DESIGNING FIELD TESTS FOR MULTIDIMENSIONAL CLASSIFICATION MODELS

by

SELAY ZOR

(Under the Direction of Laine Bradshaw)

Diagnostic classification models (DCMs) are multidimensional latent variable models that can provide diagnostic information about the mastery state of examinees' knowledge components (Rupp, Templin & Henson, 2010). DCMs classify examinees based on specified knowledge components, and provide multidimensional feedback about examinees' strength and weaknesses. However, recent large-scale assessments have not designed to diagnose, and few practical applications of DCMs exists. Creating multidimensional assessments is needed to meet demands of more detailed feedback. It introduces new challenges to educational assessment research. Field testing is an essential step in creating assessments. Field testing items for unidimensional vs. multidimensional assessments are not the same. Different field test designs result in sparse data, and research has not studied sparse data conditions for DCMs. I investigate the impact of sparse data, due to different field test designs, on the estimation accuracy for DCMs. Results provide needed guidelines for designing DCM-based field tests.

INDEX WORDS: Diagnostic classification models, Log-linear cognitive diagnosis model, field-testing, item tryouts, planned missing data design

DESIGNING FIELD TESTS FOR MULTIDIMENSIONAL CLASSIFICATION MODELS

by

SELAY ZOR

B.S., Balikesir University, Turkey, 2013

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF ARTS

ATHENS, GEORGIA

2018

DESIGNING FIELD TESTS FOR MULTIDIMENSIONAL CLASSIFICATION MODELS

by

SELAY ZOR

Major Professor:	Laine Bradshaw
Committee:	Amanda E. Ferster
	Shiyu Wang

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2018

DEDICATION

To my family..

ACKNOWLEDGEMENTS

First, and most of all, I would like to thank my adviser Dr. Laine Bradshaw, for her expertise, guidance, patience and encouragement throughout the process of writing this thesis. Without her help and support this thesis would not have been possible. I would like to thank my committee members, Dr. Amanda Ferster and Dr. Shiyu Wang, for their suggestions and feedback.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

CHAPTER 1

INTRODUCTION

Diagnostic classification models (DCMs) are multidimensional latent variable models that can provide categorical feedback about examinees' knowledge states (Rupp and Templin, 2008). In recent years, DCMs have received increased attention among educational researchers and psychometricians. These models are suitable to analyze the multidimensional content of various assessment tools to provide information about examinees' knowledge components. Although DCMs have been proposed as viable models for practice (e.g., Bradshaw et al. 2014; Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010), few practical applications of these models exist. Mostly, data from large-scale assessments that were designed as unidimensional have been used for applications of DCMs.

Designing multidimensional assessments from the ground up introduces new challenges for educational assessment. One challenge is how to design a field test for gaining initial empirical evidence about the statistical properties and diagnostic quality of newly developed items that comprise the assessment. While field testing is an essential step in creating assessments, field test designs that constructed under planned missingness will yield sparse data conditions that may impose calibration challenges for already-complex multidimensional assessments. The impact of sparse data conditions on estimation for DCMs has not yet been studied. In this study, I will investigate the impact of sparse data on estimation accuracy for DCMs. The planned missing designs will be created based upon different, realistic, field test

designs—those that are relevant to practice. I will seek answers to key questions about designing a field test for a multidimensional diagnostic assessment in practice. These questions include:

(1) How many items, per dimension, does an examinee participating in the field test need to answer to get a stable calibration of the model parameters?

(2) From how many dimensions does an examinee participating in the field test need to answer items to get stable calibration of the model parameters?

(3) How many responses per item are required for stable calibration under various testing scenarios?

Based on results from a set of simulation studies, I will provide recommendations to practitioners for how to design field tests that will yield sufficiently accurate and precise estimates of item parameters, attribute levels, and attribute correlations.

CHAPTER 2

LITERATURE REVIEW

This investigation of field testing in DCMs is based on literature in field testing, missing data, and diagnostic classification models. I begin with a description of field testing and methods to conduct a field test. Since field testing results in missing data, I, then, review missing data mechanisms and planned missing data designs. In the last section of this chapter, I introduce DCMs and describe a general DCM, the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009), in detail.

FIELD TESTING

Once test items have been developed, reviewed, and revised, they are put through field testing under conditions that reflect the actual test. Field testing is often referred to as *item tryouts.* The main purpose of field testing, in which new items are administered to examinees prior to operational use, is to ensure that items perform as expected. Field testing can indicate that there is a flaw in an item or the item is confusing. After field testing the items, item analysis is conducted to determine statistically effective items. Statistics such as item difficulty and item discrimination are calculated from response data to identify potentially problematic items. If items are either too easy or too difficult, and/or don't discriminate between high and low ability levels, they may be revised or removed. Field testing helps to determine the systematic error within an item. Design decisions, such as revising or removing items, can be made based on the results from field testing to minimize the errors.

Field testing is also essential for generating various forms of the test in static testing or creating the item pool in adaptive testing. Establishing statistical properties of items based on field testing helps to select which items will be used in future operational tests. For static tests, items are placed in an item pool when they are judged to be acceptable, and then items that meet the content and statistical criteria for the tests are then chosen from this pool to construct test forms. Test forms may seek to balance content representation and also statistical properties of the items such as difficulty and discrimination. For adaptive tests, acceptable items are also placed in the item pool, and then items may be selected according to different statistical features according to the algorithm the test is using to select items.

Types of Field Test Designs

Welch (2006) explain two methods to field test items, embedded field testing and standalone field testing. Field tested items do not count toward the examinees' test scores; the results, instead, are used for test development purposes. In an embedded field test, the actual test includes new items. Examinees do not know which items count and which do not. Thus, their motivation to solve the actual items and field test items is expected not to differ. Since the same examinees take the new items and the operational items, this method helps ensure that the sample size is adequate and represents the population well.

While representation and motivation effects are advantages of the embedded field test design, this method may yield some problems. Embedding items within a test either makes the test longer by adding additional items or keeps the test length the same by replacing some of the actual items with field test items. When field test items are added to the test, the overall testing time is expanded. Both time and fatigue effects might negatively affect examinee performance (Kirkpatrick & Way, 2008). The latter method of replacing items may bring up the question

whether the content was covered adequately in the operational test (Wendler & Walker, 2006). Limited content coverage may decrease the amount of information gathered from the operational test. In both situations, since examinees do not know which items will count, they may spend too much time on field test items that are more likely to have confusing wording or some issue that examinees struggle to understand. Moreover, Wendler and Walker (2006) point out that embedded designs on operational assessments, using one item type, do not allow field testing different item types because examinees may recognize the field test items. For example, on an operational assessment using only multiple choice item types, innovative item types such as technology enhanced items (TEIs) may be recognized by examinees as field test items.

In stand-alone field tests, items are included in separate tests that require additional testing sessions. Stand-alone designs allow trying out many items at once because the whole test form the examinee is taking is made up of field test items, in contrast to a few field test slots that would appear on an embedded field test design. When employing this method, however, it might be difficult to motivate examinees. Examinees taking the field test know that results from the test do not count toward their score. Since they know that field test items do not count, they may not try as hard as they would on an actual test, and thus results may be inaccurate. To avoid misleading results, some incentives, such as money, and gift cards, or discounted operational assessment services, may be given to examinees or their schools to increase their motivation.

Another difficulty of administering a stand-alone field test is to find a representative sample for the test-taking population. While some testing companies distribute the field test forms on the same day that examinees take the actual test, most of them administer it other than the actual test day. In both settings, examinees voluntarily attend the field testing sessions so that obtaining an adequate sample size may be more challenging. Because stand-alone field testing

requires a special testing session, additional test forms have to be developed, and often financial incentives are given, it is more expensive than embedded field testing. While developing new assessments that differ from the current assessments, it would be better to administer all new items to examinees via the stand-alone field testing. Since the new assessment would differ according to the content, item types, and other practical issues, stand-alone field testing allow to practice all those differences at once. On the other hand, if new items are required for an existing program, then embedded field test design is an effective way to administer items to examinees.

Examples of Field Test Designs: National and International Testing Programs

Large-scale testing programs such as the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), and The National Assessment of Educational Progress (NAEP) design assessments in such a way that each examinee is administered a specific combination of the items from a larger item pool. Although these large-scale testing programs reuse most of the items from the previous assessments, they update the instruments for each new cycle by adding a percentage of new items. Before new items can appear on the operational versions of the test, the items must be field tested.

The TIMSS and PIRLS both use standalone field tests. They arrange meetings to develop new items based on the subject area, topics in that area, and existing items from the previous assessment. Twice the number of items needed are developed and field tested to ensure the field test yields a sufficient number of items that are adequate for future operational testing. Each year, approximately 40% of the assessment items on the operational test are new items, which were field tested and selected to be used as operational items. Thus, a large number of items are field tested each year. For example, the TIMMS 2015 assessment required developing and field

testing 792 new items in total: 287 items for the fourth grade assessments, 354 items for eight grade assessments, and 151 items for TIMMS Numeracy (Mullis et al., 2016, pp. 1.2). The TIMSS designed the field test in 2015 to be performed in about 30 schools per country yielding at least 200 examinees for each item in each country. Each field test item was evaluated by using approximately 10,000 responses from more than 40 countries. As another example, the PIRLS 2016 assessment developed and field tests 18 new passages and item sets: 12 passages including 203 items for the PIRLS 2016, and 6 passages including 173 items for the PIRLS Literacy (Mullis & Prendergast, 2017, pp. 1.9). Four out of 12 passages were common between PIRLS and PIRLS Literacy. The field testing was conducted in approximately 30 schools in each country. Approximately 9,000 responses to each PIRLS items and 1,000 for each PIRLS Literacy item were collected from all countries to evaluate properties of each item.

While TIMMS and PIRLS conduct standalone field testing, NAEP conducts both standalone and embedded field testing. Mostly, embedded field testing is used by including newly developed items in the NAEP operational tests. NAEP conducts two types of field tests: pilot testing and field testing ("Overview of NAEP Pre-Test Administration Types," 2011). Before items are used in operational NAEP tests, they are pilot testing to obtain information about clarity, difficulty levels, and timing. Field testing is the second step and conducted one year before the operational test. The purpose of it is to improve the analysis of the operational test by pre-calibrating items. Although the information about the number of items pilot testing or field testing is not given, it is reported that approximately 500 examinees are assigned to each pilot item, and 2,000 for each field test item that allow to perform precalibration.

Features of Field Test Designs

A field test can be designed in many ways; however, two main factors play a critical role

in the design: how to sample items and examinees. Sampling of items is required because a single examinee typically cannot complete all of the field test items; nearly always, more items need to be field tested than can be given on a single test form, whether the field test is embedded or standalone. Thus, a key decision is to determine *how many* field test items should be placed on each form; for a standalone field test, this is a question of test length and for an embedded field test, this is a question of how many field test slots will be included on the operational form. Another key decision is *which collection of items* with respect to content should appear on the same form. Greater number of field test slots and longer standalone forms allow each examinee to see a more representative sample of content on their field test and is also more cost effectives because fewer forms need to be developed. With fewer field test slots or shorter standalone forms, the test developer also needs to recruit a larger overall sample of examinees to participate in the field test. These benefits, though, must be weighed with issues of fatigue and motivation discussed above which worsen as the number of field test items the examinee sees increases.

Sampling of examinees is required because as many examinees as are needed to get stable item statistics and model calibration should take the time to complete the field test items, but not more. So all examinees will not need to take all field test items. Thus it is important to figure out *which examinees* will see which items and *how many* examinees should see each item. Field testing the items with a representative sample of the examinee population should give the best predictions about the operational use of items (Welch, 2006). Standard 3.8 of the testing standards by AERA, APA and NCME (2014) emphasizes that the sample should reflect the characteristics of the population from which they are selected.

> When item tryouts or field tests are conducted, the procedures used to select the sample(s) of test takers for item tryouts and the resulting characteristics of the sample(s)

should be documented. When appropriate, the sample(s) should be as representative of the population(s) for which the test is intended. (p. 44)

The adequacy of the sample has a significant effect on the inferences about item quality and item parameter estimates. If the item parameters are obtained from a small sample, results may be a misleading basis for estimating item parameters (Hambleton & Jones, 1993).

MISSING DATA DUE TO FIELD TEST SAMPLING

Because of the two types of sampling described above, field testing results in sparse data: Not all the items from the item pool are included in a test form, and not all of the examinees complete each test form. Compared to a full data set, it is well known that calibrating a psychometric model is more difficult with a data set with missing data. Missing data is defined as the absence of data value for a variable. It may cause a variety of problems. The lack of data decreases statistical power, can cause bias in parameter estimates, and reduce the representativeness (Rubin, 2004). Missing data may lead to invalid conclusions based on the missingness.

On the other hand, a researcher may intentionally want to collect missing data by using planned missingness. Item and examinee sampling in field test designs is planned missingness. In the following sections I review *planned missing data designs* and then discuss types of field test sampling that produce different patterns of missing data.

Planned Missing Data Designs

Planned missing data designs (PMDD) enable researchers to collect incomplete data from participants by using randomized process (Graham et al., 2006; Little & Rhemtulla, 2013). This method involves intentionally omitting a subset of the data that is predetermined before data collection. By definition, it results in missing completely at random (MCAR) data (Rubin, 1976).

Data are missing completely at random (MCAR) when the probability for a data point to be missing is not related to other variables nor it is related to the missing variable itself. The missing data are a simple random sample of the complete data (Schafer, 1997). It is rare to see MCAR data in uncontrolled environments. The two modern missing data analysis techniques that are highly recommended are multiple imputation (MI) and maximum likelihood (ML) procedures. These techniques provide unbiased estimates with MCAR data (Baraldi & Enders, 2010).

PMDDs are used to reduce participant burden, the length of the test form, and the cost of data collection (Graham et al., 2006; Graham, Taylor & Cumsille, 2001; Raghunathan & Grizzle, 1995). This is done by dividing the item pool into subset of the items and then administering these subsets to examinees. Field testing items and constructing test forms also require a planned missing data designs. After field testing items, data is expected to be MCAR because examinees are randomly administered to test forms, each including different subset of the items. In general, while field testing items, it is practical to use these designs because a researcher desires to collect data with a large number of variables but the time, examinee burden, and fatigue are concerns.

Several ways have been used to assign item sets into test forms. Graham, Hofer, and Piccinin (1994) suggested the three-form design where items are divided into four item sets. In the three form design, one of the sets is included in each form with two of the remaining sets. Then, examinees are randomly assigned to one of the test forms. Including a common set that appears in each form helps to estimate the relationship between variables in different forms. Those designs that allow to link item responses from different test forms are advantageous in terms of dealing with missingness. The higher the correlation between forms, the more information can be gathered for the estimation of missing data. Besides including a common set in each form, a researcher can use another type of design allowing each item set to be appeared

in more than one test form. While each item set appears on different test forms, it also helps to link the examinees' answers from different forms. The forms are administered randomly so that the groups of examinee for each test form are approximately equivalent in terms of ability

One of the versions of the planned missing data designs is referred to as *multiple matrix sampling* (Shoemaker, 1973) and has been applied in educational assessment to develop test forms (Sirotnik & Wellington, 1977; Beaton & Zwick, 1992; Zeger & Thomas, 1997). While large-scale assessments often aim to cover a broad content domain to measure knowledge and skills of the examinees, a limited number of items can be given to an examinee in an available testing time. Matrix sampling approach creates the test forms, each of them is missing a subset of items. After items from the item pool are divided into subset of the items, different combinations of the item sets are randomly assigned to examinees. This approach ensures that all items are taken by a specified number of times without placing a lot burden on examinees in a reasonable testing time.

Johnson (1992) suggested the *balanced incomplete blocks (BIB)* design which allows to estimate means for all item parameters, and correlations for all pairs of item parameters. BIB designs are a variant of multiple matrix sampling, and called *incomplete* when the number of item sets consisting of items from the item pool used in the form is smaller than the overall number of sets, and *balanced* because each item set and each pair of item sets are administered to an equal number of examinees (Frey et al., 2009). Johnson (1992) developed a much larger BIB design with 13 item sets and 26 forms including 3 item sets that has been used in educational assessment literature (Linden, Veldkamp & Carlson, 2004). Table 2.1 shows an example of a BIB design including a total of seven item sets and seven forms each containing three item sets. The *BIB* part of the method refers to assigning item sets into test forms. The forms can be

randomly administered to an equal number of examinees by *spiraling* them among the examinees. Spiraling involves alternating the test forms when distributing them to examinees.

The National Assessment of Educational Progress (NAEP) uses BIB design to develop the assessments so that an examinee is randomly administered a subset of the items from the item pool. To assign items into forms, NAEP uses two types of BIB designs called focused balanced incomplete block (BIB) spiraling and focused partially balanced incomplete block (PBIB) spiraling (Allen et al., 2001). The *focused* part of the method refers answering items from only one subject area. In a BIB design, each item set is paired with every other item set within the same subject so that inter-item correlations can be estimated. Every item set appears once in each possible position on a form resulting in a number of different test forms. In a PBIB design of NAEP, each item set appears twice in each of the two positions. Unlike a BIB design, each item set is not required to be paired an equal number of times with every other item set in this design (Johnson, 1992). For the item sets including items from a content area each of them is paired with every other item set. However, for item sets from different content areas, each item set is paired with only one item set, including items from the other content area.

TIMSS and PIRLS use a matrix sampling approach to assemble items into test forms. For TIMMS 2015 assessment design (Martin et al, 2013), the item pool was divided into 28 item sets in total: 14 sets of mathematics items, and 14 sets of science items. Each set of the items included approximately 10-14 items at the fourth grade and 12-18 at the eight grade. While 16 of the item sets were from the 2011 assessment, 12 sets of items were newly developed for 2015 operational test. The 28 item sets were distributed across 14 test forms, and each form consists of four sets of items: two sets of mathematics items, and two sets of science items. Each examinee completed only one test form. Each item set appeared on two different forms so that the

examinees' answers from different forms can be linked. While distributing item sets across the forms, the main goal was to maximize the content coverage with a sufficient number of items (Martin et al., 2013). It was also reported that the matrix sampling approach helps to reduce examinee burden, complexity of the design process, and cost of the test administration.

It is important to note that the size of item pool plays an important role in developing designs. BIB designs may not be available for any designated item pool size, and may be difficult to develop. However, there might be readily constructed BIB designs in that neighborhood of the item pool size. Additionally, Messick (1983) suggested that if a balanced design cannot be found for a specific situation, then unbalanced designs can be used instead. Unbalanced designs allow that item sets appear an unequal number of times in each position. Table 2.2 shows an example of balanced and unbalanced matrix design. The upper part of the table shows an unbalanced design that does not include a form containing item sets A and C. The lower part of the table shows a balanced matrix design where each item sets appears an equal number of times.

FIELD TESTING FOR MULTIDIMENSIONAL ASSESSMENTS

To this point, practices of field testing items to create a unidimensional assessment were reviewed; these practices, however, cannot be expected to be the same for unidimensional and multidimensional assessments. The key difference between the two is that in a multidimensional assessment, the relationship between dimensions must be estimated. To estimate the relationship between two dimensions, some examinees must take items from both dimensions in the same field test form. Thus, creating field test forms becomes more complex.

The primary questions of field testing discussed above remain important, including how many responses per item are needed to accurately calibrate the model. The number of responses

per item required are expected to vary by the psychometric model used. Thus, multidimensional psychometric models cannot be assumed to have the same data requirements as unidimensional models. Further, for multidimensional models, a new question arises that impacts the sampling of items and examinees in the field test: How many items from a pair of dimensions should appear on the same form to accurately estimate their correlations?

The purpose of this research is to answer these key questions for field testing for a particular kind of multidimensional assessment: the kind where each dimension is assumed to be categorical, and more specifically, binary. I will examine field test conditions useful for diagnostic classification models (DCMs), a family of parametric models under the cognitive diagnosis model umbrella. In the following sections, I review DCMs generally and introduce a general linear form of this family of models.

DIAGNOSTIC CLASSIFICATION MODELS (DCMs)

The No Child Left Behind Act (NCLB; 2001) and the Every Students Succeeds Act (ESSA; 2015-2016) state the necessity of fine-grained feedback from state assessments. Fine-grained and multidimensional feedback is intended to help educators inform their instruction and ultimately improve examinee learning. To this point, most large-scale assessments have been designed as unidimensional, and IRT models provide the basis reports for examinees, educators, and stakeholders. In this approach, one continuous variable is used to represent an overall ability, and diagnostic feedback is often based on subscores. Subscores, however, often lack reliability. (Haberman & Sinharay, 2010). While IRT models are useful, with the ability of ranking and comparing examinees, these models do not generally provide sufficient diagnostic feedback about the source of the problem or examinees' weaknesses and strengths. In contrast, based on

the mastery levels of examinees on a set of knowledge components, DCMs are able to classify examines into two groups: mastery and non-mastery.

Diagnostic classification models (DCMs) are confirmatory multidimensional latent variable models that can provide diagnostic information about the mastery state of knowledge components (Rupp, Templin & Henson, 2010). In DCMs, latent variables are assumed to be discrete, unlike commonly used psychometric models assuming continuous latent variables. These discrete latent variables have been labeled in the literature as *components*, *attributes*, *skills*, *abilities*, or *traits*. Rupp and Templin (2008) defined DCMs as:

> Diagnostic classification models (DCM) are probabilistic, confirmatory multidimensional latent-variable models with a simple or complex loading structure. They are suitable for modelling observable categorical response variables and contain unobservable (i.e., latent) categorical predictor variables. The predictor variables are combined in compensatory and noncompensatory ways to generate latent classes. DCM enable multiple criterion-referenced interpretations and associated feedback for diagnostic purposes, which is typically provided at a relatively fine-grain size. This feedback can be, but does not have to be, based on a theory of response processing grounded in applied cognitive psychology. Some DCMs are further able to handle complex sampling designs for items and examinees, as well as heterogeneity due to strategy use. (p. 226)

As the definition points out that DCMs are confirmatory models in that latent classes and the attribute loading structure are explicit. To determine the latent classes, the attributes to be measured by a test need to be specified a priori. The probability of answering an item correctly is defined as a function of examinees' *attribute profile* representing the latent classes. If a test measures A attributes, the attribute profile for examinee e is an A length vector, denoted

α_e=[$\alpha_{e1}, \alpha_{e2}, \ldots, \alpha_{eA}$], where each element indicates mastery ($\alpha_{ea} = 1$) or non-mastery (α_{ea}= 0) for attribute *a*. DCMs probabilistically classify examinees into one of 2^A attribute profiles based on their mastery of each attribute. For instance, if an examinee is classified into attribute profile [0,1,0,1], it is interpreted that the examinee has mastered Attributes 2 and 4, and not mastered Attributes 1and 3. An attribute profile is assumed to provide information about examinees' strength and weaknesses in certain attributes. The *Q-matrix* (Tatsuoka, 1983) is also specified a priori and can be seen as the loading structure for DCMs. An item-by-attribute *Q-matrix* is formed to specify whether attribute *a* is measured by item *i*, denoted q_i=[$q_{i1}, q_{i2}, \ldots, q_{iA}$], where q_{ia}= 1 indicates the item measures attribute *a*, and q_{ia}= 0 indicates the item does not measure attribute *a*. Items can be written to measure either one attribute or multiple attributes. Items measuring a single attribute are named as simple structure items, and measuring multiple attributes are complex structure items.

As a member of latent class models (LCMs; Lazarsfeld & Henry, 1968), DCMs use item responses to group examinees into latent classes. Given an examinee's class membership, DCMs make the assumption that examinee responses to items are conditionally independent. The structural model and the measurement model are the two components of the latent class models and need to be estimated. The structural component indicates the proportion of examinees within each class, and the measurement component denotes the response probabilities under each class. In a general LCM, when examinee *e* giving answers to *I* items, the probability for the item response vector $\boldsymbol{x_e}$ is:

$$P(\boldsymbol{X_e} = \boldsymbol{x_e}) = \sum_{c=1}^{C} v_c \prod_{i=1}^{I} \pi_{ic}^{x_{ei}} (1 - \pi_{ic})^{1-x_{ei}} \quad (1)$$

where $P(\boldsymbol{X_e} = \boldsymbol{x_e})$ is the observed response probability, and the structural component represented by the summation part, while the measurement component represented by the product part. In Equation 1, v_c is the proportion of examinees who have mastered the attributes required by the class c; x_{ei} is the observed response to Item i by an examinee e; and π_{ic} is the probability of answering Item i correctly by an examinee in latent class c. The product implies that item responses are independent within a latent class as a consequence of the local independence. The structural component represents latent class membership probabilities which provides the base-rate proportion of examinees in Class c.

A large number of DCMs have been developed based on the ways they parameterize the response probabilities. Those models can be categorized as compensatory and noncompensatory models. In compensatory models, mastering a subset of the attributes required by the item can compensate for the non-mastery of the remaining attributes. In noncompensatory models, however, all required attributes need to be mastered to produce a correct response. The deterministic input, noisy and gate model (DINA; Junker & Sijtsma, 2001) is an example of a noncompensatory model, and the deterministic input noisy or gate model (DINO; Templin & Henson, 2006) is an example of a compensatory model. While mastering an additional attribute does not increase the probability of giving a correct answer in a noncompensatory model, compensatory models allow the increase in the probability as mastering additional attributes. The log-linear diagnosis model (LCDM; Henson, Templin, & Willse, 2009), described in the next section, provides a general framework for DCMs. By placing constraints on the LCDM parameters, both compensatory and noncompensatory core DCMs can be specified.

The Log-Linear Cognitive Diagnosis Model

The LCDM is a general DCM that provides a framework to model the relationship

between item responses and attributes. A key feature of the model is that core DCMs can be represented when some parameters are included or constrained in the LCDM, which makes these models nested within the LCDM. To explain the item response function, assume an item measures Attribute 1 and Attribute 4, Q-matrix entries are q_{i1}= 1 and q_{i4}= 1. The LCDM item response function is

$$P(\boldsymbol{X_{ie}} = \boldsymbol{x_e} \mid \boldsymbol{\alpha_e}) = \frac{\exp\left(\lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{e1} + \lambda_{i,1,(4)}\alpha_{e4} + \lambda_{i,2,(1,4)}\alpha_{e1}\alpha_{e4}\right)}{1 + \exp\left(\lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{e1} + \lambda_{i,1,(4)}\alpha_{e4} + \lambda_{i,2,(1,4)}\alpha_{e1}\alpha_{e4}\right)} \quad (2)$$

In Equation 2, X_{ie} is the observed response to item *i* by examinee *e* whose attribute profile is α_e. The LCDM item parameters are akin to a dummy-coded ANOVA model with an intercept, a main effect for each attribute, and interaction term(s) for combinations of attributes. The subscript *i* on the λ parameters represents the item *i*. The second subscript on the λ parameters represents the parameter level: The subscript 0 is for intercept parameters, 1 is for main effects, 2 is for two-way interactions, 3 is for three-way interactions, etc. The third subscripts are in parentheses and represent the attributes to which the main effects or interactions apply. In Equation 1, $\lambda_{i,0}$ is the intercept representing the log-odds of a correct response for examinees who have not mastered either Attribute 1 or Attribute 4. $\lambda_{i,1,(1)}$ and $\lambda_{i,1,(4)}$ are the main effects that represent the increase in log-odds for examinees who possess either Attribute 1 or Attribute 4. $\lambda_{i,2,(1,4)}$ is the two-way interaction term that indicates the change in log-odds of a correct response when examinees have mastered both of the attributes.

While the item measures two attributes in the example above, the LCDM can include more that two attributes resulting in additional main effects and interactions. In the LCDM

framework, the probability of a correct response to item *i* is conditional on an examinee *e*'s attribute profile $\alpha_e = \boldsymbol{\alpha_c}$. The general form of LCDM item response function is

$$P(X_{ie} = 1 | \boldsymbol{\alpha}_c) = \frac{\exp(\lambda_{i,0} + \boldsymbol{\lambda_i^T}\mathbf{h}(\boldsymbol{\alpha_c}, \boldsymbol{q_i}))}{1 + \exp(\lambda_{i,0} + \boldsymbol{\lambda_i^T}\mathbf{h}(\boldsymbol{\alpha_c}, \boldsymbol{q_i}))} \quad (3)$$

In Equation 3, $\boldsymbol{q_i}$ represents the Q-matrix entries for item *i*, and $\lambda_{i,0}$ is the intercept parameter as described above. $\boldsymbol{\lambda_i}$ is a vector of size $(2^A - 1) \times 1$ containing main effect and interaction parameters for item *i*, and $\mathbf{h}(\boldsymbol{\alpha_c}, \boldsymbol{q}_i)$ is a vector of size $(2^A - 1) \times 1$ representing a set of linear combinations of $\boldsymbol{\alpha_c}$ and $\boldsymbol{q_i}$. $\boldsymbol{\lambda_i^T}\mathbf{h}(\boldsymbol{\alpha}_c, \boldsymbol{q}_i)$ is written as:

$$\boldsymbol{\lambda_i^T}\mathbf{h}(\boldsymbol{\alpha_c}, \boldsymbol{q}_i) = \sum_{a=1}^{A} \lambda_{i,1,(a)}\alpha_{ca}\, q_{ia} + \sum_{a=1}^{A-1}\sum_{b>a}^{A} \lambda_{i,2,(a,b)}\alpha_{ca}\alpha_{cb}q_{ia}q_{ib} + \cdots \quad (4)$$

where $\lambda_{i,1,(a)}$ represents the main effect of Attribute *a* on the item *i*, and $\lambda_{i,2,(a,b)}$ represents the two-way interaction effect of Attribute *a* and Attribute *b*. The right side of the equation includes all main effect parameters and all possible interaction parameters. Item parameters are present if the linear combination of $\boldsymbol{\alpha_c}$ and $\boldsymbol{q_i}$ equals 1. For main effects, this only occurs when an examinee has mastered the attribute ($\alpha_{ea} = 1$) and item measures the attribute ($q_{ia} = 1$). For two-way interaction terms, similarly, both attributes are needed to be mastered by an examinee ($\alpha_{ea} = 1, \alpha_{eb} = 1$) and item measures the attributes ($q_{ia} = 1, q_{ib} = 1$). The LCDM models the item responses in a similar way with ANOVA; the attributes and attribute mastery in the LCDM serve as factors and levels of the factors in an ANAOVA, respectively. Constraints are placed on the main effect parameters and interaction terms so that examinees' correct response probabilities increase for DCMs represented by the LCDM (Rupp et al., 2010).

CHAPTER 3

METHODS

This study will systematically investigate the effects of missing data due to field testing on estimation accuracy of diagnostic classification models (DCMs) using simulation. A variation of a parametric DCM has been intended to use in three large-scale testing programs: Parcc, Inc. Diagnostic (e.g., Bradshaw, 2014), Dynamic Learning Maps (Clark et al., 2014), and in Navvy Education's assessment system (L. Bradshaw, personal communication, October 14, 2017). Assessments in each of these systems are administered online. Thus, this study focuses on realizable conditions in an online testing environment.

This study will use the LCDM because it provides the most general framework and represents realistic data (e.g., Bradshaw, Izsak, Templin, & Jacobson, 2014). Its modeling flexibility allows different item response functions on different items.

SIMULATION STUDY DESIGN

I simulated the different designs for conducting a field test in a DCM framework by manipulating four key factors: field test slot length resulting from the number of field test items per attribute an examinee completes (4 levels; 16, 15, 10, 6), the number of dimensions measured by the set of items an examinee completes (2 levels; 2 or 3 attributes), the number of examinees who respond to each field test item (2 levels; 250, 500), and the design of the diagnostic assessment being field tested (i.e., the Q-matrix; 2 levels; Simple and Mixed). Crossing these four factors yielded 32 simulation conditions. In addition to these conditions, the study included 4 full-data conditions that included no missing data. The purpose of the full-data

conditions was for comparison to the missing data conditions. In total, the study contained 36 conditions.

Fixed factors

The attribute correlations for all attribute pairs was fixed at .70. As I assumed the relationship between attributes were strong, .70 allows dimensions to be sufficiently distinct. The base-rate of mastery was set to be .50 for all attributes, meaning that 50% of examinees were masters of each attribute. The diagnostic assessment was designed to measure three attributes with a total of 72 items that would need to be field tested. Thus, the design considered a large number of items per attribute reflecting item pool sizes that may be required for large-scale testing where either (a) several forms are being created for a large population of examinees or (b) larger item pools are needed to prohibit over-exposure of items in an online testing environment.

Manipulated factors

Tables 3.1 and 3.2 describe the levels of the manipulated factors for each condition. The sections below describe those conditions and factors.

Q-matrix. The first Q-matrix of this study (termed Simple) includes 72 simple structure items, each measuring one attribute. Sixteen items measure Attribute 1, 24 items measure Attribute 2, and 32 items measure Attribute 3. In the second Q-matrix (termed Mixed), 50% of the items measure two attributes and the rest measure one attribute. Of the 36 simple items, 8, 12, and 16 items measure Attributes 1-3, respectively. Of the 36 complex items, 8 measure Attribute 1 and 2, 12 measure Attribute 1 and 3, and 16 measure Attribute 2 and 3.

Missing Data. Condition 1 to 4 are full data conditions where all 72 items were administered to each examinee. In the remaining conditions, a subset of items was administered to each examinee yielding different patterns of missing data, depending on the field test form.

Field test designs. The field test design for each condition is shown in Table 3.3 through Table 3.18. Across conditions, test forms included 6, 10, 15, or 16 field test items. Forms with 15 or 16 items may be reasonable for a standalone field test, while forms with 6 or 10 items would be reasonable for embedded field tests. Within a given field test length, either an equal number of items per attribute on the form or a number of items per attribute were proportional to the total number of items in the item pool for the attribute were administered. For example, an equal number of item per attribute on a form with 3 attributes and 15 items would yield 5 items from each attribute, even though there are more items from Attribute 3 than Attribute 2 or 1 that need to be field tested (see Conditions 5 and 6 in Table 3.1); Conditions 9 and 10 show a more proportional number of items per attribute where 4, 5, and 6 items are administered for the three attributes, respectively.

As Table 3.1 shows, the number of dimensions included in a test form differs among conditions with either 2 or 3 attributes being administered per form. If a form contains two attributes, then items measure either Attribute 1 and 2, Attribute 1 and 3, or Attribute 2 and 3. In Conditions 5 to 12, field test designs measured all three attributes in each of the forms. While each form in Condition 5, 6, 7, and 8 contained 5 items from each of the three attributes yielding a total of 15 items, forms in Condition 9, 10, 11, and 12 contained 4, 5, and 6 items from Attributes 1-3, respectively. Conditions 13 through 20 field tested 16 items in total. While Conditions 13-16 measured 2 out of 3 attributes with 8 items per attribute in each form, Conditions 17-20 measured all three attributes with 5, 5, 6 items from Attribute 1, Attribute 2, and Attribute 3, respectively. In Conditions 21 to 24, each form contained 5 items per attribute for two attributes which yielded 10 items in total. Examinees either took a form that measured Attribute 1 and 2, a form that measured Attribute 2 and 3, or a form that measured Attribute 1

and 3. Conditions 25 to 28, however, measured all three attributes by 10 items, 3 for Attribute 1, 3 for Attribute 2, and 4 for Attribute 3. For the Conditions 29 to 32, I assumed each attribute pair was measured by 6 items in total, 3 items for each attribute. Conditions 33 to 36 field tested 6 items, 2 items for each of the three attributes.

As demonstrated in Tables 3.3 through Table 3.18, item and examinee sampling used a matrix sampling approach. This approach administers a set of items from the item pool in different forms, with each examinee completing one form. Each item set was assumed to be administered in at least two forms to provide to link among the examinee responses from different forms. However, when some of the item sets did not appear in two forms before the others did, the other item sets were rotated within the design, resulting in some items sets being included in more than two forms. Recall that, when the item sets appear an equal number of times in each form, the design is called balanced. Some of the simulation conditions, however, did not allow to develop a balanced design because of the rotation; thus, some conditions are unbalanced, as noted in Table 3.2.

Based on the number of field test items from each attribute in a test form and the total number of items measuring each attribute, the number of item sets differ among conditions. Based on the number of item sets within a form, the total number of forms to be designed also differed among conditions. The more slots that were included in a test form, the fewer number of test forms needed to be developed. The fewer slots that were included, the more test forms required. For example, while the field test Design G in Table 3.9 included 16 items per form (5 items from Attribute 1, 5 from Attribute 2, and 6 from Attribute 3) and yielded a total of 15 item sets and 12 different forms, Design K in Table 3.13 included 10 items per form (3 items from Attribute 1, 3 from Attribute 2, and 4 from Attribute 3) and yielded 22 item sets and 16 forms.

Table 3.2 presents the details about the designs for each condition.

To provide a detailed description for one condition, consider Table 3.7 which presents field test Design E. In this design, each examinee was assigned to a form that included two item sets where the sets measured only two of the three attributes, and each item of the 72 items were administered to 250 examinees. The design yielded an overall 1125 sample size, and it required developing 9 test forms to field test the 72 simple structure items. The X's in the table indicate the item sets indicated by the columns are included in the forms indicated by rows. Examinee 1 to 125 were each administered Items 1 through 8 from Attribute 1 and Items 17 through 24 from Attribute 2; examinees 126 to 250 are each administered Items 9 through 16 measuring Attribute 1 and Items 41 through 48 measuring Attribute 3, and so on. The number of item sets each attribute belonged in is 2, 3 and 4 for Attribute 1, Attribute 2, and Attribute 3, respectively. Each of the item sets were administered twice yielding a balanced design so that they were taken by 250 examinees.

Sample size. The sample size for the full data conditions were 250 and 500. In the full data condition, all items in the pool (72 items) were taken by each examinee. These conditions represent unrealistic conditions in which there is no sparse data; results from these conditions were used to indicate the degree to which missing data from the field test designs reflected in the other conditions impacted estimation results.

For each of the conditions mimicking field test designs that yield sparse data, I used two sample sizes: 250 and 500 responses per item. For all field test designs, each item set was administered into at least two different forms. Thus, for the 250 sample size, each form was completed by at least 125 examinees and for the 500 sample size, each form was completed by at least 250 examinees. For the unbalanced field test designs, since some item sets included in more

than two test forms, it yielded more than 250 or 500 responses per item. In order to yield at least 250 or 500 expected responses per item, different overall sample sizes were required because different conditions had different numbers of item sets and the test forms varied among conditions. As shown in Table 3.2, the number of test forms for field test designs ranged from 9 to 29. When 250 responses were collected for each item, it required sample sizes of 1125 to 4000 across the conditions. When 500 responses were collected for each item, it required sample sizes from 2250 to 8000 for field testing.

Item Parameters. For simple structure items, item parameters were drawn from intervals that allowed the probability of answering an item correctly by an examinee who has not mastered the required attribute to range from .20 to .30 and allowed the probability of answering an item correctly by an examinee who has mastered the required attribute to range from .65 to .90.

For the simple structure items in the second Q-matrix, I fixed the true values of the intercept to be -1 or -1.25, which allows the probability of answering an item correctly for non-masters to range from .2 to .3. Main effects were varied between the values of 3, 2.5, 2, and 1.5 which yielded probabilities of answering an item correctly for examinees who have mastered both attributes ranging from .60 to .90. In the item pool, each attribute was measured by at least two items with each of the main effect magnitudes. For the complex items, the intercepts were fixed to be -1 or -1.25. The main effects and interaction terms were specified to yield the desired probability of answering an item correctly for masters of all required attributes from .60 to .90. Examinees who has mastered only one of the attributes had between .44 and .88 probability of a correct response.

Analysis

Data for each field test design was concurrently calibrated using Mplus (Muthén &

Muthén, 2012), using marginal maximum likelihood estimation. For each condition, I conducted 100 replications. Estimation accuracy for the missing data and the full data sets were compared for all conditions. Four outcomes were reported for each condition: (1) item statistics consistency for item point biserial and difficulty, (2) convergence rate, (3) accuracy of item parameters, and (4) examinee classification accuracy.

Item Statistics Consistency. Field test data is often analyzed using sample summary statistics. The summary statistics are produced to flag an item for review by content experts. These experts review the content of the item to who determine whether an item needs to be revised or removed.

I evaluated two summary statistics: item difficulty, the proportion of examinees who answered the item correctly, and within attribute point biserial, the relationship between the item response and the subscore on the attribute. For simple structure items, the subscore is straightforward; each item that measures the attribute contributes 1 point towards the subscore. For complex items, I calculated the subscore the same way; each item that measured the attribute contributed 1 point towards the subscore, even if the item also measured another attribute.

Item difficulty is the proportion of examinees who answered an item correctly. The item difficulty ranges from 0 to 1.0; the item is easier as the value get closer to 1.0. Item difficulty provides information on item discrimination, too. If the item is too difficult, and most of the examinees answer the item incorrectly, or if the item is too easy and most of the examinees answer the item correctly, then the item does not provide much diagnostic information about the examinees.

The within attribute point-biserial values range from -1.00 to +1.00. Within attribute positive point biserial values indicate that examinees who scored high on the collection of items

that measure the attribute are more likely to answer the item measuring that attribute correctly. In another words, the item functions similarly to other items that measure the same attribute.

Consistency. Results from the full data will be compared with the missing data to determine the degree to which the field test designs resulting in missing data yield stable item statistics. Besides the corresponding sample size in the full data conditions (250 or 500), an additional large full data condition with sample size 10,000 was used for comparison. It was assumed that the highest sample size full data condition would be a better comparison for item analysis consistency. To do so, I will calculate the average absolute difference of the item statistics (attribute point biserial and difficulty) between the full data conditions and the missing data (field testing) conditions. As this difference increases, the consistency of the item statistics decreases.

Convergence Rate. I reported the percentage of replications (out of 100) that each condition converged. It is expected that convergence rate is affected by these factors: the field test slot length (4 levels; 16, 15, 10, 6), number of attributes in each form (2 vs. 3), sample size (250 vs. 500), and the design of the diagnostic test being field tested (Q-matrix; Simple vs. Mixed). I evaluated convergence rates based on these factors.

Item Parameter Bias. I evaluated the quality of parameter estimation using the mean absolute bias between true and estimated values of the item parameters. I reported the mean absolute bias by parameter type: intercepts, main effects, and interaction terms.

Classification Accuracy. I compared the true attribute mastery states to the estimated ones. I reported the attribute classification accuracy as the proportion of examinees whose estimated mastery state for each attribute matched their true mastery status.

CHAPTER 4

RESULTS

In this section, I provide the results of the simulation study according to the analyses described in the previous chapter. I will focus on comparing results from the full data conditions and missing data conditions. The results from full data conditions will be used to indicate the degree to which missing data from the field test designs impacted estimation results.

SIMULATION STUDY RESULTS

I report the effect of missing data that results from different field testing designs on four outcomes: (1) convergence rate, (2) accuracy of item parameters, (3) item statistics consistency and (4) classification accuracy. I analyzed results under full and sparse data conditions.

Convergence Rates

Table 4.1 shows the convergence rates for the 36 conditions. Full data conditions were conditions where examinees completed all items, so these conditions always measured 3 attributes per 'form'. The corresponding conditions that measured 2 attributes are indicated as NA in Table 4.1. The convergence rate for full data conditions with simple structure items (Condition 1 and Condition 2) with sample size of 250 and 500 was 1. Full data conditions 3 and 4 that contained both simple and complex structured items converged less than Condition 1 and 2 that included only simple structure items. The convergence rates of Condition 3 with 250 sample size and Condition 4 with 500 sample size are .29 and .43, respectively.

For the mixed Q-matrix conditions, convergence rates decreased significantly. In the mixed Q-matrix designs, convergence rates are higher when forms contain all 3 attributes instead

of only 2. The convergence rates ranged from .9 to .57 for conditions measuring 2 attributes in each form with a sample size of 250 and ranged from .14 to .89 for conditions measuring 3 attributes. Additionally, sample size had a strong effect on convergence rates; larger sample sizes of 500 resulted in higher convergence rates.

Accuracy of Item Parameters

I examined the degree to which item parameters were accurately estimated with missing data from different designs. Table 4.2 shows the mean absolute bias for the intercept and main effect estimates for the conditions that had a simple Q-matrix data design with a 250 or 500 sample size. The missing data conditions had greater bias than the corresponding full data conditions Condition 1 and Condition 2. The results showed that the models were reasonably well estimated under the missing data conditions.

The number of field test slots had a large impact on bias. Conditions 29, 30, 33, and 34 (Design M and Design O) where forms were the shortest (6 items) had greater bias than remaining conditions. Bias for these conditions ranged from .06 to .08 for the intercept and .09 to .12 for the main effects when the sample size for each item is 250, and ranged from .04 to .05 for the intercept and .06 to .08 for the main effects when the sample size is 500. The variability of the bias was also greatest for the 6-item conditions. The bias was least for the missing data Condition 13 and Condition 14 (Design E) where examinees responded to 16 items.

Results from Table 4.2 shows that when examinees answer 5 items per dimension yielding 15 items in total (Design A), the bias was .044 for intercept and .066 for the main effect. Increasing the overall number of items in a test from 15 to 16 items did not change the estimation accuracy. When examinees answered 5 items for Attribute 1, 5 for Attribute 2, and 6 for Attribute 3 yielding 16 items in total (Design G), the bias was .043 for the intercept and .064 for

the main effect. On the other hand, the bias considerably increased when examinees answered 10 items in total; 3, 3, and 4 items for Attribute 1, Attribute 2, and Attribute 3, respectively (Design K). Lastly, the bias reached the maximum value among conditions when examinees responded the fewest number of items (6 items in total). In that design, examinees answered 2 items per dimension (Design O), and the bias was .077 for the intercept and .117 for the main effect. On the other hand, when the mixed Q-matrix was used, the overall number of items affected the item parameter accuracy. As shown in Table 4.3, while measuring 16 items in a test, the bias was ranged from .049 to .051 for intercept, and ranged from .090 to .126 for main effects (Design F and H); while measuring 10 items, the bias for the intercept was ranged from .055 to .058 for the intercept and from .097 to .144 for the main effects (Design J and L). The bias was highest for the design where 6 items included in a test form; the bias was .069 for intercept and ranged from .121 to .162 for main effects (Design N and P).

The sample size also had an impact on bias. The designs with 500 sample size had smaller bias than corresponding missing data conditions with 250 sample size. As the number of responses per item increases, the bias of estimation decreases. As shown in Table 4.2, the bias for 250 sample size conditions with simple Q-matrix was ranged from .041 to .077 for intercept and ranged from .061 to .117 for main effect. When the sample size increased to 500 for per item, then the bias decreased and ranged from .029 to .052 for intercept and ranged from .043 to .079 for main effect. Similar results were obtained from the designs where mixed Q-matrix was used for 250 and 500 sample sizes. While the bias for intercept ranged from .049 to .069 for 250 sample size, it ranged from .037 to .049 for 500 sample size conditions. The bias for main effects ranged from .078 to .208 for 250 sample size conditions, it ranged from .062 to .156 for 500 sample size conditions.

The complexity of the Q-matrix had a significant impact on bias. Table 4.3 shows the mean absolute bias for the intercept and main effect estimates for the conditions that were designed to field test both simple and complex structure items (mixed) with the sample size of 250 and 500. Overall, the parameters are reasonably well estimated; however, the bias in the mixed Q-matrix conditions is greater than the bias in the simple Q-matrix conditions. For example, when a test measured 15 items, the bias for intercept was ranged from .031 to .044 for simple Q-matrix conditions (Design A and C) and .101 to .135 for mixed Q-matrix conditions (Design B and D). Additionally, while the minimum value of bias for intercept was .029 for simple Q-matrix conditions (Design E), it was .037 for mixed Q-matrix conditions (Design H).

The number of attributes measured within a field test form also made a significant difference in the bias. When holding the number of field test slots constant, the field test designs that included two attributes instead of three in the test form yielded smaller bias. For example, as shown in Table 4.3, when the form contained 16 items, the bias for the intercept was .049 for form measuring 2 attributes (Design F), and .051 for the form measuring 3 attributes (Design H). The bias for the main effects ranged from .078 to .111 for the forms measuring 2 attributes, and .090 to .126 for the forms measuring 3 attributes. When test forms contained two attributes, the number of items for each attribute in the test form was greater than that of forms contained three attributes. For example, when the field test slot length is 16 and the test form contained 2 attributes, each attribute was measured by either 8 items with simple Q-matrix design or by 12 items with mixed Q-matrix design. However, when the test form contained 3 attributes, each attribute was measured by 5, 5, and 6 items or 7, 8, and 9 items for Attribute 1, Attribute 2, and Attribute 3, respectively.

Overall, the bias was lowest when the test form included the highest number of items (16 items), the Q-matrix was simple, the test form included two attributes instead of three, and the sample size per item was 500. Based on these results, field test design Design E (Table 3.7) was the very best condition. On the other hand, the bias was greatest when the test form included overall 6 items, the Q-matrix was mixed, the test form included two attributes instead of three, and the sample size per item was 250. Based on these results, filed test design Design N (Table 3.16) was the very worst condition among others.

Item Statistics Consistency

Full data and missing data conditions were compared to see if the missing data resulting from field test designs yielded stable item statistics. Each missing data condition was compared with 10,000 sample size full data condition and corresponding sample size in the full data conditions (e.g. 250 or 500). 10,000 sample size condition was an additional large full data condition that was used for analysis of item statistics consistency. The mean absolute difference for the item difficulty and within attribute point biserial values were reported on Table 4.4, Table 4.5 and Table 4.6.

The field test designs yielded relatively stable item statistics when the overall number of items included in a test is either 15 or 16: While the average *p*-value (difficulty) differences ranged from .033 to .037 for the conditions field-testing 15-16 items with 250 sample size, the average point-biserial differences ranged from .060 to .123. Those designs where test forms contained 6 items yielded the least stable item statistics for the point-biserial values: While the average *p*-value difference ranged from .030 to .034, the average point-biserial difference ranged from .127 to .266.

If the consistency is high, then the item statistics for the missing data conditions are similar to corresponding full data conditions. The impact of the sample size on the item statistics also can be seen from the tables. The consistency is the least when the sample size is 250, indicating that summary statistics are more stable with the large sample size of 500. For example, while the average *p*-value difference ranged from .028 to .037 for small sample size, it ranged from .02 to .024 for large sample size. Similar results were shown for the average point-biserial differences; it ranged from .056 to .266 for the smaller sample size and from .047 to .184 for larger sample size.

Classification Accuracy

Next, the impact on classification accuracy was examined. Classification accuracy for individual attributes is the proportion of examinees whose estimated mastery state for the attribute matched with their true mastery state. Table 4.7 shows the classification accuracy rates for the designs where simple structure items were field tested. Accuracy was relatively high in all conditions except Conditions 29 and 33 where the test forms included 6 items (Design M and O). Comparing Condition 29 and Condition 33 with corresponding full data conditions, the classification accuracy decreased by 16-21% and 15-20%, respectively. The sparse data from Condition 21 and Condition 25 (Design I and K) also yielded noticeable decreases in classification accuracy of 11-17% and 10-13%, respectively. On the other hand, in Conditions 5 and 9 (Design A and C), test designs yielded strong accuracy rates, and sparse data decreased the classification accuracy by 6-8%.

Although Conditions 13-17, 21-25, and 29-33 field tested the same number of items in each form, the results showed that the design where three attributes are included in a test form had greater accuracy than including only a pair of the attributes. The same result was obtained

from the Condition 5 measuring 3 attributes with a total of 15 items (Design A) and Condition 13 measuring 2 attributes with a total of 16 items (Condition E), even though for the attributes only 5 items per attribute instead of 8 were administered, and the Condition 5 field test design was one item shorter.

These results came from the designs where the number of responses per item was 250 and did not largely differ from the results of field test designs with 500 sample size. Obtaining similar accuracy rates with 250 and 500 responses per items indicates that sample size did not significantly affect the classification accuracy. Results also show when all three attributes were included in the test form, the overall number of items answered had the largest impact on classification accuracy.

Table 4.8 shows the results for the field test designs where both simple and complex structure items are used. Overall, accuracy was high in all conditions; however, it was a little lower than the corresponding simple data conditions. For example, Conditions 10 and 12 measuring 3 attributes with 15 items in total. In Condition 10 (Design C), Attribute 1, Attribute 2, and Attribute 3 was measured by 4, 5, and 6 simple structure items, respectively. On the other hand, each attribute was measured by 7, 7, and 9 simple and complex structure items in Condition 12 (Design D). Although the number of items per attribute in Condition 10 is less than that of Condition 12, the accuracy in Condition 10 is higher. The accuracy rates for Condition 10 ranged from 91 to 93%, and for Condition 12 ranged from 89 to 92%.

CHAPTER 5

CONCLUSIONS

Simulation results demonstrated the impact of sparse data, due to different field testing designs, on the estimation accuracy for DCMs. I simulated 36 field testing designs and found that sample item statistics, item parameter estimates, and classification accuracy levels varied across these designs.

Convergence results demonstrated that both sample sizes may be too low to yield high rates of convergence with mixed Q-matrices. Though, with simple Q-matrices, 250 is a sufficient sample size.

The overall number of items answered in a test had a considerable effect on the estimation accuracy. As the number of items increases, the bias of item parameter estimation decreases. Additionally, field test designs yielded relatively stable item statistics as the field test slot length was getting longer. The item statistics for those designs were similar to corresponding full data conditions. Overall, the results from the simulation study indicates that if an attribute is measured by 5 items or more, then item parameter estimation achieves reasonable levels of accuracy.

The design of the diagnostic assessment is another factor that affects the estimation. The simulation study results show that in a design where both simple and complex structure items are field tested, item parameters have greater bias than that of designs field testing only simple structure items. The item parameter bias increases when the Q-matrix is mixed. Additionally, when the test forms contain two attributes instead of three, the parameter estimates are less

biased. Including two attributes in a form yields more items per attribute than including three attributes in a form; thus, designing a field test form with a maximum number of attributes should not be the goal. Instead, as long as a pair of attributes are included on the test form to estimate attribute relationships, it is more important to focus on increasing the number of items per attribute taken by an examinee.

The number of responses per item has a significant impact on item parameter estimates. When the sample size increases, the item parameter estimation bias decreases. Sample size per item has also a significant impact on the consistency. Large sample size yielded the most stable item statistics. However, increasing the sample size from 250 to 500 did not largely affect the classification accuracy rates. The overall number of items taken by examinees has a considerable impact on classification accuracy rates: as the number of items increases, classification accuracy increases. On the other hand, the results from the simulation study shows that when a field test design included three attributes instead of two in the test form, classification accuracy increases.

Results from this study can be used as an initial guide for test developers. Currently test developers do not have guidelines for how to structure a field test for a multidimensional assessment that uses DCMs. Field testing is a crucial step in creating assessments to gather sufficient data to accurately specify the Q-matrix and estimate model parameters. By examining various practical conditions, this study aims to help to fill the need for guidelines for test developers.

Limitations and Future Directions

This study was an initial exploration of designing filed tests for multidimensional assessments. The study is limited by some factors, including calculation of the subscores for complex items to analyze within point biserial. Subscores for complex items could be calculated

in different ways. For example, if an item measures two attributes, then it could count ½ point towards each attribute subscore instead of 1 point as done in this study. On the other hand, the Q-matrices were known and correct. In practice, however, some entries of the Q-matrix may not be correct, especially in the field testing stage.

In addition, designing a filed test design for multidimensional tests depends on the interplay of many factors. This study examined a small set of conditions as an initial examination. While these practical conditions can help be an initial guide for test developers, many variations of field testing exist. As a future study, many other combinations of the factors that play an important role in designing field tests can be created to explore.

Table 2.1

An Example Balanced Incomplete Blocks Design

	Item Sets						
Forms	**A**	**B**	**C**	**D**	**E**	**F**	**G**
1	X	X		X			
2		X	X		X		
3			X	X		X	
4				X	X		X
5	X				X	X	
6		X				X	X
7	X		X				X

Note. X indicates that the item set is included in the form.

Table 2.2

An Example of Balanced and Unbalanced Matrix Design

	Item Sets		
Form	**A**	**B**	**C**
Unbalanced Design			
1	X	X	
2		X	X
Balanced Design			
3	X	X	
4		X	X
5	X		X

Note. X indicates that the item set is included in the form.

Table 3.1
Simulation Conditions

Conditions	Attribute	Items in Form of Total Items per Attribute	Number of Attributes on Form	Number of Items in a Form	Q-matrix	Field Test Design	Expected RPI Min
1-2	1, 2, 3	72 of 72	3	72	Simple	NA	250, 500
3-4	1, 2, 3	72 of 72	3	72	Mixed	NA	250, 500
5-6	1 2 3	5 of 16 5 of 24 5 of 32	3	15	Simple	A	250, 500
7-8	1 2 3	8 of 30 8 of 36 8 of 42	3	15	Mixed	B	250, 500
9-10	1 2 3	4 of 16 5 of 24 6 of 32	3	15	Simple	C	250, 500
11-12	1 2 3	7 of 30 7 of 36 9 of 42	3	15	Mixed	D	250, 500
13-14	1 2 3	8 of 16 8 of 24 8 of 32	2	16	Simple	E	250, 500
15-16	1 2 3	12 of 30 12 of 36 12 of 42	2	16	Mixed	F	250, 500
17-18	1 2 3	5 of 16 5 of 24 6 of 32	3	16	Simple	G	250, 500
19-20	1 2 3	7 of 30 8 of 36 9 of 42	3	16	Mixed	H	250, 500
21-22	1 2 3	5 of 16 5 of 24 5 of 32	2	10	Simple	I	250, 500
23-24	1 2 3	7 of 30 7 of 36 8 of 42	2	10	Mixed	J	250, 500
25-26	1 2 3	3 of 16 3 of 24 4 of 32	3	10	Simple	K	250, 500
27-28	1 2 3	4 of 30 5 of 36 6 of 42	3	10	Mixed	L	250, 500
29-30	1 2 3	3 of 16 3 of 24 3 of 32	2	6	Simple	M	250, 500
31-32	1 2 3	4 of 30 4 of 36 5 of 42	2	6	Mixed	N	250, 500
33-34	1 2 3	2 of 16 2 of 24 2 of 32	3	6	Simple	O	250, 500
35-36	1 2 3	3 of 30 3 of 36 3 of 42	3	6	Mixed	P	250, 500

Note. RPI= responses per item. The two conditions in the same row correspond to the two sample sizes, with other factors being the same.

Table 3.2

Field Test Design Factors

Conditions	Field Test Design	Measured # of attributes in a form	Field test slots per form	# of item sets	# of items in item set	# of forms	Type of Design
5-6	A	3	15	16	5	14	Unbalanced
7-8	B	3	15	12	6,9	16	Balanced
9-10	C	3	15	15	4, 5, 6	12	Unbalanced
11-12	D	3	15	11	7, 8	12	Unbalanced
13-14	E	2	16	9	8	9	Balanced
15-16	F	2	16	18	8, 8	12	Balanced
17-18	G	3	16	15	5, 5, 6	12	Unbalanced
19-20	H	3	16	17	8, 8	12	Balanced
21-22	I	2	10	16	5	16	Balanced
23-24	J	2	10	24	5, 5	16	Unbalanced
25-26	K	3	10	22	3, 3, 4	16	Unbalanced
27-28	L	3	10	24	5, 5	18	Unbalanced
29-30	M	2	6	25	3	25	Balanced
31-32	N	2	6	39	3, 3	26	Balanced
33-34	O	3	6	36	2	29	Unbalanced
35-36	P	3	6	39	3, 3	26	Balanced

Table 3.3

Field Test Design A

Examinee		**Item**														
		Attribute 1				**Attribute 2**					**Attribute 3**					
250	**500**	1-5	6-10	11-15	12-16	17-21	22-26	27-31	32-36	36-40	41-45	46-50	51-55	56-60	61-65	66-70	68-72
1-125	**1-250**	X				X					X						
126-250	**251-500**		X				X					X					
251-375	**501-750**			X				X					X				
376-500	**751-1000**				X				X					X			
501-625	**1001-1250**	X								X					X		
626-750	**1251-1500**		X			X										X	
751-875	**1501-1750**			X			X										X
876-1000	**1751-2000**				X			X			X						
1001-1125	**2001-2250**	X							X			X					
1126-1250	**2251-2500**		X							X			X				
1251-1375	**2501-2750**			X		X								X			
1376-1500	**2751-3000**				X		X								X		
1501-1625	**3001-3250**	X						X								X	
1626-1750	**3251-3500**		X						X								X

Note. Each row represents a different form. X indicates that the item set is included in the form. Items 1-16 measure Attribute 1; Items 17-40 measure Attribute 2; Items 41-72 measure Attribute 3.

Table 3.4

Field Test Design B

Examinee		Item Set											
		Attribute 1&2&3								Attribute 1&2&3			
250	**500**	S1	S2	S3	S4	S5	S6	S7	S8	C1	C2	C3	C4
1-125	**1-250**	X								X			
126-250	**251-500**		X								X		
251-375	**501-750**			X								X	
376-500	**751-1000**				X								X
501-625	**1001-1250**					X				X			
626-750	**1251-1500**						X				X		
751-875	**1501-1750**							X				X	
876-1000	**1751-2000**								X				X
1001-1125	**2001-2250**	X											X
1126-1250	**2251-2500**		X									X	
1251-1375	**2501-2750**			X							X		
1376-1500	**2751-3000**				X					X			
1501-1625	**3001-3250**					X							X
1626-1750	**3251-3500**						X					X	
1751-1875	**3501-3750**							X			X		
1876-2000	**3751-4000**								X	X			

Note. Each row represents a different form. X indicates that the item set is included in the form. S1-S8=Simple structure item sets that include 6 items. C1-C4=Complex structure item sets that include 9 items. All S and C sets include items from all 3 attributes.

Table 3.5

Field Test Design C

Examinee		**Item Set**														
		Attribute 1				**Attribute 2**					**Attribute 3**					
250	**500**	1-4	5-8	9-12	13-16	17-21	22-26	27-31	32-36	36-40	41-46	47-52	53-58	59-64	65-70	67-72
1-125	**1-250**	X				X					X					
126-250	**251-500**		X				X					X				
251-375	**501-750**			X				X					X			
376-500	**751-1000**				X				X					X		
501-625	**1001-1250**	X								X					X	
626-750	**1251-1500**		X			X										X
751-875	**1501-1750**			X			X				X					
876-1000	**1751-2000**				X			X				X				
1001-1125	**2001-2250**	X							X				X			
1126-1250	**2251-2500**		X							X				X		
1251-1375	**2501-2750**			X		X									X	
1376-1500	**2751-3000**				X		X									X

Note. Each row represents a different form. X indicates that the item set is included in the form. Items 1-16 measure Attribute 1; Items 17-40 measure Attribute 2; Items 41-72 measure Attribute 3.

Table 3.6

Field Test Design D

Examinee		**Item Set**										
		Attribute 1&2&3						**Attribute 1&2&3**				
250	**500**	S1	S2	S3	S4	S5	S6	C1	C2	C3	C4	C5
1-125	**1-250**	X						X				
126-250	**251-500**		X						X			
251-375	**501-750**			X						X		
376-500	**751-1000**				X						X	
501-625	**1001-1250**					X						X
626-750	**1251-1500**						X	X				
751-875	**1501-1750**	X							X			
876-1000	**1751-2000**		X							X		
1001-1125	**2001-2250**			X							X	
1126-1250	**2251-2500**				X							X
1251-1375	**2501-2750**					X		X				
1376-1500	**2751-3000**						X		X			

Note. Each row represents a different form. X indicates that the item set is included in the form. S1-S6=Simple structure item sets that include 7 items. C1-C5=Complex structure item sets that include 8 items. All S and C sets include items from all 3 attributes.

Table 3.7

Field Test Design E

Examinee		**Item**								
		Attribute 1		**Attribute 2**			**Attribute 3**			
250	**500**	1-8	9-16	17-24	25-32	33-40	41-48	49-56	57-64	65-72
1-125	**1-250**	X		X						
126-250	**251-500**		X				X			
251-375	**501-750**				X			X		
376-500	**751-1000**					X			X	
501-625	**1001-1250**	X								X
626-750	**1251-1500**		X				X			
751-875	**1501-1750**			X				X		
876-1000	**1751-2000**				X				X	
1001-1125	**2001-2250**					X				X

Note. Each row represents a different form. X indicates that the item set is included in the form. Items 1-16 measure Attribute 1; Items 17-40 measure Attribute 2; Items 41-72 measure Attribute 3.

Table 3.8

Field Test Design F

Examinee		**Item Set**																	
		Attribute 1&2				**Attribute 1&3**				**Attribute 2&3**				**Attribute 1&2**		**Attribute 1&3**		**Attribute 2&3**	
250	**500**	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	C1	C2	C3	C4	C5	C6
1-125	**1-250**	X												X					
126-250	**251-500**		X												X				
251-375	**501-750**			X										X					
376-500	**751-1000**				X										X				
501-625	**1001-1250**					X										X			
626-750	**1251-1500**						X										X		
751-875	**1501-1750**							X								X			
876-1000	**1751-2000**								X								X		
1001-1125	**2001-2250**									X								X	
1126-1250	**2251-2500**										X								X
1251-1375	**2501-2750**											X						X	
1376-1500	**2751-3000**												X						X

Note. Each row represents a different form. X indicates that the item set is included in the form. S1-S12=simple structure item sets, C1-C6=complex structure item sets. Each set includes 8 items. S1, S2, S3, and S4 include items from Attributes 1 and 2; S5, S6, S7, and S8 include items from Attributes 1 and 3; S9, S10, S11 and S12 include items from Attributes 2 and 3. C1 and C2 include items from Attributes 1 and 2; C3 and C4 include items from Attribute 1 and 3; C5 and C6 include items from Attribute 2 and 3.

Table 3.9

Field Test Design G

Examinee		**Item**														
		Attribute 1				**Attribute 2**					**Attribute 3**					
250	**500**	1-5	6-10	11-15	12-16	17-21	22-26	27-31	32-36	36-40	41-46	47-52	53-58	59-64	65-70	67-72
1-125	**1-250**	X				X					X					
126-250	**251-500**		X				X					X				
251-375	**501-750**			X				X					X			
376-500	**751-1000**				X				X					X		
501-625	**1001-1250**	X								X					X	
626-750	**1251-1500**		X			X										X
751-875	**1501-1750**			X			X				X					
876-1000	**1751-2000**				X			X				X				
1001-1125	**2001-2250**	X							X				X			
1126-1250	**2251-2500**		X							X				X		
1251-1375	**2501-2750**			X		X									X	
1376-1500	**2751-3000**				X		X									

Note. Each row represents a different form. X indicates that the item set is included in the form. Items 1-16 measure Attribute 1; Items 17-40 measure Attribute 2; Items 41-72 measure Attribute 3.

Table 3.10

Field Test Design H

Examinee		**Item Set**																
		Attribute 1&2				**Attribute 1&3**				**Attribute 2&3**				**Attribute 1&2&3**				
250	**500**	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	C1	C2	C3	C4	C5
1-125	**1-250**	X												X				
126-250	**251-500**		X												X			
251-375	**501-750**			X												X		
376-500	**751-1000**				X												X	
501-625	**1001-1250**					X												X
626-750	**1251-1500**						X							X				
751-875	**1501-1750**							X							X			
876-1000	**1751-2000**								X							X		
1001-1125	**2001-2250**									X							X	
1126-1250	**2251-2500**										X							X
1251-1375	**2501-2750**											X		X				
1376-1500	**2751-3000**												X		X			

Note. Each row represents a different form. X indicates that the item set is included in the form. S1-S12=Simple structure item sets, C1-C5=Complex structure item sets. Each simple and complex structure item sets include 8 items. S1, S2, S3, and S4 include items from Attributes 1 and 2; S5, S6, S7, and S8 include items from Attributes 1 and 3; S9, S10, S11 and S12 include items from Attributes 2 and 3. All Cs sets include items from all 3 attributes.

Table 3.11

Field Test Design I

Examinee		**Item Set**															
		Attribute 1				**Attribute 2**					**Attribute 3**						
250	**500**	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1-125	**1-250**	X				X											
126-250	**251-500**		X				X										
251-375	**501-750**			X							X						
376-500	**751-1000**				X							X					
501-625	**1001-1250**							X					X				
626-750	**1251-1500**								X					X			
751-875	**1501-1750**									X					X		
876-1000	**1751-2000**	X														X	
1001-1125	**2001-2250**		X														X
1126-1250	**2251-2500**					X					X						
1251-1375	**2501-2750**						X					X					
1376-1500	**2751-3000**							X					X				
1501-1625	**3001-3250**								X					X			
1626-1750	**3251-3500**									X					X		
1751-1875	**3501-3750**			X												X	
1876-2000	**3751-4000**				X												X

Note. Each row represents a different form. X indicates that the item set is included in the form. Set 1 contains Items 1-5, Set 2 contains Items 6-10, and the remaining sets continue in this pattern, each containing the next five items in the pool of 72 items. Sets 1-4 contain items that measure Attribute 1, Sets 5-9 contains items that measure Attribute 2, and Sets 10 to 16 contains items that measure Attribute 3.

Table 3.12

Field Test Design J

Examinee		Item Set																							
		Attribute 1&2				Attribute 2&3						Attribute 1&3						Attribute 1&2		Attribute 2&3		Attribute 1&3			
250	**500**	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	C1	C2	C3	C4	C5	C6	C7	C8
1-125	**1-250**	X																X							
126-250	**251-500**		X																X						
251-375	**501-750**			X														X							
376-500	**751-1000**				X														X						
501-625	**1001-1250**					X														X					
626-750	**1251-1500**						X														X				
751-875	**1501-1750**							X														X			
876-1000	**1751-2000**								X											X					
1001-1125	**2001-2250**									X											X				
1126-1250	**2251-2500**										X											X			
1251-1375	**2501-2750**											X											X		
1376-1500	**2751-3000**												X											X	
1501-1625	**3001-3250**													X											X
1626-1750	**3251-3500**														X								X		
1751-1875	**3501-3750**															X								X	
1876-2000	**3751-4000**																X								X

Note. Each row represents a different form. X indicates that the item set is included in the form. S1-S16=simple structure item sets, C1-C8=complex structure item sets. Each set includes 5 items. S1, S2, S3, and S4 include items from Attributes 1 and 2; S5, S6, S7, S8, S9, and S10 include items from Attributes 2 and 3; S11, S12, S13, S14, S15, and S16 include items from Attributes 1 and 3. C1 and C2 include items from Attributes 1 and 2; C3 and C4 include items from Attribute 2 and 3; C5 and C8 include items from Attribute 1 and 3.

Table 3.13

Field Test Design K

Examinee		**Item**																			
		Attribute 1						**Attribute 2**								**Attribute 3**							
250	**500**	1-3	4-6	7-9	10-12	13-15	14-16	17-19	20-22	23-25	26-28	29-31	32-34	35-37	38-40	41-44	45-48	49-52	53-56	57-60	61-64	65-68	69-72
1-125	**1-250**	X						X								X							
126-250	**251-500**		X						X								X						
251-375	**501-750**			X						X								X					
376-500	**751-1000**				X						X								X				
501-625	**1001-1250**					X						X								X			
626-750	**1251-1500**						X						X								X		
751-875	**1501-1750**	X												X								X	
876-1000	**1751-2000**		X												X								X
1001-1125	**2001-2250**			X				X								X							
1126-1250	**2251-2500**				X				X								X						
1251-1375	**2501-2750**					X				X								X					
1376-1500	**2751-3000**						X				X								X				
1501-1625	**3001-3250**	X										X								X			
1626-1750	**3251-3500**		X										X								X		
1751-1875	**3501-3750**			X										X								X	
1876-2000	**3751-4000**				X										X								X

Note. Each row represents a different form. X indicates that the item set is included in the form. Items 1-16 measure Attribute 1; Items 17-40 measure Attribute 2; Items 41-72 measure Attribute 3.

Table 3.14

Field Test Design L

Examinee		Item Set																							
		Attribute 1&2				Attribute 2&3						Attribute 1&3						Attribute 1&2&3							
250	**500**	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	C1	C2	C3	C4	C5	C6	C7	C8
1-125	**1-250**	X																					X		
126-250	**251-500**		X																					X	
251-375	**501-750**			X																					X
376-500	**751-1000**				X															X					
501-625	**1001-1250**	X																			X				
626-750	**1251-1500**		X																			X			
751-875	**1501-1750**					X												X							
876-1000	**1751-2000**						X												X						
1001-1125	**2001-2250**							X										X							
1126-1250	**2251-2500**								X														X		
1251-1375	**2501-2750**									X														X	
1376-1500	**2751-3000**										X														X
1501-1625	**3001-3250**											X						X							
1626-1750	**3251-3500**												X						X						
1751-1875	**3501-3750**													X					X						
1876-2000	**3751-4000**														X					X					
2001-2125	**4001-4250**															X					X				
2126-2250	**4251-4500**																X					X			

Note. Each row represents a different form. X indicates that the item set is included in the form. S1-S16=Simple structure item sets, C1-C8=Complex structure item sets. Each simple and complex structure item sets include 5 items. S1, S2, S3, and S4 include items from Attributes 1 and 2; S5, S6, S7, S8, S9 and S10 include items from Attributes 2 and 3; S11, S12, S13, S14, S15, and S16 include items from Attributes 1 and 3. All C sets include items from all 3 attributes.

Table 3.15

Field Test Design M

Examinee		Item Set																								
		Attribute 1						Attribute 2								Attribute 3										
250	**500**	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1-125	**1-250**	X						X																		
126-250	**251-500**		X						X																	
251-375	**501-750**			X						X																
376-500	**751-1000**				X											X										
501-625	**1001-1250**					X											X									
626-750	**1251-1500**						X											X								
751-875	**1501-1750**										X								X							
876-1000	**1751-2000**											X								X						
1001-1125	**2001-2250**												X								X					
1126-1250	**2251-2500**													X								X				
1251-1375	**2501-2750**														X								X			
1376-1500	**2751-3000**	X																						X		
1501-1625	**3001-3250**		X																						X	
1626-1750	**3251-3500**			X																						X
1751-1875	**3501-3750**				X											X										
1876-2000	**3751-4000**					X											X									
2001-2125	**4001-4250**						X											X								
2126-2250	**4251-4500**							X											X							
2251-2375	**4501-4750**								X											X						
2376-2500	**4751-5000**									X											X					
2501-2625	**5001-5250**										X											X				
2626-2750	**5251-5500**											X											X			
2751-2875	**5501-5750**												X											X		
2876-3000	**5751-6000**													X											X	
3001-3125	**6001-6250**														X											X

Note. Each row represents a different form. X indicates that the item set is included in the form. Set 1 contains Items 1-3, Set 2 contains Items 4-6, and the remaining sets continue in this pattern, each containing the next three items in the pool of 72 items. Sets 1-6 contain items that measure Attribute 1, Sets 7-14 contains items that measure Attribute 2, and Sets 15 to 25 contains items that measure Attribute 3.

Table 3.16

Field Test Design N

Examinee		Item Set																																				
		Attribute 1&2								Attribute 1&3								Attribute 2&3										Attribute 1&2				Attribute 1&3				Attribute 2&3				
250	**500**	S 1	S 2	S 3	S 4	S 5	S 6	S 7	S 8	S 9	S 10	S 11	S 12	S 13	S 14	S 15	S 16	S 17	S 18	S 19	S 20	S 21	S 22	S 23	S 24	S 25	S 26	C 1	C 2	C 3	C 4	C 5	C 6	C 7	C 8	C 9	C 10	C 11	C 12	C 13
1-125	**1-250**	X																										X												
126-250	**251-500**		X																										X											
251-375	**501-750**			X																										X										
376-500	**751-1000**				X																										X										
501-625	**1001-1250**					X																						X												
626-750	**1251-1500**						X																						X											
751-875	**1501-1750**							X																						X										
876-1000	**1751-2000**								X																						X										
1001-1125	**2001-2250**									X																						X									
1126-1250	**2251-2500**										X																						X								
1251-1375	**2501-2750**											X																						X							
1376-1500	**2751-3000**												X																						X						
1501-1625	**3001-3250**													X																			X							
1626-1750	**3251-3500**														X																			X						
1751-1875	**3501-3750**															X																			X					
1876-2000	**3751-4000**																X																			X				
2001-2125	**4001-4250**																	X																			X			
2126-2250	**4251-4500**																		X																			X		

Table 3.16 Continued

Examinee		**Item Set**																																						
		Attribute 1&2								**Attribute 1&3**								**Attribute 2&3**										**Attribute 1&2**				**Attribute 1&3**				**Attribute 2&3**				
250	**500**	S 1	S 2	S 3	S 4	S 5	S 6	S 7	S 8	S 9	S 10	S 11	S 12	S 13	S 14	S 15	S 16	S 17	S 18	S 19	S 20	S 21	S 22	S 23	S 24	S 25	S 26	C 1	C 2	C 3	C 4	C 5	C 6	C 7	C 8	C 9	C 10	C 11	C 12	C 13
2376-2500	**4751-5000**																						**X**																	
2501-2625	**5001-5250**																							X																
2626-2750	**5251-5500**																								X													X		
2751-2875	**5501-5750**																									X													X	
2876-3000	**5751-6000**																										X													X
3001-3125	**6001-6250**																												X											
3126-3250	**6251-6500**																													X										

Note. Each row represents a different form. X indicates that the item set is included in the form. S1-S26=simple structure item sets, C1-C13=complex structure item sets. Each set includes three items. S1-S8 include items from Attributes 1 and 2; S9-S16 include items from Attributes 1 and 3; S17-S26 include items from Attributes 2 and 3. C1-C4 include items from Attributes 1 and 2; C5- C8 include items from Attribute 1 and 3; C9 and C13 include items from Attribute 2 and 3.

Table 3.17

Field Test Design O

Examinee		Item Set																																	
		Attribute 1								Attribute 2												Attribute 3															
250	**500**	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1-250	**1-500**	X								X												X															
251-500	**501-1000**		X								X												X														
501-750	**1001-1500**			X								X												X													
751-1000	**1501-2000**				X								X												X												
1001-1250	**2001-2500**					X								X												X											
1251-1500	**2501-3000**						X								X												X										
1501-1750	**3001-3500**							X								X												X									
1751-2000	**3501-4000**								X								X												X								
2001-2250	**4001-4500**	X																X												X							
2251-2500	**4501-5000**		X																X												X						
2501-2750	**5001-5500**			X																X												X					
2751-3000	**5501-6000**				X																X												X				
3001-3250	**6001-6500**					X				X																								X				
3251-3500	**6501-7000**						X				X																								X			
3501-3750	**7001-7500**							X				X																								X		
3751-4000	**7501-8000**								X				X																									X

Note. Each row represents a different form. X indicates that the item set is included in the form. Set 1 contains Items 1 and 2, Set 2 contains Items 2 and 3, and the remaining sets continue in this pattern, each containing the next two items in the pool of 72 items. Sets 1-8 contain items that measure Attribute 1, Sets 9-20 contains items that measure Attribute 2, and Sets 21- 36 contains items that measure Attribute 3.

Table 3.18

Field Test Design P

Examinee		Item Set																																						
		Attribute 1 &2								Attribute 1&3								Attribute 2&3										Attribute 1&2&3												
250	500	S 1	S 2	S 3	S 4	S 5	S 6	S 7	S 8	S 9	S 10	S 11	S 12	S 13	S 14	S 15	S 16	S 17	S 18	S 19	S 20	S 21	S 22	S 23	S 24	S 25	S 26	C 1	C 2	C 3	C 4	C 5	C 6	C 7	C 8	C 9	C 10	C 11	C 12	C 13
1-125	1-250	X																														X								
126-250	251-500		X																														X							
251-375	501-750			X																														X						
376-500	751-1000				X																														X					
501-625	1001-1250					X																														X				
626-750	1251-1500						X																														X			
751-875	1501-1750							X																															X		
876-1000	1751-2000								X																															X	
1001-1125	2001-2250									X																															X
1126-1250	2251-2500										X																	X													
1251-1375	2501-2750											X																	X												
1376-1500	2751-3000												X																X												
1501-1625	3001-3250													X																X											
1626-1750	3251-3500														X																						X				
1751-1875	3501-3750															X																						X				
1876-2000	3751-4000																X																						X			
2001-2125	4001-4250																	X																						X		
2126-2250	4251-4500																		X																							X
2251-2375	4501-4750																			X								X													

Table 3.18 Continued

Examinee		Item Set																																						
		Attribute 1 &2								**Attribute 1&3**								**Attribute 2&3**										**Attribute 1&2&3**												
250	**500**	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21	S22	S23	S24	S25	S26	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
2376-2500	**4751-5000**																				X								X											
2626-2750	**5251-5500**																						X								X									
2751-2875	**5501-5750**																							X								X								
2876-3000	**5751-6000**																								X								X							
3001-3125	**6001-6250**																									X								X						
3126-3250	**6251-6500**																										X								X					

Note. Each row represents a different form. X indicates that the item set is included in the form. S1-S26=simple structure item sets, C1-C13=complex structure item sets. Each set includes 2 items. S1-S8 include items from Attributes 1 and 2; S9-S16 include items from Attributes 1 and 3; S17-S26 include items from Attributes 2 and 3. All C sets include items from all 3 attributes.

Table 4.1

Convergence Rates for each Field Test Design Condition

Q-matrix	Number of Attribute in a Form	Sample Size	Slot Length				
			72	**16**	**15**	**10**	**6**
Simple	2	250	NA	1	NA	1	1
		500	NA	1	NA	1	1
Simple	3	250	1	.99	1 - 1	1	1
		500	1	.99	1 - 1	1	1
Mixed	2	250	NA	.57	NA	.95	.16
		500	NA	.81	NA	.99	.82
Mixed	3	250	.29	.79	.37 - .35	.41	.14
		500	.43	.91	.90 - .94	.92	.17

Table 4.2

Mean Absolute Bias of Item Parameters for Field Test Designs with Simple Q-matrix

Q-matrix	Sample size	Condition	Number of Items & Attributes in a Form	Intercept	Main Effect
Simple	250	1	72 - 3	.011 (.021)	.030 (.031)
		5	15 - 3	.044 (.055)	.066 (.082)
		9	15 - 3	.043 (.054)	.064 (.079)
		13	16 - 2	.041 (.051)	.061 (.076)
		17	16 - 3	.043 (.053)	.064 (.079)
		21	10 - 2	.047 (.058)	.069 (.086)
		25	10 - 3	.047 (.058)	.070 (.086)
		29	6 - 2	.059 (.073)	.087 (.109)
		33	6 - 3	.077 (.096)	.117 (.146)
Simple	500	2	72 - 3	.006 (.014)	.015 (.022)
		6	15 - 3	.032 (.039)	.046 (.058)
		10	15 - 3	.031 (.038)	.045 (.056)
		14	16 - 2	.029 (.036)	.043 (.054)
		18	16 - 3	.031 (.038)	.045 (.056)
		22	10 - 2	.033 (.041)	.049 (.061)
		26	10 - 3	.033 (.041)	.049 (.061)
		30	6 - 2	.041 (.051)	.062 (.077)
		34	6 - 3	.052 (.066)	.079 (.099)

Note. Standard errors for mean absolute bias are given in parenthesis.

Table 4.3

Mean Absolute Bias of Item Parameters for Field Test Designs with Mixed Q-matrix

Q-Matrix	Sample Size	Condition	Number of Items-Attributes in a Form	Intercept	Main Effect 1	Main Effect 2	Main Effect 3	Interaction
Mixed	250	3	72 - 3	.024 (.023)	.038 (.049)	.035 (.043)	.035 (.042)	.059 (.074)
		7	15 - 3	.055 (.062)	.135 (.121)	.097 (.118)	.095 (.115)	.173 (.200)
		11	15 - 3	.051 (.059)	.126 (.114)	.091 (.111)	.089 (.108)	.158 (.187)
		15	16 - 2	.049 (.058)	.111 (.112)	.078 (.110)	.091 (.107)	.139 (.185)
		19	16 - 3	.051 (.058)	.126 (.113)	.091 (.111)	.090 (.108)	.158 (.187)
		23	10 - 2	.055 (.063)	.133 (.123)	.098 (.120)	.097 (.116)	.170 (.202)
		27	10 - 3	.058 (.066)	.138 (.129)	.144 (.125)	.135 (.122)	.182 (.213)
		31	6 - 2	.069 (.082)	.162 (.162)	.125 (.157)	.123 (.153)	.208 (.268)
		35	6 - 3	.069 (.080)	.159 (.158)	.121 (.153)	.122 (.149)	.204 (.259)
Mixed	500	4	72 - 3	.027 (.033)	.012 (.042)	.045 (.046)	.022 (.043)	.058 (.064)
		8	15 - 3	.039 (.044)	.105 (.085)	.068 (.083)	.064 (.081)	.119 (.140)
		12	15 - 3	.037 (.041)	.101 (.080)	.062 (.079)	.062 (.076)	.112 (.132)
		16	16 - 2	.038 (.786)	.100 (.470)	.063 (.754)	.062 (.473)	.111 (.711)
		20	16 - 3	.037 (.041)	.100 (.080)	.063 (.782)	.062 (.076)	.113 (.132)
		24	10 - 2	.040 (.044)	.106 (.087)	.068 (.085)	.067 (.082)	.120 (.143)
		28	10 - 3	.042 (.046)	.110 (.091)	.070 (.089)	.070 (.086)	.126 (.150)
		32	6 - 2	.049 (.058)	.126 (.114)	.087 (.111)	. 087 (.108)	.156 (.188)
		36	6 - 3	.047 (.056)	.123 (.112)	.084 (.108)	.087 (.104)	.149 (.183)

Note. Standard errors for mean absolute bias are given in parenthesis. The interaction terms between different pairs of attributes were not analyzed separately; thus, results represent the mean across all two-way interactions between each pair of attributes.

Table 4.4

Item Statistics Consistency Across Conditions with Simple Q-matrix – Compared with Corresponding Sample Size and 10000 Sample Size Full Data Condition

Q-Matrix	Condition	Number of Items in a form	Within Attribute Point Biserial						Difficulty	
			Attribute1		Attribute 2		Attribute 3			
			250	10000	250	10000	250	10000	250	10000
	5	15	.075 (.058)	.083 (.035)	.102 (.059)	.105 (.044)	.123 (.053)	.106 (.040)	.033 (.021)	.022 (.017)
	9	15	.094 (.070)	.108 (.047)	.109 (.058)	.111 (.042)	.098 (.053)	.080 (.036)	.035 (.022)	.027 (.019)
	13	16	.060 (.043)	.056 (.026)	.069 (.046)	.065 (.035)	.080 (.056)	.061 (.042)	.037 (.058)	.029 (.020)
	17	16	.086 (.058)	.086 (.034)	.109 (.059)	.111 (.048)	.098 (.053)	.080 (.036)	.036 (.022)	.027 (.020)
Simple	21	10	.073 (.067)	.083 (.049)	.087 (.057)	.085 (.051)	.122 (.069)	.106 (.045)	.028 (.025)	.019 (.015)
	25	10	.111 (.080)	.129 (.059)	.163 (.068)	.165 (.053)	.151 (.067)	.134 (.048)	.029 (.025)	.019 (.014)
	29	6	.127 (.077)	.143 (.053)	.176 (.070)	.177 (.050)	.189 (.072)	.172 (.049)	.034 (.027)	.028 (.019)
	33	6	.195 (.091)	.213 (.073)	.251 (.073)	.253 (.058)	.266 (.080)	.249 (.059)	.030 (.029)	.019 (.015)
			500	10000	500	10000	500	1000	500	10000
	6	15	.079 (.061)	.082 (.038)	.109 (.035)	.097 (.029)	.104 (.051)	.103 (.041)	.021 (.017)	.014 (.012)
	10	15	.100 (.060)	.103 (.040)	.103 (.037)	.090 (.031)	.084 (.043)	.082 (.033)	.022 (.015)	.016 (.011)
	14	16	.047 (.043)	.042 (.032)	.067 (.034)	.053 (.029)	.054 (.038)	.052 (.030)	.021 (.016)	.016 (.012)
	18	16	.078 (.058)	.084 (.038)	.104 (.036)	.091 (.030)	.084 (.043)	.081 (.033)	.021 (.015)	.016 (.010)
Simple	22	10	.069 (.053)	.071 (.035)	.101 (.047)	.088 (.039)	.095 (.036)	.094 (.032)	.022 (.017)	.018 (.012)
	26	10	.124 (.073)	.127 (.050)	.187 (.057)	.174 (.050)	.126 (.047)	.127 (.038)	.024 (.018)	.017 (.012)
	30	6	.134 (.082)	.137 (.061)	.184 (.049)	.171 (.041)	.171 (.056)	.172 (.048)	.022 (.017)	.017 (.011)
	34	6	.307 (.064)	.304 (.045)	.272 (.061)	.286 (.055)	.292 (.074)	.291 (.059)	.020 (.015)	.013 (.010)

Note. Standard errors for mean absolute difference are given in parenthesis.

Table 4.5

Item Statistics Consistency Across Conditions with Mixed Q-matrix – Compared with Corresponding Sample Size Full Data Condition

Q-Matrix	Condition	Number of Items in a Form	Within Attribute Point Biserial									Difficulty
			Simple Structure Items			Complex Structure Items						
			Attribute 1	Attribute 2	Attribute 3	Attribute 1&2		Attribute 1&3		Attribute 2&3		
						Attribute1	Attribute2	Attribute1	Attribute3	Attribute2	Attribute3	
Mixed	7	15	.066 (.053)	.074 (.054)	.065 (.051)	.093 (.050)	.101 (.039)	.058 (.038)	.070 (.052)	.054 (.053)	.069 (.037)	.028 (.020)
	11	15	.075 (.058)	.098 (.051)	.077 (.044)	.087 (.051)	.097 (.042)	.099 (.060)	.094 (.063)	.062 (.057)	.056 (.043)	.034 (.028)
	15	16	.048 (.021)	.069 (.055)	.066 (.044)	.093 (.039)	.095 (.036)	.065 (.055)	.071 (.067)	.069 (.049)	.059 (.044)	.029 (.024)
	19	16	.067 (.037)	.101 (.064)	.068 (.053)	.058 (.021)	.051 (.034)	.045 (.057)	.061 (.050)	.048 (.034)	.050 (.033)	.030 (.024)
	23	10	.059 (.048)	.092 (.064)	.080 (.054)	.120 (.051)	.118 (.035)	.089 (.058)	.082 (.046)	.077 (.043)	.083 (.054)	.029 (.019)
	27	10	.104 (.024)	.123 (.075)	.097 (.040)	.134 (.055)	.108 (.052)	.103 (.059)	.064 (.044)	.084 (.069)	.058 (.034)	.029 (.022)
	31	6	.124 (.056)	.116 (.067)	.132 (.062)	.187 (.062)	.154 (.041)	.155 (.092)	.131 (.081)	.110 (.068)	.098 (.055)	.033 (.027)
	35	6	.322 (.065)	.243 (.110)	.304 (.095)	.271 (.082)	.263 (.057)	.299 (.077)	.296 (.072)	.303 (.066)	.300 (.069)	.031 (.020)
Mixed	8	15	.075 (.050)	.052 (.036)	.057 (.047)	.053 (.035)	.079 (.031)	.049 (.030)	.061 (.034)	.064 (.030)	.059 (.031)	.022 (.017)
	12	15	.107 (.026)	.087 (.042)	.082 (.035)	.052 (.036)	.067 (.027)	.068 (.055)	.059 (.034)	.057 (.031)	.050 (.033)	.027 (.022)
	16	16	.054 (.038)	.036 (.019)	.059 (.029)	.041 (.041)	.066 (.039)	.051 (.037)	.058 (.035)	.028 (.024)	.036 (.026)	.027 (.021)
	20	16	.086 (.037)	.075 (.040)	.068 (.044)	.048 (.031)	.023 (.011)	.057 (.033)	.037 (.023)	.032 (.028)	.041 (.031)	.027 (.022)
	24	10	.088 (.047)	.062 (.046)	.074 (.043)	.081 (.050)	.090 (.043)	.077 (.053)	.075 (.048)	.075 (.048)	.076 (.048)	.024 (.016)
	28	10	.133 (.066)	.085 (.053)	.092 (.050)	.087 (.050)	.081 (.050)	.091 (.060)	.057 (.035)	.077 (.040)	.050 (.037)	.024 (.017)
	32	6	.148 (.089)	.099 (.042)	.109 (.057)	.140 (.045)	.130 (.044)	.155 (.068)	.133 (.058)	.116 (.067)	.098 (.052)	.022 (.017)
	36	6	.299 (.090)	.301 (.057)	.304 (.073)	.310 (.067)	.293 (.040)	.289 (.081)	.287 (.070)	.325 (.063)	.323 (.067)	.022 (.017)

Note. Standard errors for mean absolute difference are given in parenthesis.

Table 4.6

Item Statistics Consistency Across Conditions with Mixed Q-matrix – Compared with 10000 Sample Size Full Data Condition

Q-Matrix	Condition	Number of Items in a Form	Within Attribute Point Biserial									Difficulty
			Simple Structure Items			Complex Structure Items						
			Attribute 1	Attribute 2	Attribute 3	Attribute 1&2		Attribute 1&3		Attribute 2&3		
						Attribute1	Attribute2	Attribute1	Attribute3	Attribute2	Attribute3	
	7	15	.077 (.026)	.044 (.036)	.049 (.042)	.068 (.040)	.076 (.025)	.047 (.036)	.065 (.040)	.054 (.030)	.061 (.026)	.018 (.014
	11	15	.088 (.033)	.071 (.038)	.064 (.040)	.064 (.042)	.071 (.040)	.097 (.049)	.089 (.050)	.064 (.032)	.057 (.027)	.024 (.020
	15	16	.049 (.024)	.046 (.029)	.047 (.035)	.060 (.039)	.070 (.029)	.047 (.042)	.063 (.047)	.068 (.033)	.066 (.041)	.020 (.017
	19	16	.080 (.024)	.073 (.034)	.052 (.043)	.045 (.030)	.035 (.024)	.046 (.041)	.060 (.040)	.041 (.029)	.033 (.027)	.020 (.018
Mixed	23	10	.072 (.036)	.056 (.035)	.067 (.033)	.096 (.046)	.093 (.046)	.087 (.049)	.083 (.061)	.077 (.034)	.057 (.036)	.020 (.015
	27	10	.118 (.017)	.095 (.036)	.087 (.047)	.110 (.040)	.082 (.040)	.103 (.057)	.062 (.040)	.090 (.061)	.061 (.043)	.022 (.016
	31	6	.137 (.071)	.090 (.046)	.120 (.054)	.162 (.047)	.129 (.041)	.153 (.081)	.128 (.065)	.116 (.063)	.110 (.050)	.028 (.021
	35	6	.309 (.078)	.271 (.075)	.313 (.098)	.295 (.056)	.288 (.049)	.300 (.064)	.300 (.077)	.297 (.056)	.293 (.059)	.019 (.015
	8	15	.068 (.024)	.053 (.034)	.049 (.029)	.073 (.029)	.087 (.029)	.051 (.035)	.069 (.030)	.070 (.038)	.063 (.035)	.013 (.010
	12	15	.100 (.028)	.087 (.041)	.062 (.042)	.063 (.038)	.074 (.030)	.077 (.049)	.068 (.037)	.071 (.038)	.053 (.043)	.019 (.014
	16	16	.045 (.016)	.037 (.023)	.044 (.025)	.058 (.025)	.067 (.028)	.048 (.036)	.063 (.034)	.037 (.025)	.045 (.035)	.017 (.015
	20	16	.079 (.013)	.078 (.032)	.058 (.033)	.033 (.030)	.029 (.019)	.051 (.031)	.041 (.023)	.031 (.031)	.035 (.026)	.017 (.014
Mixed	24	10	.082 (.030)	.062 (.033)	.059 (.033)	.102 (.036)	.098 (.038)	.085 (.053)	.073 (.054)	.090 (.042)	.067 (.038)	.019 (.014
	28	10	.127 (.047)	.085 (.047)	.082 (.038)	.099 (.044)	.091 (.048)	.100 (.056)	.065 (.044)	.084 (.042)	.054 (.040)	.015 (.012
	32	6	.137 (.068)	.100 (.044)	.103 (.033)	.162 (.034)	.137 (.035)	.163 (.068)	.141 (.065)	.129 (.066)	.111 (.058)	.015 (.012
	36	6	.306 (.070)	.301 (.057)	.309 (.054)	.288 (.066)	.286 (.057)	.281 (.075)	.279 (.068)	.306 (.064)	.304 (.070)	.014 (.011

Note. Standard errors for mean absolute difference are given in parenthesis.

Table 4.7

Classification Accuracy Percentages Across Conditions with Simple Q-matrix

Q-matrix	Sample Size	Condition	Number of Items in Form	Number of Attributes in Form	Classification Accuracy %		
					Attribute 1	Attribute 2	Attribute 3
Simple	250	1	72	3	99.56	99.71	99.95
		5	15	3	93.71	91.29	91.90
		9	15	3	91.90	91.32	93.48
		13	16	2	84.72	88.35	90.74
		17	16	3	89.82	91.11	93.41
		21	10	2	82.18	84.17	88.90
		25	10	3	88.98	86.58	89.93
		29	6	2	79.14	80.50	84.03
		33	6	3	84.74	82.61	83.15
Simple	500	2	72	3	99.53	99.75	99.97
		6	15	3	93.76	91.22	92.04
		10	15	3	91.95	91.29	93.50
		14	16	2	84.68	88.36	90.78
		18	16	3	90.06	91.13	93.44
		22	10	2	82.21	84.19	88.90
		26	10	3	89.02	86.64	89.90
		30	6	2	79.88	81.15	84.58
		34	6	3	84.76	81.39	83.60

Table 4.8

Classification Accuracy Percentages Across Conditions with Mixed Q-matrix

Q-matrix	Sample Size	Condition	Number of Items in a Form	Number of Attributes in a Form	Classification Accuracy %		
					Attribute 1	Attribute 2	Attribute 3
Mixed	250	3	72	3	98.69	99.59	99.74
		7	15	3	89.92	91.10	90.68
		11	15	3	89.71	90.59	92.89
		15	16	2	87.66	87.92	87.82
		19	16	3	90.33	91.27	91.43
		23	10	2	82.92	84.76	86.96
		27	10	3	85.18	87.00	88.08
		31	6	2	78.84	81.60	81.35
		35	6	3	81.20	83.22	83.73
Mixed	500	4	72	3	98.80	99.56	99.84
		8	15	3	89.82	91.29	90.69
		12	15	3	89.81	90.66	92.96
		16	16	2	87.69	87.88	87.75
		20	16	3	90.45	91.05	91.46
		24	10	2	83.00	84.89	86.99
		28	10	3	86.14	90.11	85.11
		32	6	2	78.94	81.34	81.66
		36	6	3	81.28	82.95	83.93

REFERENCES

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). The NAEP 1998 technical report (NCES 2001-509). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

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

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of school psychology*, *48*(1), 5-37.

Beaton, A., and Zwick, R. (1992). Overview of the National Assessment of Educational Progress. Journal of Educational Statistics, 17, 95-109.

Bradshaw, L. (2014). PARCC Diagnostic Assessments: Design Research for Diagnostic Classification Model-based PARCC Diagnostic Assessments in Mathematics Comprehension and Decoding. Technical Report. Pearson Education.

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.

Clark, A., Kingston, N., Templin, J., & Pardos, Z. (2014). *Summary of results from the fall 2013 pilot administration of the Dynamic Learning Maps® Alternate Assessment System* (Technical Report No. 14-01). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

Every Student Succeeds Act, Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016).

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, *28*(3), 39-53.

Graham, J. W., Hofer, S. M., & Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. *NIDA research monograph*, *142*, 13-13.

Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Planned missing-data designs in analysis of change. In L. M. Collins & A. G. Sayer (Eds.), *Decade of behavior. New methods for the analysis of change* (pp. 335-353). Washington, DC, US: American Psychological Association.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*, 323–343. doi:10.1037/1082-989X.11.4.323.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*(2), 209-227.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74(2),* 191–210.

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29,* 95–110.

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.

Kirkpatrick, R., & Way, W. D. (2008, March). Field testing and equating designs for state educational assessments. In *annual meeting of the American Educational Research Association, New York, NY*.

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives, 7 (*4), 199-204. DOI: 10.1111/cdep.12043

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Martin, M.O., Mullis, I.V.S., & Foy, P. (2013). TIMSS 2015 assessment design. In I.V.S. Mullis & M.O. Martin (Eds), *TIMSS 2015 assessment frameworks* (pp. 85–99). Chestnut Hill, MA: International Association for the Evaluation of Educational Achievement.

Mullis, I.V.S., Cotter, K.E., Fishbein, B.G., Centurino, V.A.S. (2016). Developing the TIMSS 2015 achievement items. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), Methods and Procedures in TIMSS 2015 (pp.1.1-1.22). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timss.bc.edu/publications/timss/2015-methods/chapter-1.html

Messick, S. (1983). National Assessment of Educational Progress Reconsidered: A New Design for a New Era.

Muthén, L. K., & Muthén, B. O. (1998-2012). Mplus user's guide. (7th ed.). Los Angeles, CA: Muthén & Muthén.

Mullis, I. V. S., & Prendergast, C. O. (2017). Developing the PIRLS 2016 achievement items. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), Methods and Procedures in PIRLS 2016 (pp.1.1-1.29). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-1.html

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, 115 Stat/ 1449-1452 (2002).

Overview of NAEP Pre-Test Administration Types. (2011, March 14). Retrieved from https://nces.ed.gov/nationsreportcard/tdw/overview/naep_pretests_2009.aspx

Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, *90*, 54–63. doi:10.1080/01621459.1995.10476488.

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

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

Rupp, A. A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, 6:4, 219-262.

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

Schafer, J. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling*. Cambridge: Ballinger Publishing Company.

Sirotnik, K., and Wellington, R. (1977). Incidence sampling: An integrated theory for matrix sampling. Journal of Educational Measurement, 14, 343-399.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345-354.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305.

Welch, C. (2006). Item and Prompt Development in Performance Testing. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. (pp. 303-327) Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Wendler, C. L. W, & Walker, M. E. (2006) Practical Issues in Designing and Maintaining Multiple Test Forms for Large-Scale Programs. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. (pp. 445-468) Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement, 28,* 317–331.

Zeger, L.M., and Thomas, N. (1997). Efficient matrix sampling for correlated latent traits: Examples from the National Assessment of Educational Progress. Journal of the American Statistical Association, 92, 416-425.