, D. B. (1976). Inference and missing data. *Biometricka, 63*, 581-592.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*, 219-262. doi: 10.1080/15366360802490866

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

Savalei, V. (2008). Is the ML chi-square ever robust to nonnormality? A cautionary note with missing data. *Structural Equation Modeling, 15*, 1-22. doi: 10.1080/10705510701758091

Schafer, J. L., & Graham, J. W. (2002). Missing data. Our view of the state of the art. *Psychological Methods, 7*, 147–177. doi: 10.1037//1082-989X.7.2.147

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*, 545-571. doi: 10.1207/s15327906mbr3304_5

Sheehan, K. M., Tatsuoka, K. K., & Lewis, C. (1993). *A diagnostic classification model for document processing skills* (Report No. 93-39-ONR). Princeton, NJ: Educational Testing Service. Retrieved from https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1993.tb01550.x

Sijtsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research, 38*, 505-528. doi: 10.1207/s15327906mbr3804_4

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the Analysis of missing data. *Psychological Methods, 6*, 317-329. doi: 10.1037//1082-989X.6.4.317

Skaggs, G., Wilkins, J. L., & Hein, S. F. (2016). Grain size and parameter recovery with TIMSS and the general diagnostic model. *International Journal of Testing, 16*, 310-330. doi: 10.1080/15305058.2016.1145683

Song, X. Y., & Lee, S. Y. (2006). A maximum likelihood approach for multisample nonlinear structural equation models with missing continuous and dichotomous data. *Structural Equation Modeling, 13,* 325-351. doi: 10.1207/s15328007sem1303_1

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods, 19*, 506-532.doi: 10.1177/1094428116630065

Sünbül, S. Ö. (2017) The impact of different missing data handling methods on DINA model. *International Journal of Evaluation and Research in Education, 7*, 77-86. doi: 10.11591/ijere.v1i1.11682

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

Templin, J. (2016, November 1). R functions for estimation of the LCDM in Mplus. Retrieved from https://jonathantemplin.com/functions-estimation-lcdm-mplus/

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*, 317-339. doi: 10.1007/S11336-013-9362-0

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*, 251-275. doi: 10.1007/s00357-013-9129-4

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305. doi: 10.1037/1082-989X.11.3.287

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice, 32*(2), 37-50. doi: 10.1111/emip.12010

van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2007). Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results, *Multivariate Behavioral Research, 42*, 387-414. doi: 10.1080/00273170701360803

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Report No. RR-05-16). Princeton, NJ: Educational Testing Service. Retrieved from https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.2005.tb01993.x

Xin, T., & Zhang, J. (2015). Local equating of cognitively diagnostic modeled observed scores. *Applied Psychological Measurement, 39*, 44-61. doi: 10.1177/0146621614542427

Yi, Y. (2012). *Implementing a cognitive diagnostic assessment in an institutional test: A new networking model in language testing and experiment with a new psychometric model and task type*. (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (Order No. 3633779)

Zhang, J. (2014). *Relationships between missing responses and skill mastery profiles of cognitive diagnostic assessment* (Doctoral dissertation). Retrieved from ProQuest. (UMI AAINR96139)

APPENDIX A

Balanced Q-Matrix BAL-3

Item	FA	Attribute 2	Attribute 3
1	1		
2		1	
3			1
4	1	1	
5		1	1
6	1		1
7	1	1	1
8	1		
9		1	
10			1
11	1	1	
12		1	1
13	1		1
14	1	1	1
15	1		
16		1	
17			1
18	1	1	
19		1	1
20	1		1

Note. Entries of 0 were removed for the ease of view.

APPENDIX B

Unbalanced Q-Matrix PAIR-3

Item	FA	Attribute 2	Attribute 3
1		1	
2			1
3	1	1	
4	1		1
5		1	1
6	1	1	1
7		1	
8			1
9	1	1	
10	1		1
11		1	1
12	1	1	1
13		1	
14			1
15	1	1	
16	1		1
17		1	1
18	1	1	1
19	1		1
20	1	1	

Note. Entries of 0 were removed for the ease of view.

APPENDIX C

Unbalanced Q-Matrix ALL-3

Item	FA	Attribute 2	Attribute 3
1		1	
2			1
3	1	1	
4	1		1
5		1	1
6	1	1	1
7		1	
8			1
9	1	1	
10	1		1
11		1	1
12	1	1	1
13		1	
14			1
15	1	1	
16	1		1
17		1	1
18	1	1	1
19	1	1	1
20	1	1	1

Note. Entries of 0 were removed for the ease of view.

APPENDIX D

MAR Data Set Generation Procedure

This appendix gives an example of how the MAR data sets were generated in the study. The Q-matrix and estimated marginal probabilities of attribute mastery are created to illustrate this procedure. Table D1 shows a hypothetical matrix with three items. Each item measures different combinations of three attributes. Table D2 shows the estimated marginal attribute mastery probabilities from the complete data.

Table D1

Q-Matrix

Item	Attribute 1	Attribute 2	Attribute 3
1	1	1	0
2	0	1	0
3	1	1	1

Table D2

Estimated Marginal Probabilities of Attribute Mastery

Respondent	Attribute 1	Attribute 2	Attribute 3
1	.2	.1	.5
2	.6	.1	.1
3	.3	.4	.5

Information from Table D1 and Table D2 are then used to generate an initial matrix that will indicate the missingness of each response. For example, item 1 measures two attributes, and the marginal probability of mastering the two attributes for respondent 1 respectively are 0.2 and 0.1, the corresponding cell in the missing propensity matrix would be 1.7 (calculated by 1-0.2+1-0.1) (Table D3). Because the propensity matrix contained values larger than 1, it was transformed so that all values were between 0 and 1 (Table D4). The transformed propensity matrix is then used in combination with the set missing percentage to generate missing responses.

Table D3

Initial Missing Propensity Matrix

Respondent	Item 1	Item 2	Item 3
1	.8+.9=1.7	1-.1=0.9	.8+.9+.5=2.2
2	.4+.9=1.3	1-.1=0.9	.4+.9+.9=2.2
3	.7+.6=1.3	1-.4=0.6	.7+.6+.5=1.8

Table D4

Final Propensity Matrix

Respondent	Item 1	Item 2	Item 3
1	0.69	0.19	1
2	0.44	0.19	1
3	0.44	0	0.75