Georgia State University
Digital Archive @ GSU

Educational Policy Studies Dissertations | Department of Educational Policy Studies

8-12-2009

# A Monte Carlo Study Investigating Missing Data, Differential Item Functioning, and Effect Size

Phyllis Lorena Garrett
pgarrett101@yahoo.com

Follow this and additional works at: http://digitalarchive.gsu.edu/eps_diss

Recommended Citation
Garrett, Phyllis Lorena, "A Monte Carlo Study Investigating Missing Data, Differential Item Functioning, and Effect Size" (2009). *Educational Policy Studies Dissertations.* Paper 35.

This Dissertation is brought to you for free and open access by the Department of Educational Policy Studies at Digital Archive @ GSU. It has been accepted for inclusion in Educational Policy Studies Dissertations by an authorized administrator of Digital Archive @ GSU. For more information, please contact digitalarchive@gsu.edu.

ACCEPTANCE

This dissertation, A MONTE CARLO STUDY INVESTIGATING MISSING DATA, DIFFERENTIAL ITEM FUNCTIONING, AND EFFECT SIZE, by PHYLLIS GARRETT, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree Doctor of Philosophy in the College of Education, Georgia State University.

The Dissertation Advisory Committee and the student's Department Chair, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty. The Dean of the College of Education concurs.

_____
Carolyn Furlow, Ph.D.
Committee Chair

_____
Dennis Thompson, Ph.D.
Committee Member

_____
Phillip E. Gagné, Ph.D.
Committee Member

_____
T. Chris Oshima, Ph.D.
Committee Member

_____
Date

_____
Sheryl A. Gowen, Ph.D.
Chair, Department of Educational Policy Studies

_____
R.W. Kamphaus, Ph.D.
Dean and Distinguished Research Professor
College of Education

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education's director of graduate studies and research, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

_____

Phyllis Garrett

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Phyllis Garrett
947 Lake Drive Court
Stone Mountain, GA 30088

The director of this dissertation is:

Dr. Carolyn Furlow
Department of Educational Policy Studies
College of Education
Georgia State University
Atlanta, GA 30303–3083

VITA

Phyllis Garrett

ADDRESS: 947 Lake Drive Court
Stone Mountain, GA 30088

EDUCATION:

| | | |
|---|---|---|
| Ph. D. | 2009 | Georgia State University<br>Educational Policy Studies |
| M.Ed. | 2003 | Georgia State University<br>Behavior-Learning Disabilities |
| B.A. | 1998 | Louisiana State University<br>Communication Disorders |

PROFESSIONAL EXPERIENCE:

| | |
|---|---|
| 2008 | Data Manager, Georgia State University, Atlanta |
| 2006-2008 | Cognitive Development Specialist, Georgia State University, Atlanta |
| 2002-2006 | Interrelated Special Education Teacher, DeKalb County School System, GA |
| 2001-2002 | Interrelated Special Education Teacher, Gwinnett County Public Schools, GA |

PROFESSIONAL ORGANIZATIONS:

| | |
|---|---|
| 2007-Present | American Educational Research Association |
| 2008-Present | National Council on Measurement in Education |

PRESENTATIONS AND PUBLICATIONS:

Garrett, P., Furlow, C. F., & Gagné, P. (2008, March). The impact of missing data and DIF detection method on the identification of DIF using polytomous IRT models. Poster presentation at the annual meeting of the American Educational Research Association, New York.

ABSTRACT

A MONTE CARLO STUDY INVESTIGATING MISSING DATA, DIFFERENTIAL ITEM FUNCTIONING, AND EFFECT SIZE
by
Phyllis Garrett

The use of polytomous items in assessments has increased over the years, and as a result, the validity of these assessments has been a concern. Differential item functioning (DIF) and missing data are two factors that may adversely affect assessment validity. Both factors have been studied separately, but DIF and missing data are likely to occur simultaneously in real assessment situations. This study investigated the Type I error and power of several DIF detection methods and methods of handling missing data for polytomous items generated under the partial credit model. The Type I error and power of the Mantel and ordinal logistic regression were compared using within-person mean substitution and multiple imputation when data were missing completely at random. In addition to assessing the Type I error and power of DIF detection methods and methods of handling missing data, this study also assessed the impact of missing data on the effect size measure associated with the Mantel, the standardized mean difference effect size measure, and ordinal logistic regression, the R-squared effect size measure. Results indicated that the performance of the Mantel and ordinal logistic regression depended on the percent of missing data in the data set, the magnitude of DIF, and the sample size ratio. The Type I error for both DIF detection methods varied based on the missing data method used to impute the missing data. Power to detect DIF increased as DIF magnitude

increased, but there was a relative decrease in power as the percent of missing data increased. Additional findings indicated that the percent of missing data, DIF magnitude, and sample size ratio also influenced the effect size measures associated with the Mantel and ordinal logistic regression. The effect size values for both DIF detection methods generally increased as DIF magnitude increased, but as the percent of missing data increased, the effect size values decreased.

# A MONTE CARLO STUDY INVESTIGATING MISSING DATA, DIFFERENTIAL ITEM FUNCTIONING, AND EFFECT SIZE

by
Phyllis Garrett

A Dissertation

Presented in Partial Fulfillment of Requirements for the
Degree of
Doctor of Philosophy
in
Educational Policy Studies
in
the Department of Educational Policy Studies
in
the College of Education
Georgia State University

Atlanta, GA
2009

Copyright by
Phyllis Garrett
2009

ACKNOWLEDGMENTS

There are many people I would like to thank that have supported me throughout my years in the doctoral program. I've had the support of my family from the beginning of this endeavor. My mother, Betty Garrett, my fiancé, Chris, brother, sister and niece have always given me the support and encouragement I needed to keep going through the good times and bad times. My family has supported me throughout this process, but there are several other individuals I must thank as well. Allison, Judy, Randi, Dominique, Denzel, and George were my good friends that helped me get to this point by allowing me to borrow their PCs for my data analysis. Because of your kindness, I was able to complete my dissertation within my scheduled timeframe. If it weren't for your selflessness, I would not have completed this study. You are definitely true friends.

TABLE OF CONTENTS

Page
List of Tables ..... iv
List of Figures ..... v
List of Abbreviations ..... vi

Chapter
1 INTRODUCTION ..... 1
Purpose of Study ..... 2
Significance of the Study ..... 2

2 LITERATURE REVIEW ..... 4
Polytomous Items ..... 4
Differential Item Functioning ..... 11
Effect Size Measures ..... 14
Polytomous DIF Detection Methods and Associated Effect Size Measures ..... 16
DIF Research for Polytonous Items ..... 26
Missing Data ..... 35
Statement of the Problem ..... 42

3 METHOD ..... 44
Study Design Conditions ..... 44
Data Generation ..... 49
Data Analysis ..... 52

4 RESULTS ..... 54
Type I Error Results ..... 55
Power ..... 59
Effect Size ..... 66

5 DISCUSSION ..... 74
Power ..... 74
Type I Error ..... 77
Standardized Mean Difference Effect Size ..... 78
R-squared Effect Size Measure ..... 79
Limitations and Future Research ..... 80
Summary ..... 81

References ..... 83
Appendixes ..... 90

LIST OF TABLES


| Table | | Page |
|---|---|---|
| 1 | An Example of a Contingency Table Found at Each Score Interval | 19 |
| 2 | Type I Error for Item 4 with No Missing Data and No Impact | 55 |
| 3 | Type I Error for Item 4 with No Missing Data and Impact | 56 |
| 4 | Type I Error for Item 4 with Missing Data and No Impact | 57 |
| 5 | Type I Error for Item 4 with Missing Data and Impact | 57 |
| 6 | Type I Error for Item 17 with No Missing Data and No Impact | 58 |
| 7 | Type I Error for Item 17 with No Missing Data and Impact | 58 |
| 8 | Type I Error for Item 17 with Missing Data and No Impact | 60 |
| 9 | Type I Error for Item 17 with Missing Data and Impact | 60 |
| 10 | Power for Item 4 with No Missing Data | 61 |
| 11 | Power for Item 4 with Missing Data | 62 |
| 12 | Power for Item 17 with No Missing Data | 64 |
| 13 | Power for Item 17 with Missing Data | 65 |
| 14 | Average SMDES for Item 4 with No Missing Data | 67 |
| 15 | Average SMDES for Item 17 with No Missing Data | 67 |
| 16 | Average SMDES for Item 4 with Missing Data | 68 |
| 17 | Average SMDES for Item 17 with Missing Data | 69 |
| 18 | Average R-Squared Effect Sizes for Item 4 with No Missing Data | 70 |
| 19 | Average R-Squared Effect Sizes for Item 17 with No Missing Data | 71 |
| 20 | Average R-Squared Effect Sizes for Item 4 with Missing Data | 72 |
| 21 | Average R-Squared Effect Sizes for Item 17 with No Missing Data | 73 |

LIST OF FIGURES

Figure Page

1 ICC for a Dichotomous Item.....5

2 Category Response Curves for a Five-Category Polytomous Item .....6

3 An Item Showing Uniform DIF .....23

4 An Item Showing Nonuniform DIF .....24

## ABBREVIATIONS

| | |
|---|---|
| ASA | Average Signed Area |
| CCA | Complete Case Analysis |
| DFIT | Differential Functioning of Items and Tests |
| DIF | Differential Item Functioning |
| HD | Hot-decking |
| ICC | Item Characteristic Curve |
| IRT | Item Response Theory |
| GMH | Generalized Mantel-Haenszel |
| GRM | Graded Response Model |
| GPCM | Generalized Partial Credit Model |
| LDFA | Logistic Discriminant Function Analysis |
| MAR | Missing At Random |
| MCAR | Missing Completely At Random |
| MCMC | Markov Chain Monte Carlo Method |
| MI | Multiple Imputation |
| MNAR | Missing Not At Random |
| PCM | Partial Credit Model |
| Poly-SIBTEST | Polytomous Simultaneous Item Bias Test |
| SIBTEST | Simultaneous Item Bias Test |
| SMD | Standardized Mean Difference |
| SMDES | Standardized Mean Difference Effect Size |
| WMS | Within-Person Mean Substitution |

# CHAPTER 1

# INTRODUCTION

The use of polytomous items in assessment has greatly increased over the years (Dodeen, 2004; Welch & Hoover, 1993; Zwick, Donoghue, & Grima, 1993). Although the use of polytomous items has increased, there has been limited research regarding the validity of these items and the validity of the measures that incorporate these items (Dorans & Schmitt, 1991; Welch & Miller, 1995; Zwick et al., 1993). To determine a measure's validity, assessment developers often conduct validation studies. These studies are conducted to assemble evidence regarding the strength of an assessment's inferences (Crocker & Algina, 1986). If the assembled evidence indicates a measure has strong inferences, then that measure is labeled highly valid. A measure's validity, however, can be adversely affected by a number of factors. One such factor is differential item functioning (DIF). DIF is an unexpected performance difference among examinees with the same ability. The unexpected performance difference may cause inaccurate trait inferences for certain examinees, in turn, adversely affecting test validity. DIF, however, is not the only factor that may lead to a reduction in test validity.

Missing data, the omission of item values, is another factor that may lead to inaccurate trait inferences. Assessment developers often utilize various statistical procedures to replace examinees' missing values. These methods of handling missing data provide measurement information needed to calculate examinees' scores and reduce faulty assessment inferences. The impact of DIF and missing data has been studied

separately in simulation research, but with real data, they are likely to occur concurrently. To reduce the potential threats of DIF and missing data to test validity, it is important to determine the effectiveness of DIF detection methods in the presence of missing data. In many DIF analyses, DIF detection methods are used in conjunction with effect size measures, measures that calculate the size of DIF. Since effect size measures are frequently used in conjunction with DIF detection methods, it is equally important to determine the effectiveness of effect size measures in the presence of missing data.

## Purpose of the Study

The purpose of this Monte Carlo simulation study was to investigate DIF detection methods and methods of handling missing data when DIF and missing data occur simultaneously. This study compared the Type I error and power of several DIF detection methods and methods of handling missing data for polytomous items generated under the partial credit model. The DIF detection rates of the Mantel and ordinal logistic regression (OLR) were compared using within-person mean substitution and multiple imputation when data were missing completely at random. In addition to assessing the Type I error and power of DIF detection methods and methods of handling missing data, this study assessed the impact of missing data on effect size measures associated with specific DIF detection methods. Since effect size measures are frequently used with DIF detection methods, this study investigated the impact of missing data on effect size measures associated with the Mantel and OLR.

## Significance of the Study

Simulation studies investigating the performance of DIF detection methods for polytomous items has increased over the years, however, there had yet to be a study

investigating the performance of DIF detection methods in the presence of missing data. Missing data are a common occurrence in real assessment situations, and there are several methods analysts can choose to handle missing data. The type of missing data and the method used to handle the missing data can have a unique effect on statistical outcomes. As a result, it is important to determine how missing data and the methods used for handling missing data will affect DIF detection. Effect size measures are tools analysts often use when conducting DIF analyses, and to date, there had been no simulation research assessing how missing data may impact effect size measures for polytomous DIF detection methods.

It is hoped that the results of this study will provide insight into the effectiveness of the Mantel and OLR when there are missing data handled with within-person mean substitution and multiple imputation. In addition to determining the effectiveness of DIF detection in the presence of missing data, this study will also provide insight into how missing data may affect the effect size measures associated with the Mantel and OLR, the standardized mean difference effect size measure and the R-squared effect size measure.

# CHAPTER 2

# LITERATAURE REVIEW

This chapter begins by describing the use of polytomous items in measurement and the item response theory (IRT) models used to model polytomous data. The next sections will describe DIF, effect size measures, polytomous DIF detection methods and their related effect size measures, missing data mechanisms, and common methods for handling missing data.

## Polytomous Items

In recent years, polytomous items have been frequently chosen over dichotomous items to assess certain traits. Although dichotomous items have a long history in measurement, polytomous items have the advantage of providing more information about the underlying latent trait because of their multiple response categories (Ostini & Nering, 2006). This unique characteristic of polytomous items can be explained using an IRT framework. In IRT, examinee performance on a dichotomous item is modeled graphically using an item characteristic curve (ICC). An ICC models the probability of a correct response given an examinee's ability and item characteristics.

The ICC for an item is contained at the point on the ability scale where an examinee has a 0.5 probability of getting the item correct. Consequently, a dichotomous item provides the most information for only one ability level. Figure 1 shows an ICC for a dichotomous item generated under the simplest IRT model, the one-parameter logistic

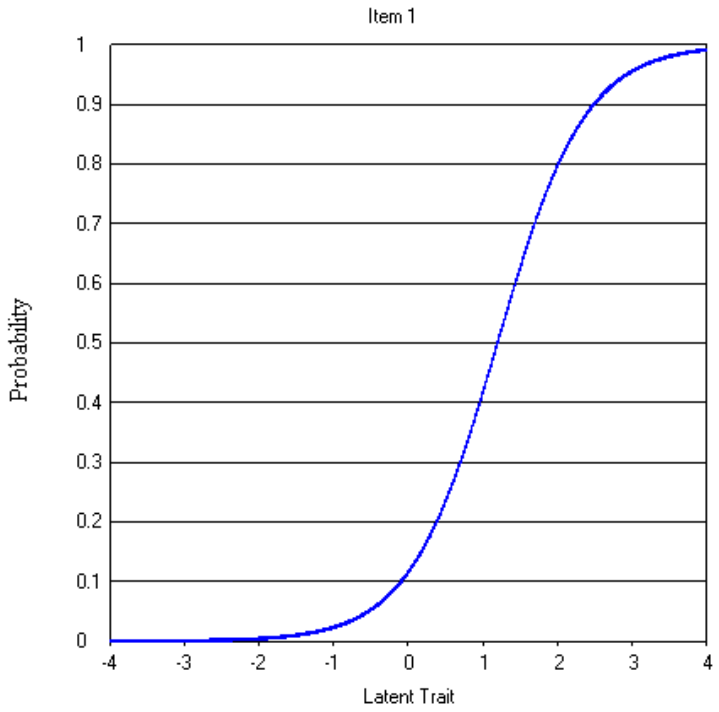


*Figure 1.* ICC for a dichotomous item where b = 1.0.

model. The item's difficulty, or *b*- parameter, is the only characteristic that describes an item generated under this model, and it indicates an ICC's location on the latent trait scale. In IRT models, a small *b*- parameter value indicates less ability is required to answer an item correctly, whereas a large *b*- parameter value indicates more ability is required to answer an item correctly.

A polytomous item, unlike a dichotomous item, has multiple response categories, and examinee performance on a polytomous item is modeled graphically with category response curves (Embretson & Reise, 2000). The category response curves for a polytomous item are contained at various ability levels, thus a polytomous item provides multiple pieces of information along the trait continuum (Ostini & Nering, 2006). Figure 2 displays category response curves for a polytomous item. This polytomous item

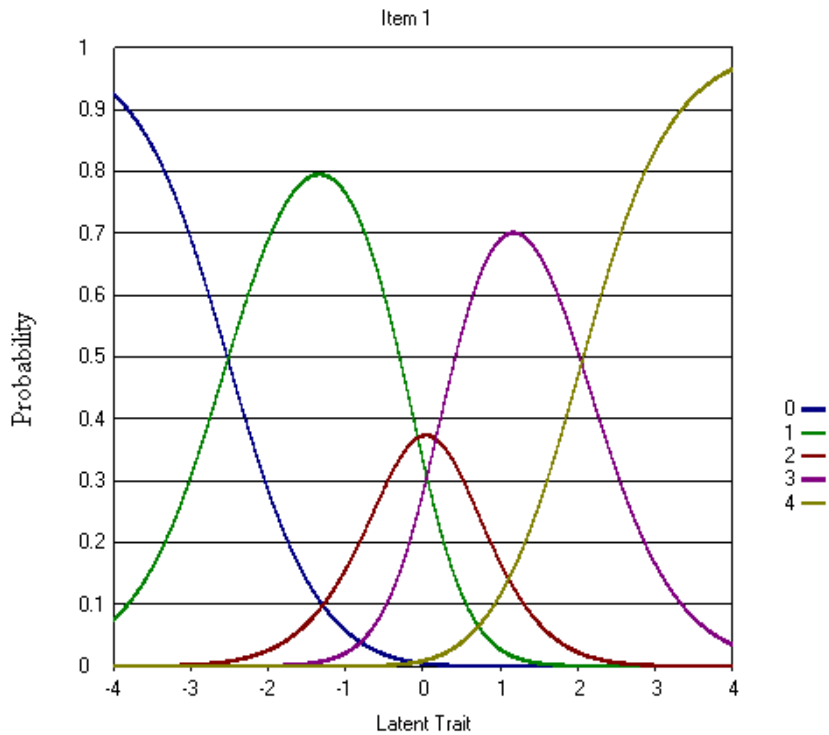


*Figure 2.* Category response curves for a five-category polytomous item where $b_1 = -2.519$, $b_2 = -0.063$, $b_3 = 0.170$, and $b_4 = 2.055$.

has multiple *b*-parameters that are located at each category response curve intersection. Each *b*-parameter value indicates the point on the latent trait scale where a category response becomes more likely for an examinee with a specific ability level (Embretson & Reise). This greater measurement capacity makes polytomous items more attractive than dichotomous items in many settings.

Two settings in which polytomous items are frequently used are ability assessment and attitude assessment. Ability assessments measure cognitive traits such as reading comprehension, written expression, or math. These assessments can be comprised entirely of polytomous items or contain a mixture of dichotomous and

polytomous items (Shaeffer, Henderson-Montero, Julian, & Bené, 2002; Thissen & Wainer, 2001). Assessment developers often take advantage of both types of items on assessments because dichotomous items assess more content across an academic domain and are less time-consuming to score, whereas polytomous items more effectively assess an examinee's intricate understanding of a specific concept or skill within a content area, but typically require more time to score (Dorans & Schmitt, 1991; Sykes & Hou, 2003; Thissen & Wainer, 2001).

Unlike ability assessments, attitude assessments usually contain only polytomous items. Their widespread use in attitude assessment is the result of the type of traits these assessments measure. Attitude assessments (e.g., surveys and questionnaires) are used to gauge an individual's feeling or stance on a position (Dodeen, 2004). Ostini and Nering (2006) called these types of assessments predilection measures, because an individual's typical performance is measured not his or her maximum performance. Polytomous items are vital in predilection measures because dichotomous responses of correct or incorrect are inappropriate responses to indicate feelings and opinions, and while dichotomous responses of true or false could be used for attitude assessment, they do not provide the appropriate range of choices for respondents (Ostini & Nering, 2006). As a result, polytomous items are typically chosen for these types of assessments.

Polytomous items are available in a variety of presentation formats. Ability assessments typically contain constructed response or performance task items. These item formats require examinees to complete some type of lengthy work to demonstrate their knowledge or understanding of a concept or skill. The constructed response item requires examinees to write a response to a question or statement, whereas the performance task

item requires examinees to create a unique product to demonstrate their ability to evaluate, analyze, and synthesize information. The essay and student portfolio, respectively, are examples of these types of items. A third type of polytomous item often used in ability assessment is the multistep item. The multistep item requires examinees to solve a multistep problem showing their work at each problem solving step. The multistep item is typically used for math assessment, but it can also be used to assess science skills as well (Dodd, DeAyala, & Koch, 1995). Constructed response, performance task, and multistep items are categorized as polytomous items because they are scored using a multiple-category scale. Scores for these items are typically determined by raters who assess examinees' written responses, products, or problems and indicate the degree of correctness or level of understanding of a concept or skill using a multiple-category scale.

The polytomous items in attitude assessments are presented and scored quite differently than items in ability assessments. Items in attitude assessments typically contain a statement in which respondents rate their level of agreement or disagreement with a statement using a multiple-category scale. The responses on the scale may be constant for all items or they may vary from item to item. Examples of this type of assessment are Likert-type items in which statements are presented and respondents indicate a response to these statements on a continuum from strongly agree to strongly disagree.

After items have been scored, the data are analyzed using a specific measurement framework. The framework frequently chosen to model examinee or respondent data on both dichotomous and polytomous items is IRT (see Hambleton, Swaminathan, &

Rogers, 1991 for the advantages of IRT). There are a series of polytomous IRT models available when item responses have an ordered scale. A commonly used polytomous model for ordered responses is the partial credit model (PCM). This model is frequently used to model polytomous data, because it is a Rasch family model.

Polytomous Rasch family models have two main advantages. The first advantage of these models is that only one equation is needed to estimate model parameters. These 'direct' models (e.g., the PCM (Masters, 1982) and Andrich's rating scale model (Andrich, 1978)) are less computationally complicated than the 'indirect' models (e.g., the graded response model (Samejima, 1969) and the generalized partial credit model (Muraki, 1992)), which require two equations in a two-step process to determine the probability of an examinee or respondent responding in a particular category (Embretson & Reise, 2000; Ostini & Nering, 2006). The second advantage of polytomous Rasch family models, which is unique to this group of models, is that the total score is a sufficient statistic to estimate examinee or respondent performance (Embretson & Reise, 2000; Thissen & Wainer, 2001). Consequently, examinees or respondents with the same total score possess the same amount of the studied trait (see Ostini & Nering, 2006; Rasch, 1960; and Thissen & Wainer, 2001 for a detailed description of the Rasch model).

*The Partial Credit Model*

The PCM is a polytomous Rasch family IRT model primarily used for items that can be given partial credit. These items are typically multistep problems, but they can also be attitudinal items in which respondents rate their feelings and opinions using a multiple-category scale (Masters & Wright, 1996). Items generated under the PCM must have an ordered score scale. DeAyala (1993) provided an example of a partial credit item

using this multistep problem: (6/3) + 2. In his example, DeAyala stated that in order to solve this problem and receive full credit, an examinee must complete two steps. Step1 involves solving the division portion of the problem, 6/3, and step 2 involves solving the addition portion of the problem, 2 + 2. To solve this problem correctly, these steps must be done in order. DeAyala continued by stating that if only step 1 is answered correctly, then partial credit is awarded, but if neither step 1nor step 2 is answered correctly, then no credit is awarded. Using this scoring scale, this problem possesses three response categories: 0, 1, 2. Category 0 is no credit, category 1 is partial credit, and category 2 is full credit. DeAyala maintained that step 1 is more difficult than step 2 (division is more difficult than adding), but ordering from easiest to hardest is not a requirement of the model. The only requirement to use the PCM is that each item has ordered steps.

The PCM has two parameters: a person parameter and an item parameter. The person parameter, θ (theta), is the latent trait being assessed, whereas the item parameter, δ (delta), is the item step difficulty. The item step difficulty parameter is the "point on the latent trait scale at which two consecutive category response curves intersect" (Embretson & Reise, 2000, p. 106), and as a result, it is often called the category intersection parameter. The category intersection parameter represents the amount of difficulty involved in choosing one response category over the next response category. For any PCM item, there will be one less category intersection parameter than the number of response categories. For example, an item with four response categories will have three category intersection parameters. This characteristic holds for any PCM item.

For the PCM, the probability of responding in a specific response category can be written as

$$P_{ix}(\theta) = \frac{\exp[\sum_{j=0}^{x}(\theta - \delta_{ij})]}{\sum_{r=0}^{m_i}[\exp\sum_{j=0}^{r}(\theta - \delta_{ij})]}, \quad (1)$$

where $\delta_{ij}$ is the category intersection parameter for a specific category response *x* for item *i* and *j* is an examinee's category response curve score for item *i*. Equation 1 states that "the probability of an examinee responding in category *x* on a $m_i$ step item is a function of the difference between an examinee's trait level and a category intersection parameter" (Embretson & Reise, 2000; p.106). For a PCM item, there are $m_i$ number of category intersection parameters for $m_i$+1 category responses.

Both dichotomous and polytomous items are often subject to factors that can lead to differential group performance. Groups of examinees may have differential item performance due to differing ability levels (i.e., impact), but when individuals with the same ability from different groups have different item performance, then DIF is present.

## Differential Item Functioning

DIF is defined as an item-level performance difference between groups of examinees or respondents matched on a latent trait. Subgroups typically studied in DIF analyses are minority ethnic groups such as African-Americans, Hispanics, and Asians, but subgroups based on other examinee characteristics such as gender, religion, and socioeconomic status are also studied. For ability testing, DIF is defined as an item level performance difference between groups of examinees matched on ability. The definition of DIF for surveys (an attitude measure) is slightly different because respondents on surveys are matched on overall agreement level instead of ability (Dodeen, 2004).

DIF is typically identified using inferential DIF detection methods. Inferential DIF detection methods use a significance test to determine if an item possesses DIF. The numerical value obtained from the inferential DIF method indicates that an item is more difficult for a particular subgroup than originally intended (Camilli & Shepard, 1994). For attitude items, DIF indicates that a particular subgroup responded more positively to an item than another subgroup. Dodeen and Johansson (2003) stated that "the correct answer in the cognitive context is similar to the positive affect of attitude toward the item" (p. 129).

Items on ability and attitude assessments may exhibit DIF for several reasons. Assessment developers must evaluate the DIF item to determine the cause of DIF (i.e., the source of DIF). Item evaluation typically involves a panel review of the item to determine possible reasons why examinees or respondents may have answered or endorsed an item differently. If the source of DIF cannot be explained after the item is evaluated, Camilli and Shepard (1994) stated that the item was likely flagged for DIF due to a Type I error of the statistical method used to detect DIF. Although Type I error can account for unexplained DIF, it is likely that other reasons could be hypothesized as to why the DIF could not be explained. For example, DIF may go unexplained due to a lack of understanding of the latent trait being measured, or it may also go unexplained due to a lack of understanding of the examinee population taking the assessment. DIF that can be explained, however, can be interpreted to be caused by item multidimensionality.

Item multidimensionality occurs when an item simultaneously measures two or more constructs. Shealy and Stout (1993) posed a theory stating that DIF results from an item measuring an additional construct or dimension. According to this theory of DIF,

examinees have the same ability or agreement level on the matching dimension (the primary dimension of interest), but differ in their ability or agreement level on the second dimension (Gierl, 2005). Ackerman (1992) noted that it is impossible to match examinees simultaneously on two dimensions without special scaling. Consequently, if special scaling is not performed, differing distributions on the multiple dimensions will lead to DIF. The presence of DIF, however, is not always detrimental to a group of examinees or respondents trait inferences. For DIF to affect examinee or respondent inferences adversely, the second dimension must be caused by a specific factor.

The two factors that can cause item multidimensionality are a construct relevant factor and a construct irrelevant factor. A construct relevant factor has no adverse affect on item measurement, whereas a construct irrelevant factor changes the measurement intent of an item (Camilli & Shepard, 1994). Gierl (2005) illustrated a construct relevant factor and a construct irrelevant factor using a math example. In his example, Gierl stated that a construct relevant factor associated with a math item might be critical thinking. Critical thinking, Gierl maintained, is an intentional or related factor measured by the math item. In order to answer the math item, a related secondary dimension, critical thinking, will also be assessed with the primary dimension, math ability.

In contrast to the construct relevant factor, a construct irrelevant factor is an unintended factor measured by an item that undermines the intended construct an item is suppose to measure (Camilli & Shepard, 1994). In his math example, Gierl stated that a construct irrelevant factor may be test-taking ability. Assessing test-taking skills is not the intent of this math item, thus the multidimensionality in this item may have an adverse affect on a particular group of examinees. In Gierl's example, DIF items were

evaluated to determine the cause of DIF. This is true in any DIF analysis. Thus, after an item is evaluated and the cause of DIF is determined to be from a construct irrelevant factor, that item is labeled biased.

Bias is "systematic error in how a test measures for members of a particular group" (Camilli & Shepard, 1994, p. 8). Bias is important to identify because it has a detrimental affect on examinee scores if undetected. Kamata and Vaughn (2004) stated that bias disadvantages a subgroup and results in test score differences between groups of examinees with the same ability. Affected examinees may be shown to possess higher or lower trait levels than is actually true. Typically, a biased item is either removed from an assessment or altered to eliminate its negative effect on examinee inferences.

Not all items flagged as statistically significant for DIF are evaluated to determine if they are biased. Frequently, after items are flagged for DIF, an additional step is taken to measure the size of DIF (i.e., the amount of DIF an item possesses). Measures that estimate the size of DIF are called descriptive DIF methods or effect size measures. Potenza and Dorans (1995) stated that "descriptive measures of an item's degree of DIF are essential to DIF assessment" (p. 28).

## Effect Size Measures

Effect size measures are statistical tools used to determine the practical significance of DIF. Practically significant DIF is DIF that is large enough to possibly have a substantial effect on a group of examinees or respondents. Many times, items are flagged as having statistically significant DIF, but the item contains a small amount of DIF. Small amounts of DIF may not have much impact on examinee or respondent scores, whereas large amounts of DIF will likely have more impact on examinee or

respondent scores (Jodoin & Gierl, 2001; Meyer, Huynh, & Seaman, 2004; Zumbo, 1999). Due to the possibility of flagging items with small amounts of DIF, inferential DIF detection methods are often combined with effect size measures to determine the practical significance of DIF. In this case, the effect size measure is used as a complement to an inferential DIF detection method.

There are times, however, when effect size measures are used as the sole method of DIF detection. This situation frequently occurs when a large sample is analyzed for DIF. When a DIF analysis is conducted on a large sample, there is a tendency for many of the items to exhibit statistically significant DIF (Kim, Cohen, Alagoz, & Kim, 2007; Kirk, 1996; Zumbo, 1999). Due to the increase in statistical power resulting from the large number of individuals in the sample, small performance differences between groups can lead to high levels of statistically significant DIF items (Kim, Cohen, Alagoz, & Kim, 2007; Zumbo, 1999). Jodoin and Gierl (2001) stated that since the significance tests of inferential DIF detection methods are sensitive to sample size, it is important to use effect size measures as the indicator of DIF. Currently, there is no rule stating how large a sample must be in order to use an effect size measure as the only DIF detection method.

After the effect size of a DIF item is calculated whether the effect size measure is used with an inferential DIF method or as the sole DIF detection method, the item is placed into one of three size categories to determine the practical significance of the DIF. These effect size categories are predetermined size guidelines associated with a specific effect size measure. The names of the size categories may vary, but the categories always indicate whether an item has a small, medium, or large amount of DIF. After an item is

placed into an effect size category, the item may be referred for further evaluation as a potentially biased item.

Although categorizing DIF is necessary to determine practical significance, the determination of which effect size values will be categorized as small, medium, or large is likely to be context specific (Cohen, 1977). An assessment developer for a statewide exam, for example, may decide an effect size value should be placed in the large DIF category, whereas an assessment developer for a schoolwide attitude inventory may decide this same effect size value should be placed in the medium DIF category. Since the effect size values related to a small, medium, and large DIF category may vary depending on the assessment context, the interpretation of the practical significance of the DIF will also vary.

There have been various effect size measures developed for polytomous items, but many of these measures do not have guidelines to interpret the size of DIF. Potenza and Dorans (1995) stated that "to be used effectively, a DIF detection technique needs an interpretable measure of the amount of DIF" (p. 33). When an effect size measure does not have interpretable size guidelines, faulty conclusions about the impact of DIF on examinees or respondents may occur.

## Polytomous DIF Detection Methods and Associated Effect Size Measures

To reduce the effect of DIF on assessment validity, it is necessary to identify potentially biased items. In order to determine item bias, an item must first be found to be functioning differently between groups, and DIF detection methods are typically used to determine differential item performance. There are numerous DIF detection methods available for practitioners (see Millsap & Everson, 1993, and Potenza & Dorans, 1995 for

a review of DIF detection methods). Potenza and Dorans (1995) organized DIF detection methods by (a) the presence or absence of a specified model illustrating a relationship between a latent trait and the matching variable and (b) the type of matching variable. DIF detection methods that do not specify a relationship between the latent trait and the matching variable are called nonparametric methods. Nonparametric methods are model-free methods. On the other hand, DIF detection methods that specify a relationship between the latent trait and the matching variable are called parametric methods. Parametric DIF detection methods are model-based methods.

The most frequently used matching variable for nonparametric and parametric methods is the latent trait. Typically, the latent trait is obtained from the assessment in which items are being analyzed for DIF. The latent trait is represented by either an observed score or a latent trait estimate. Observed score DIF detection methods use a total test score as the matching variable, whereas latent trait DIF detection methods use a latent trait measure as the matching variable (Potenza & Dorans, 1995). When the two types of DIF detection methods are crossed, four different categories of DIF detection methods are created: (a) nonparametric observed score methods such as the Mantel-Haenszel test; (b) parametric observed score methods such as logistic regression; (c) nonparametric latent trait methods such as Simultaneous Item Bias Test (SIBTEST); and (d) parametric latent trait methods such as the one-parameter logistic model (Penfield & Lam, 2000).

Analysts choose the DIF detection method appropriate for their assessment, and two commonly chosen polytomous DIF detection methods are the Mantel and OLR. These DIF detection methods are polytomous extensions of two popular DIF detection

methods for dichotomous items: the Mantel-Haenszel test (Holland & Thayer, 1988) and logistic regression. These two polytomous DIF detection methods require inexpensive software, little computational intensity, and ease of implementation (Mapuranga, Dorans, & Middleton, 2008). The Mantel and OLR as well as their associated effect size measures are described below.

*The Mantel Test*

The Mantel test (Mantel, 1963), a nonparametric observed score method; tests for DIF under the assumption that responses on the score scale are ordered. The test provides a statistic with a chi-square distribution with one degree of freedom when the null hypothesis of no DIF is true (Meyer, Huynh, & Seaman, 2004; Zwick, Donoghue, & Grima, 1993). This method uses a 2 x $T$ x $K$ contingency table to test for DIF, where $T$ is the number of response categories in an item, and $K$ is the number of score interval levels. With the contingency table methods, one 2 x $T$ table is constructed at every score interval level. Various score interval levels must be created to match examinees or respondents, and the researcher typically determines the number of score interval levels needed for the DIF analysis.

Consider, for example, a 20-item test with 4 response categories per item for which the score categories are 1, 2, 3, and 4. The total score for this test can range from 20 through 80. A researcher may create score interval levels using each total score in this range (i.e., thin matching) or a researcher may create score interval levels by combining several total scores to create wider score interval levels (i.e., thick matching). These score levels are typically constructed such that there is at least one observation in each score category (Donoghue & Allen, 1993). Table 1 shows an example of a contingency table

Table 1

*An Example of a Contingency Table Found at Each Score Interval Level*

| | Item Score | | | | | |
|---|---|---|---|---|---|---|
| Group | $y_1$ | $y_2$ | $y_3$ | .... | $y_T$ | Total |
| Reference | $n_{R1k}$ | $n_{R2k}$ | $n_{R3k}$ | .... | $n_{RTk}$ | $n_{R+k}$ |
| Focal | $n_{F1k}$ | $n_{F2k}$ | $n_{F3k}$ | .... | $n_{FTk}$ | $n_{F+k}$ |
| Total | $n_{+1k}$ | $n_{+2k}$ | $n_{+3k}$ | .... | $n_{+Tk}$ | $n_{++k}$ |

*Note.* This table was taken from Zwick et al. (1993).

that is constructed at each score interval level, where $y_1, y_2, \ldots, y_k$ are the category responses for an item, *R* represents the reference group, *F* represents the focal group, and '+' is the summation of a particular row or column at a specific score interval level.

The equation for the Mantel statistic is

$$\chi^2_{Mantel} = \frac{[\sum_K F_K - \sum_K E(F_K)]^2}{\sum_K Var(F_k)}, \tag{2}$$

where $F_k$ is the total of the focal group scores for the $k^{th}$ level of the matching variable, $E(F_k)$ is the expected value for the focal group, and $Var(F_k)$ is the variance the focal group scores.

The equation for $F_k$ is

$$F_k = \sum y_k n_{FTk}\ ; \tag{3}$$

for $E(F_k)$ is

$$E(F_k) = \frac{n_{F+k}}{n_{++k}} \sum y_k n_{+Tk} \; ; \tag{4}$$

and for $Var(F_k)$ is

$$Var(F_k) = \frac{n_{R+k} n_{F+k}}{n_{++k}^2 (n_{++k} - 1)} \{ (n_{++k} \sum y_T^2 n_{+Tk}) - (\sum y_T n_{+Tk})^2 \} . \tag{5}$$

The null hypothesis for the Mantel test is that there is no association between the row mean scores of the studied group (i.e., the focal group) and the comparison group (i.e., the reference group). If the focal group has a lower row mean score than the reference group, then the focal group did not perform as well on the studied item as the reference group. Conversely, a higher row mean score for the focal group indicates the focal group performed better than the reference group on the studied item. A difference between the row mean scores at any score interval level indicates the presence of DIF for the studied item.

*Standardized Mean Difference Effect Size Measure*

The standardized mean difference effect size measure (SMDES) is a common descriptive measure of DIF. This measure is related to the standardized mean difference (SMD) DIF detection method. The SMD (Dorans & Schmitt, 1991; Zwick & Thayer, 1996) is a descriptive DIF detection procedure that describes the difference between the mean item score of the focal group and the weighted mean item score of the reference group. Having a weighted mean item score for the reference group allows for a comparison of similar numbers of focal and reference group members (since there are usually fewer members in the focal group than in the reference group). The equation for the SMD is

$$SMD = \overline{x_F} - \sum \hat{p}_{Fk} \overline{x_{Rk}} \ , \tag{6}$$

where $\overline{x_F}$ is the mean item score for all the focal group members on the studied item, $\overline{x_{Rk}}$ is the mean item score of the reference group in the $k^{th}$ score interval level, and $\hat{p}_{Fk}$ is the proportion of individuals in the focal group in the $k^{th}$ score interval level (Meyer, Huynh, & Seaman, 2004). A negative SMD value indicates the focal group has a lower mean item score than the reference group and did not perform as well as the reference group on the item, whereas a positive SMD value indicates the focal group has a higher mean item score than the reference group and performed better than the reference group on the item (Zwick et al., 1993).

To obtain the SMDES estimate, the SMD value is divided by the standard deviation of the reference and focal groups. The equation for the SMDES is

$$SMDES = \frac{SMD}{S_{CG}} \ , \tag{7}$$

where SMD is the value of the standardized mean difference DIF method and $S_{CG}$ is the pooled standard deviation of the reference group and focal group (Meyer, Huynh, & Seaman, 2004).

ETS developed classification guidelines for the SMDES that have been used by practitioners to determine the practical significance of DIF. ETS's classification guidelines for the SMDES are:

1. Small or *A* level DIF: the absolute value of the effect size is ≤ 0.17,
2. Medium or *B* level DIF: the absolute value of the effect size is > 0.17 and ≤ 0.25,
3. Large or *C* level DIF: the absolute value of the effect size is > 0.25.

*Ordinal Logistic Regression*

OLR (Agresti, 1984; French & Miller, 1996) tests for the presence of DIF using a prediction equation. OLR is a parametric observed score method that specifies a specific parametric model for the relationship between the latent trait and the matching variable. The OLR equation uses the total score, group membership, and the interaction between the total score and group membership to test for DIF (Zumbo, 1999). The advantage of this method is that both uniform and nonuniform DIF can be modeled within the same equation (French & Miller, 1996; Zumbo, 1999).

Su and Wang (2005) described uniform and nonuniform DIF in terms of ability and group membership. They stated that uniform DIF indicates no relationship between the ability or agreement level of examinees or respondents and group membership. When an item possesses uniform DIF, all the individuals in one group perform better on an item than all the individuals in the other group. The equation used to test for uniform DIF is shown in Equation 8.

$$Y' = \alpha + b_1 \text{ totscore} + b_2 \text{ group}. \tag{8}$$

In Equation 8, $\alpha$ represents the intercept, $b_1$ represents the slope for the total score variable, and $b_2$ represents the slope for the grouping variable. An illustration of uniform DIF is shown in Figure 3. The space between the curves in Figure 3 indicates the presence of uniform DIF for that item. In Figure 3, the examinees in group 1 performed better on this item than the examinees in group 2 which indicates that this item was easier for group 1 than group 2.

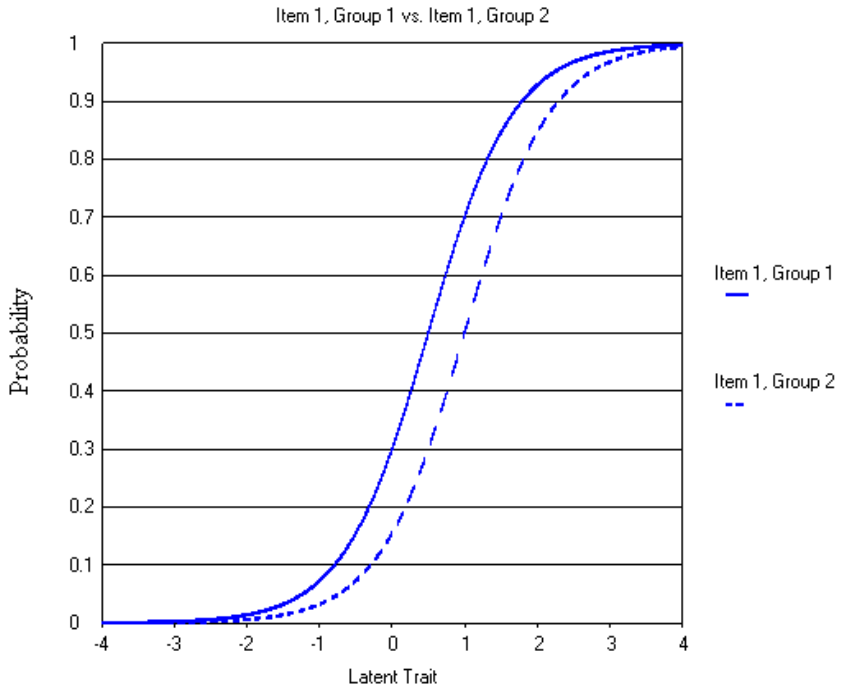


*Figure 3.* An item showing uniform DIF.

Conversely, nonuniform DIF indicates a relationship between the ability or agreement level of examinees or respondents and group membership. When an item possesses nonuniform DIF, the item favors one group at lower ability levels, but at higher ability levels, the item favors the other group. The equation used to test for nonuniform DIF is shown in Equation 9 and an illustration is shown Figure 4. The $\alpha$, $b_1$, and $b_2$ in Equation 9 represent the same intercept and slopes as in Equation 8, but $b_3$ in Equation 9 represents the slope for the interaction between the total score and group membership. The graph in Figure 4 shows the crossing curves of nonuniform DIF. The item in Figure 4 is easier for group 1 than group 2 at lower ability levels, but the item is harder for group 1 than group 2 at higher ability levels. As with uniform DIF, the space between the curves in Figure 4 indicates the presence of nonuniform DIF for that item

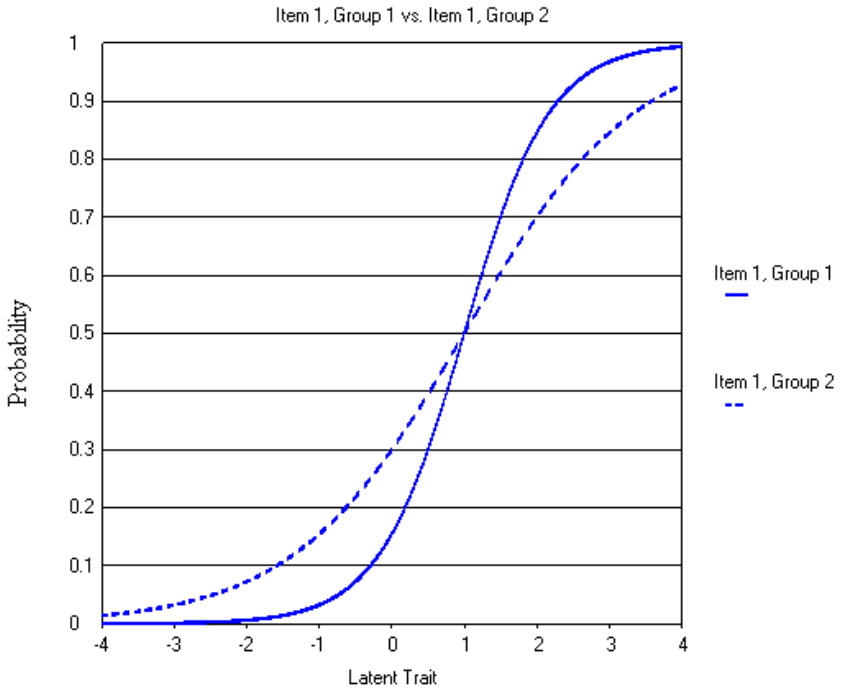


*Figure 4.* An item showing nonuniform DIF.

$$Y' = \alpha + b_1 \text{ totscore} + b_2 \text{ group} + b_3 \text{ (totscore*group)}. \quad (9)$$

For an item with nonuniform DIF, the DIF for one group, positive DIF, can compensate for some or all of the DIF against the other group, negative DIF. While this effect happens within an item, the same effect can also happen within a test. In the differential functioning of items and tests framework (DFIT, Raju, van der Linden, & Fleer, 1995), a parametric latent trait DIF detection method, positive DIF for one item may compensate for negative DIF in another item. A test, therefore, may contain several DIF items, but the adverse effect of DIF across the test may be minimal.

*R-Squared Effect Size Measure*

Zumbo (1999) provided an effect size measure for the logistic regression procedure that can be extended to OLR. Zumbo stated that an effect size measure for the

logistic regression procedure can be obtained if the R-squared values for nested models are compared. Three models are typically used in the logistic regression procedure. Model 1 contains only the total score. This model, as shown in Equation 10, tests for the presence of DIF, not the type of DIF

$$Y' = \alpha + b_1 \text{totscore}. \tag{10}$$

Model 2 (seen in Equation 8, which tests for uniform DIF) contains the total score and grouping variable. Model 3, the final model, contains the total score, grouping variable, and the interaction between the total score and the grouping variable (seen in Equation 9, which tests for nonuniform DIF).

To calculate the R-squared effect size value for an item, the difference between two nested models is found. The R-squared effect size measure for uniform DIF is obtained when the R-squared value for model 1 is subtracted from the R-squared value for model 2, whereas the R-squared effect size measure for nonuniform DIF is obtained when the R-squared value for model 2 is subtracted from the R-squared value for model 3. Equation 11 illustrates the R-squared measure for uniform DIF

$$R^2_{uniform} = R^2_2 - R^2_1 . \tag{11}$$

Equation 12 illustrates the R-squared measure for nonuniform DIF

$$R^2_{nonuniform} = R^2_3 - R^2_2 . \tag{12}$$

Kim, Cohen, Alagoz, and Kim (2007) and Jodoin and Gierl (2001) provided classification guidelines for the R-squared effect size measure. The guidelines provided in Kim et al. (2007) used Cohen's (1992) effect size guidelines. These guidelines are:

1. Small effect size: < 0.13,
2. Medium effect size: 0.13≥ and <0.26,

3. Large effect size: ≥0.26.

The R-squared classification guidelines provided by Jodoin and Gierl (2001) were developed using the effect size value ranges from SIBTEST's effect size guidelines to predict a corresponding R-squared effect size value range for the R-squared effect size guidelines. The guidelines from Jodoin and Gierl are:

1. Negligible effect size: < .035,
2. Moderate effect size: .035 ≤ and < .070,
3. Large effect size: ≥ .070.

## DIF Research for Polytomous Items

The performances of the Mantel and OLR have been investigated in both applied and simulation studies. Early simulation studies compared the Mantel with other polytomous DIF detection methods using data containing both dichotomous and polytomous items. Zwick et al. (1993) conducted a simulation study comparing the performance of the Mantel and the generalized Mantel-Haenszel (GMH). The GMH, like the Mantel, is an extension of the Mantel-Haenszel test, but the GMH does not assume ordered response categories. The GMH tests for DIF by comparing the number of examinee responses in each score category between the reference and focal groups across each score interval level.

There were 20 dichotomous items and 5 four-category polytomous items. The dichotomous items in the study were generated using the three-parameter logistic model (a dichotomous IRT model with an *a*-, *b*-, and *c*-parameter where the *a*-parameter represents the ability of the item to discriminate between low and high performing examinees and the *c*-parameter represents guessing). The polytomous items in the study

were generated using the PCM. The dichotomous items were used to match reference and focal group members, and the 25th item, a polytomous item, was used as the studied item. Several factors were varied in the study including the values of the studied item's parameters, the DIF magnitude, the pattern of DIF (i.e., the placement of DIF along an item's category intersection parameters), and the average ability of the focal group.

Three different sets of category intersection parameters were created for the studied item. The category intersection parameters varied within each studied item and across each studied item. As a result, the studied item was easy or hard depending on which set of parameters was investigated. Two DIF magnitudes were investigated, 0.1 and 0.25. The authors stated that these two magnitudes were typically found when investigating DIF for dichotomous items.

There were four patterns of DIF. Constant DIF was created when the DIF magnitude being investigated was added to each category intersection parameter of the studied item. Balanced DIF was created when the DIF magnitude was added to one category intersection parameter of the studied item, but that same DIF magnitude was subtracted from another category intersection parameter of the same studied item. Having the DIF magnitude added and subtracted to a category intersection parameter within the studied item, balances the DIF across the intersection parameters of the studied item. The next two patterns of DIF created a DIF condition where the DIF magnitude was added to the lowest category intersection parameter of the studied item (i.e., shift-low DIF), and a DIF condition where the DIF magnitude was added to the highest category intersection parameter of the studied item (i.e., shift-high DIF).

The average ability of the focal group was generated to have either a mean of 0 and a standard deviation of 1 or a mean of -1 and a standard deviation of 1. The mean and standard deviation of the reference group remained the same across all conditions with a mean of 0 and a standard deviation of 1. There were 1,000 examinees with 500 examinees in the reference group and focal group, respectively; and each condition was replicated 600 times.

The researchers averaged the results of the Type I error and power studies for the two ability conditions because the results for both sets of data were similar. The study found that the Mantel had higher power than the GMH when there was constant DIF with a magnitude of 0.25. Although the Mantel outperformed the GMH in these conditions, the GMH also had its highest power when there was constant DIF with a magnitude of .25. These results indicated that it was more difficult to detect DIF with a DIF magnitude of 0.1 for both DIF detection methods. The study also found that both DIF detection methods had Type I error rates near the nominal rate of 0.05 in conditions with no DIF.

In another simulation study, Chang, Mazzeo, and Roussos (1996) compared the Mantel, polytomous Simultaneous Item and Bias Test (Poly-SIBTEST) (i.e., a nonparametric latent trait method that uses a classical test theory latent trait variable, the true score, to match examinees or respondents), and the SMD procedure. These three DIF detection methods were compared using data generated under the three-parameter logistic model, the PCM, and the generalized partial credit model (GPCM). The GPCM has several *b*-parameters like the PCM, but this model also has an *a*-parameter to describe examinee performance. This one study was constructed to contain two smaller simulation studies. The first simulation study used the same research conditions as Zwick et al.

(1993) which had the studied item generated with the PCM (e.g., the item had no discrimination parameter) and a test length of 25 items. The second simulation study used the GPCM to generate the studied item (e.g., the item had a discrimination parameter, or *a*-parameter). The authors decided to conduct the second simulation study because they concluded that real testing situations are likely to have items with a discrimination parameter, so it was important to determine the effect of a discrimination parameter on DIF detection.

The conditions for the Mantel and SMD had a test length of 25 items, while the conditions for Poly-SIBTEST had a test length of 24 items. Though the number of test items differed across conditions, there was only one studied item investigated for each DIF detection method. For the second simulation study, three sample sizes were used, 1,000, 2,000, and 4,000, and each sample size condition had equal sample size ratios. As a result, the sample size ratios were 500:500, 1,000:1,000, and 2,000:2,000, respectively. For each condition, 1,000 replications were conducted.

In conditions with no DIF (baseline conditions), study 1 found that the Mantel and SMD had Type I error rates close to the nominal rate of 0.05, whereas Poly-SIBTEST had Type I error rates slightly above the nominal rate. This occurred in baseline conditions when the focal group had a mean of 0 and a mean of -1. On average, across all DIF conditions, all three DIF detection methods had similar power to detect constant DIF, but their performance varied across conditions with other DIF patterns. All three DIF methods had the highest power for the constant DIF condition with a magnitude of .25. The Type I error results for study 2 indicated that the Type I error rates for the Mantel and SMD increased as the discrimination parameter value became less similar to

the average discrimination parameter value of the entire test. The Type I error rates for these two methods also increased as the sample size increased. The Type I error rates for Poly-SIBTEST also increased as the discrimination value became less similar than the average discrimination value of the test, but the increase was not as extreme. The increase in sample size seemed to not have an effect on Type I error.

Su and Wang (2005) also conducted a simulation study on DIF detection using an assessment that contained dichotomous and polytomous items. The study compared the performance of the Mantel, GMH, and logistic discriminant function analysis (LDFA), a DIF detection method which uses a logistic function to predict group membership instead of the item score, using data modeled under the PCM and the graded response model (GRM). Several variables were varied in the study such as test length, the percentage of DIF items, DIF pattern, impact, DIF magnitude, and test purification (i.e., the extent to which DIF items are excluded from the matching variable). The results indicated that average signed area (ASA) was important in determining the Type I error of the three DIF methods. ASA is the average of the difference between the item difficulties of the reference group and focal group. An ASA of zero indicates an assessment advantages no group, whereas an ASA that is positive indicates an assessment advantages the reference group, but an ASA that is negative indicates an assessment advantages the focal group. The Type I error for the three DIF detection methods were close to the nominal rate of 0.05 when ASA was near zero, but when ASA increased, the Type I error of the three DIF methods also increased. Type I error was closer to the nominal level for the items generated with the PCM than for those items generated with the GRM. The study also found that the Mantel and LDFA had higher power to detect DIF than the GMH method

under all patterns of DIF except when the pattern of DIF was balanced. When the DIF was balanced, the GMH performed better than the Mantel and LDFA.

Recently, there has been a shift in the literature on DIF detection for polytomous items. Whereas previous DIF detection studies investigated DIF when the assessment contained both dichotomous and polytomous items, current DIF detection studies have investigated the performance of several DIF detection methods when the assessment contain only polytomous items (e.g., Kim, Cohen, Alagoz, & Kim, 2007; Kristjansson, Aylesworth, McDowell, & Zumbo, 2005; Meyer, Huynh, & Seaman, 2004; Wang & Su, 2004). These studies are important to polytomous DIF detection research because they help determine how DIF methods will function when the entire assessment contains polytomous items.

Kim, Cohen, Alagoz, and Kim (2007) conducted an applied study in which two types of IRT likelihood ratio tests (i.e., tests that use a likelihood ratio statistic to test for DIF), OLR, the Mantel, and the GMH were compared to determine their ability to detect DIF items. The data were taken from a state-wide test which contained 10 three-category polytomous items. The data set contained 105,731 examinees. Since this was an applied study, there were no conditions in which there was prior knowledge of the presence of DIF items. The results of the DIF analysis indicated that most of the inferential DIF detection methods flagged most of the test items as DIF items. This likely occurred due to the large student sample. Due to the extremely large sample size, descriptive DIF measures were later used to identify potentially biased items.

In another applied study, Meyer, Huynh, and Seaman (2004) compared two DIF detection methods typically used with small sample sizes, the Wilcoxon Rank Sum test

and the van der Waerden Normal Scores test, with the Mantel. The study investigated whether the small sample DIF methods detected DIF at higher rates than the Mantel which is ordinarily used with large sample sizes. The data for the study were obtained from an attitude measure designed to assess attitudes regarding math. The measure contained 30 four-category polytomous items, and there were a total of 375 respondents in the data set. The study found that the Mantel flagged more DIF items than either of the exact small sample methods. Ten items were flagged for DIF using the Mantel, but four items were flagged for DIF using the Wilcoxon Rank Sum test and the van der Waerden Normal Scores Test.

Wang and Su (2004), however, compared the Mantel to the GMH using data modeled with the PCM and the GRM. The study manipulated the same variables as Su and Wang (2005). The study found that both DIF detection methods had better Type I error rates under the PCM than the GRM, but when ASA, or the space between two ICCs, increased beyond zero, the Type I error for both DIF methods increased. The Mantel was found to have higher power than the GMH under all DIF patterns except when the DIF was balanced.

Kristjansson et al. (2005) also studied the performance of the Mantel, and their study compared the Mantel, the GMH, LDFA, and OLR for polytomous data generated using the GPCM. The assessment investigated contained 26 four-category polytomous items, and the 26$^{th}$ item was the studied item. The type of DIF, the discrimination parameter value of the studied item, impact, and sample size ratio were varied in the study. To study the type of DIF, a DIF magnitude of .25 was added to each category intersection parameter in the studied item to create uniform DIF with a constant DIF

pattern. Nonuniform DIF was created by varying the *a*-parameter value of the reference group parameters for the studied item.

To study impact, two levels of focal group ability were investigated. Focal group members had the same average ability as the reference group members (mean of 0 and standard deviation of 1) in the first level of focal group ability. Focal group members had a lower average ability than the reference group members in the second level of focal group ability. The focal group members in the second level had a mean of -0.5 and a standard deviation of 1, whereas the reference group members had a mean of 0 and a standard deviation of 1.

The sample size for the study was set at 4,000 examinees, but two sample size ratios were investigated, 2,000:2,000 and 3,200:800, to study the impact of sample size ratio on DIF detection. These two sample size ratios resulted in a set of conditions where the reference group and the focal group had an equal number of examinees and another set of conditions where the reference group had more examinees than the focal group. There were 400 replications for each condition.

The study found that all four DIF detection methods had Type I error rates that adhered to the nominal level. The Type I error rates for conditions with impact were slightly higher than conditions with no impact. All four DIF detection tmethods had high power to detect uniform DIF, however, OLR and GMH had the highest power to detect nonuniform DIF. Power rates for all DIF methods were higher for the sample ratio of 1:1 (i.e., when the sample size ratio was 2,000:2,000).

While there are many studies investigating DIF detection methods for polytomous items, many of these studies have only researched inferential DIF detection methods. As

discussed earlier, Kim, Cohen, Alagoz, and Kim (2007) and Meyer, Huynh, and Seaman (2004) investigated inferential DIF detection methods for polytomous items, but these studies also investigated effect size measures associated with the polytomous DIF detection methods.

Kim, Cohen, Alagoz, and Kim (2007) compared two types of likelihood ratio tests, OLR, the Mantel, and the GMH. They found that with a very large sample size most of the items on the assessment were flagged as DIF items. As a result, descriptive DIF measures were used to identify potentially biased items. In the study, observed score effect size measures and model-based effect size measures were studied. The study found that of the ten items flagged for DIF, only one of the items would need to be investigated for item bias. Overall, when effect size measures were used in conjunction with inferential DIF detection methods, fewer items were flagged as DIF items.

Meyer, Huynh, and Seaman (2004) compared the Wilcoxon Rank Sum test, the van der Waerden Normal Scores test, and the Mantel, and found that the Mantel identified more DIF items than the other two DIF detection methods. Once the statistical significance results were found, the researchers obtained an effect size measure for each assessment item to determine if the same items would be flagged for DIF using a descriptive DIF measure. In the study, SMDES was used as the effect size measure. ETS classification guidelines were used to sort the items into small, medium, and large DIF categories. The results indicated that when the SMDES was paired with the previous statistical significance results, all three DIF methods flagged the same DIF items when the item was placed in the large DIF category. This indicates there was less agreement

among the DIF detection methods when the DIF item was placed in the small or medium DIF category.

## Missing Data

While DIF is one factor that may adversely affect test validity, another factor that may adversely affect the validity of a test is missing data. Missing data, the omission of item values, causes a reduction of available information to accurately analyze the data which may lead to faulty inferences regarding examinee or respondent performance. Little and Rubin (1987) described three mechanisms of missing data: data missing completely at random, data missing at random, and data missing not at random. Data missing completely at random (MCAR) occurs from a factor not related to the variable being assessed or any other variable related to the assessment (Peugh & Enders, 2004). Individuals inadvertently not responding to an item or individuals not needing to respond to an item based on a previous response will produce data that are MCAR. Another example of data that are MCAR occurs during matrix sampling where entire sections of items are intentionally not presented to examinees during an assessment (Peugh & Enders, 2004).

Data missing at random (MAR) are due to a variable related to an assessment, but that is not related to the value of the variable with the missingness. If a person with less education is less likely to answer an item regarding income, the data would be MAR because a factor related to the individual, level of education, produced the missingness (Sinharay, Stern, & Russell, 2001).

Alternatively, data missing not at random (MNAR) occurs because of the value of the item being assessed. If respondents with high levels of income are more likely not to

answer an item that asks about their level of income, then the missing data would be MNAR (Sinharay, et al., 2001). Sinharay et al. (2001) stated that during data analysis, there is no method to accurately decipher whether data are MAR or MNAR. Graham and Donaldson (1993) stated that a sufficient strategy for determining the missing data mechanism is to simply ask nonrespndents the reason why they did not respond to an item.

There are several imputation techniques available to lessen the effects of missing data. Peugh and Enders (2004) described these missing data methods as *traditional* or *modern*. Traditional imputation methods include mean substitution and hot-decking imputation, while a modern method of handling missing data is multiple imputation. These three missing data methods are described below.

*Mean Substitution*

The mean substitution method replaces a missing value with the average of complete responses for an item. Huisman (2000) described two types of mean substitution: item mean substitution and within-person mean substitution (WMS). Item mean substitution imputes a missing value with the average of responses across all individuals on an item. In contrast, WMS imputes a missing value with the average of responses taken from the person with the missing value. Mean substitution is relatively easy to implement, but replacing missing values with a mean attenuates a variable's variance and covariances with other variables in the assessment (Peugh & Enders, 2004; Schafer & Graham, 2002; Sinharay et al., 2001). Although an attenuation of variances and covariances have been noted, Furlow, Fouladi, Gagné, and Whittaker (2007) stated that there is less attenuation with WMS compared with item mean substitution because

WMS incorporates an individual's response patterns into the imputed value. WMS can be easily implemented in SAS (SAS Institute, 2005) with minimal programming knowledge, thus, it is a highly accessible missing data method that both experienced and novice analysts can use.

*Hot-Deck Imputation*

The hot-deck imputation (HD) method replaces an individual's missing value with a value from another individual with similar response patterns. The donor value is usually a value randomly chosen from a set of individuals with a complete response on the missing item in question (Peugh & Enders, 2004). Individuals with complete and missing data are typically matched on specific criteria. These criteria can vary across data analyses. Schafer and Graham (2002) stated that the method of randomly choosing a response from a set of complete responses in HD does not have as much effect on variability as mean substitution, but this method can still alter other measures of association for a variable. Sinharay et al. (2001) also stated that when a dataset has a high percentage of missing data, HD may be difficult to implement because of the scarcity of complete donor cases.

Currently, HD has had limited use due to the software required or the vast programming knowledge needed to implement the method. The U.S. Census Bureau, however, commonly uses this method to impute missing data (Furlow et al., 2007). SAS can be used to implement HD, but the analyst would need to write a HD program in order to perform the comparison between individuals with missing values and those individuals with no missing values. This comparison often takes considerable time to conduct

especially when a large data set is used, and there is a large amount of missing data. SOLAS 3.0 (Statistical Solutions, 2006), however, has HD as an available feature.

*Multiple Imputation*

Multiple imputation (MI) is currently one of the most recommended missing data methods in the literature (Peugh & Enders, 2004). Unlike WMS and HD which impute a single value, MI imputes multiple values for a missing response. MI takes the uncertainty of replacing missing values into account when imputing missing values (Schafer & Graham, 2002; Sinharay et al., 2001). Collins, Schafer, and Kam (2001) stated that "the missing values are replaced by $m$>1 sets of simulated imputed values, resulting in $m$ plausible but different versions of the complete data" (p.335). The creation of each $m$ data set is the first step in the MI process.

Each $m$ data set is created by a two-step process. Allison (2002), Peugh and Enders (2004), and McKnight, McKnight, Sidani, and Figueredo (2007) described this process. The first step, called the I-phase, is the step where the missing data are imputed. In the first I-phase, an EM algorithm is used to estimate an initial mean vector and covariance matrix that is used to form a series of regression equations. The regression equations are used to calculate a predicted value for each missing value in the original data set. Once each predicted value is calculated, a residual value is added to the predicted value to create a final value that will be imputed into the original data set. This initial complete data set is then used to create the next complete data set.

The second step in the imputation process is called the P-phase. The P-phase creates a new mean vector and covariance matrix modeled from the previous complete data set in the I-phase. This new mean vector and covariance matrix is then used in the

next I-phase to produce new regression equations to predict new values for the next data set. Again, a residual value is added to the predicted values, and those final values are then imputed to make the second complete data set. The I- and P-phases are repeated many times to obtain the required number of *m* data sets.

Several repetitions, or iterations, of these two phases are required to obtain the appropriate number of *m* data sets because each new imputed data set is related to the previous data set. This correlation between consecutive data sets occurs because the P-phase uses the complete data set created in the I-phase to create new mean vectors and covariance matrices for the next I-phase. Since each new data set is related to the previous data set, consecutive data sets are not retained to make an *m* data set. Data sets are retained after several iterations of the I- and P-phases to obtain the required *m* data sets.

After the required *m* data sets are obtained, each data set is analyzed separately using normal statistical procedures (e.g., conducting a DIF analysis), the second step in the MI process. The last step in the MI process requires combining the statistical results from each analyzed *m* data set. This last step produces one final result. Rubin (1987) developed rules for combining the *m* statistical results. The equation is

$$\bar{q} = \frac{1}{m}\sum_{i=1}^{m}\hat{q}_i \,, \tag{13}$$

where $\hat{q}_i$ is the result calculated for each *m* dataset that is averaged across all *m* data sets. The total variance of $\bar{q}$ is found using equation 14

$$t = \bar{u} + \frac{1}{(1+m)}b \,, \tag{14}$$

where $\bar{u}$ is the variance within each imputed data set and $b$ is the variability between each imputed data set. The equation for $\bar{u}$ is

$$\bar{u} = \frac{1}{m}\sum_{i=1}^{m} u_i \ , \tag{15}$$

and the equation for $b$ is

$$b = \frac{1}{(m-1)}\sum_{i=1}^{m}[\hat{q}_i - \bar{q}]^2 \ . \tag{16}$$

When using MI, typically "three to five imputations are sufficient to obtain excellent results" (Sinharay et al., 2001, p. 320). MI can be implemented with several accessible software programs, such as SAS and SPSS (SPSS, 2006), which may account for its increased use in handling missing data (Furlow et al., 2007).

There have been several simulation studies which have investigated the performance of various missing data methods using a polytomous Rasch model (e.g., DeAyala, 2003; Furlow, Fouladi, &Whittaker, 2003). The purpose of both studies was to determine the most effective missing data method to use when estimating respondent attitude level. DeAyala (2003) investigated the ignoring the omitted response, selecting the "midpoint" response category, HD, and a Likelihood-based approach method, where various possible responses are imputed and the likelihood of the response pattern is calculated, when there were three levels of missing data, 7%, 13%, and 20%. DeAyala used real data (N=4,282) to generate a data set for the simulation study. Only individuals in the real data set with complete responses were used in the simulation data set, thus, the simulation data set contained 3,473 respondents. All item responses to the fifteen four-category polytomous items were generated under Andrich's RSM, and later calibrated to obtain the theta estimates. The study found that the accuracy of trait estimation was

related to the amount of omitted responses and the method used to handle the missing data. The study also found that HD had the greatest accuracy in trait estimation of all the imputation methods.

Furlow et al. (2003) investigated the effect that different missing data mechanisms have on theta estimation using polytomous items generated under the PCM and the RSM. Data were generated to contain MCAR and MAR data, and WMS and MI were used as the methods of handling the missing data. The results of this study indicated that MI provided the most accurate trait estimates.

A recent simulation study by Furlow et al. (2007) investigated the impact of DIF and missing data on respondent theta estimation when data were generated under the PCM and the RSM. The study investigated four missing data methods, complete case analysis (CCA), WMS, HD, and MI, when data were MCAR. CCA was the only missing data method implemented that did not impute missing values. CCA is unique, because this method allows for the deletion of individual cases when a value is missing.

Several factors were varied in the study to determine their effect on theta estimation when DIF and missing data were present. There were four levels of DIF magnitude: .25, .50, .75, and 1.00. There were two levels of the amount of test items containing DIF: 10% DIF and 20% DIF and three levels of missing data: 10%, 25%, and 40%. There were 20 four-category items generated for the PCM and the RSM, respectively. Each study condition was replicated 500 times with 1,000 simulees in each replication. Each sample size of 1,000 simulees had 900 reference group members to 100 focal group members.

The study found that the RSM estimated respondents' theta more adequately than the PCM when missing data were present, and MI had the best performance of the four methods for handling missing data. CCA was appropriate only when small amounts of data were missing. The amount of missing data influenced the strength of the correlation between the conditions with no missing data and those with missing data as well as the precision of theta estimates. DIF, however, also influenced theta estimates. DIF resulted in underestimated thetas (i.e., a decrease in the ability of the respondents) for respondents in the focal group and lower standard deviations for these respondents as well. This effect became more pronounced as the magnitude of DIF increased.

## Statement of the Problem

Although there have been simulation studies investigating DIF detection methods using polytomous items and separate studies investigating missing data and methods used to handle missing data, to date, there has been no research that demonstrates the effectiveness of DIF detection methods when missing data are present. This study will investigate the Type I error and power of several DIF detection methods using several methods of handling missing data for polytomous items generated under the PCM. The Type I error and power of the Mantel and OLR will be compared using within-person mean substitution and multiple imputation when data are MCAR.

In addition to assessing the Type I error and power of DIF detection methods under various methods of handling missing data, this study will also assess the impact of missing data on effect size measures associated with polytomous DIF detection methods. Effect size measures are typically used in conjunction with DIF detection methods, and to

date, there has been no research assessing how missing data may impact effect size measures for the Mantel and OLR.

# CHAPTER 3

# METHOD

The purpose of this Monte Carlo simulation study was to investigate the performance of the Mantel and OLR using WMS and MI when missing data were MCAR. In addition to investigating the performance of the Mantel and OLR, this simulation study also investigated effect size measures associated with the Mantel and OLR, the SMDES and the R-squared effect size measure. Several variables were manipulated in this study, including the DIF magnitude, the percent of missing data, the presence or absence of group ability differences (impact), the sample size, and the ratio of reference group sample size to focal group sample size. For each of the manipulated study conditions, the performances of the Mantel and OLR were compared for each of the two methods for handling missing data, and the effect size values for each percent of missing data were compared for the SMDES and the R-squared effect size measure.

## Study Design Conditions

In this study, there were conditions investigating Type I error rates and conditions investigating power. The conditions investigating Type I error had items with no DIF, whereas the conditions investigating power had items with DIF. The Type I error and power conditions had factors that were held constant and factors that were varied.

### *Factors Held Constant*

*Polytomous IRT model*. The PCM was used to generate the data for the reference group and the focal group. This model has been used in many simulation studies on DIF

detection (e.g., Chang, Mazzeo, & Roussos, 1996; Su & Wang, 2005; Zwick, Donoghue, & Grima, 1993; Zwick & Thayer, 1996). This model is versatile because it can be used for both cognitive assessments and attitude assessments, and it possesses the Rasch characteristic of the total score being a sufficient statistic to determine ability. Using the PCM in this study may help determine the impact of its Rasch characteristics on DIF detection in the presence of missing data.

*Test Length*. There were 20 items generated under the PCM. This is a common test length in simulation studies investigating DIF. This test length is also similar to DIF detection studies investigating the impact of DIF for attitude items (Dodeen, 2004; Dodeen & Johanson, 2003). Attitude assessments typically do not contain as many items as cognitive assessments.

*Percent of Items with DIF*. Ten percent of the 20 items in this simulation study contained DIF; therefore, two items from the assessment were DIF items. Several simulation studies have investigated DIF detection methods when there were more than one DIF item (e.g., Bolt, 2002; Chang, Mazzeo, & Roussos, 1996; Su & Wang, 2005; Wang & Su, 2004; Zwick, Thayer, & Mazzeo, 1997). Wang and Su (2004) stated that it is more realistic to have a simulation study that contains more than one DIF item.

*Type of DIF*. Uniform DIF was the only type of DIF investigated in this study since the PCM has only *b*-parameters. To investigate nonuniform DIF, an item must contain an *a*-parameter, and examinees or respondents in the reference group must have a different *a*-parameter than the examinees or respondents in the focal group.

*Factors Varied*

*Percentage of Missing Data.* There were three levels of missing data in this study. These levels included 10%, 25%, and 40% of the data that were MCAR. Furlow, et al. (2007) used these three levels of missingness in their study investigating missing data and the presence of DIF when estimating theta levels. These three percentages of missing data were compared to determine how well each DIF detection method performed when there were different amounts of missing data.

*DIF Magnitude.* Three levels of DIF magnitude were investigated, .25, .50, and .75. These values represented the amount of DIF that occurred within the focal group. The magnitude of .25 has been frequently used in several DIF studies (e.g., Kristjansson et al, 2005; Wang & Su, 2004; Zwick, Donoghue, & Grima, 1993; Zwick, Thayer, & Mazzeo, 1997), however, it is important to study how larger magnitudes of DIF may affect DIF detection methods.

*Impact*. There were two levels of impact investigated in this study. Kristjansson et al. (2005) stated that impact may affect the Type I error rates of DIF detection methods. When the reference and focal groups had the same ability, the data were generated to have a mean of 0 and a standard deviation of 1. When ability distributions between the focal group and the reference group differ, the data for the focal group were generated to have a mean of -0.5 and a standard deviation of 1, whereas the data for the reference group were generated to have a mean of 0 and a standard deviation of 1.

*Sample Size and Sample Size Ratio*. Sample size is a factor that was considered when designing this simulation study. Su and Wang (2005) and Wang and Su (2004) used a sample size of 1,000 examinees in their simulation studies with a sample size ratio

of 500:500 for the reference and focal groups. In their study designs, the sample size ratios were equal. This same sample size of 1,000 simulees with a sample size ratio of 500:500 was used in this study. Kristjansson et al. (2005), however, investigated the effect of unequal sample size ratios on DIF detection, and their results indicated that sample size ratios had an effect of the number of items flagged as DIF items. Conditions with equal sample size ratios had higher power than conditions with unequal sample size ratios. Due to this finding, this study also investigated the effect of unequal sample size ratios on DIF detection by investigating the sample size ratios of 700:300 and 900:100 in addition to the sample size ratio of 500:500.

Kristjansson et al.'s simulation study investigated sample size ratios for only one sample size. Raiford-Ross (2007), however, studied the effect of sample size ratios on DIF detection using two different sample sizes. Raiford-Ross's study investigated two sample sizes, 2,000 and 5,000, where each sample size was divided into conditions with an equal and an unequal sample size ratio.

Within a sample size, Raiford-Ross found that conditions with equal sample size ratios had higher power to detect DIF than conditions with unequal sample size ratios. In other words, when the sample size of 2,000 was investigated, conditions with a sample size ratio of 1,000:1,000 performed better than conditions with a sample size ratio of 1,800:200, and when the sample size of 5,000 was investigated, conditions with 2,500:2,500 performed better than conditions with 4,500:500. Across sample sizes, Raiford-Ross found that conditions with a sample size of 2,000 and an equal sample size ratio of 1,000:1,000 had higher power than conditions with a sample size of 5,000 and an unequal sample size ratio of 4,500:500.

The conditions with equal sample size ratios had a harmonic mean that was equal to the sample sizes of the reference and focal groups. For example, the sample size condition with 2,000 simulees and an equal sample size ratio of 1,000:1,000 had a harmonic mean of 1,000 which was equal to the sample size of the reference group the sample size of the focal group. The equation for the harmonic mean is

$$\text{Harmonic mean} = \frac{2n_1 n_2}{n_1 + n_2}, \tag{17}$$

where $n_1$ is the sample size ratio for the reference group and $n_2$ is the sample size ratio for the focal group. The conditions with unequal sample size ratios had a harmonic mean that was less than the sample size of the reference group but more than the sample size of the focal group. For example, the sample size condition with 5,000 simulees and an unequal sample size ratio of 4,500:500 had a harmonic mean of 900.

This study investigated two additional sample sizes, 1,200 and 1,500, that had the same harmonic mean as the 500:500 sample size ratio condition, but unequal sample size ratios. Conditions with a sample size of 1,200 had a sample size ratio of 845:355, whereas the conditions with a sample size of 1,500 had a sample size ratio of 1183:317. Overall, this study had five sample size ratios.

*Overview of Final Study Design*

For the power study, there were three factors that were fully crossed 3 (magnitudes of DIF) x 3 (percents of missing data) x 5 (sample size ratios) = 45 fully crossed conditions. Each of these 45 conditions was compared for each of the two DIF detection methods across the two methods for handling missing data.

A set of baseline conditions were also examined in this study in order to examine DIF detection with no DIF, then with no missing data, and then when both occur

simultaneously. For the Type I error portion of the study where no DIF is present but data were missing, three factors were fully crossed: 3 (percents of missing data) x 2 (presence or absence of impact) x 5 (sample size ratios) = 30 fully crossed conditions. In each of these 30 conditions, the results of the two DIF detection methods were compared across the two methods for handling missing data.

For the second part of the Type I error portion of the study, conditions were also examined where there were no DIF and no missing data: 2 (presence or absence of impact) x 5 (sample size ratios) = 10 fully crossed conditions. The two methods of DIF detection were compared for each of these 10 conditions. In addition to the Type I error conditions that contained items with no DIF, there will also be baseline conditions that contain no missing data (but DIF is present) in order to observe how each of the two methods perform for DIF detection when there are no missing data. Three factors were fully crossed in these conditions: 3 (magnitudes of DIF) x 5 (sample size ratios) = 15 fully crossed conditions. In each of these 15 conditions the two methods of DIF detection were compared. These baseline conditions resulted in an additional 55 conditions.

## Data Generation

The item parameters for the PCM used to generate the data were taken from Table 2 in Masters (1982). The table contains 14 items, but to lengthen the assessment, 6 of these 14 items were repeated to obtain a 20 item assessment. The items in the table were from the Fine- Motor/Cognitive section of the Developmental Indicators for the Assessment of Learning test (Mardell & Goldenberg, 1972, 1975). Each item has a 4-point scale resulting in three category intersection parameters per item.

In order to generate DIF items, specific item parameters for the focal group were altered according to the condition under examination (while the reference group parameters remained unaltered). The category intersection parameters for two items in the focal group had DIF added to them, and that amount varied depending on the study condition. The two DIF items in this study were item 4 and item 17. The *b*-parameters for item 4 are 1.81, 1.07, and 1.46, and the *b*-parameters for item 17 are 0.32, 0.86, and 2.21 (see Appendix A for all item parameters).

Once DIF was added to the items for the focal group, the data were generated with the PCM. In this study, a total of 1,000 replications were completed for each condition. Responses for the reference group and focal group members were generated separately and were later combined to create one data set containing both reference group and focal group responses. The sample size for the reference group and focal group were taken from the sample size ratio being studied. For example, for the 900:100 conditions, there were 900 reference group members and 100 focal group members.

IRTGEN (Whittaker, Fitzpatrick, Dodd, & Williams, 2003), a SAS/IML program, used the item parameters for the PCM to generate a sample of item responses and thetas at the subject level, assuming a normal distribution. In conditions with no impact, the thetas had a mean of 0 and a standard deviation of 1 for both groups. In conditions with impact, the reference group had thetas generated to have a mean of 0 and a standard deviation of 1, whereas the focal group thetas were generated to have a mean of -0.5 and a standard deviation of 1. After the item responses and thetas were generated for all replications, missingness was added to the data for the reference group and the focal group using SAS according to the appropriate study condition. All missing data were

generated to be MCAR, so the missingness was spread randomly across all items and all response categories.

After missingness was added to the data sets, the two missing data methods were used to replace missing item values. For WMS, when simulees had a missing item value, SAS was used to average the available item values for that simulee and impute the mean. After the missing item values had been replaced, a DIF analysis was run on these imputed data sets using the Mantel and OLR methods. For MI, the PROC MI statement in SAS was utilized.

In this study, five imputed data sets ($m$ = 5) were used for the MI procedure. PROC MI uses a Markov Chain Monte Carlo (MCMC) method to generate imputed item values. For MCMC, two steps were involved: the I-phase and the P-phase. Iterations of these two steps created the data sets needed to obtain the five multiply-imputed data sets. The five imputed data sets were taken from every 200$^{th}$ iteration (the default in SAS). These five data sets were then analyzed separately using the two DIF detection methods. Each DIF detection procedure yielded chi-square values that were then combined using a procedure described by Schafer (1997). An F-ratio statistic was used to determine the statistical significance of the combined chi-square values from the $m$ data sets. The chi-square value, $\chi^2$, and the degrees of freedom, $df$, from each analyzed $m$ data set was used to determine if the final chi-square value was statistically significant. The equation for the F-ratio statistic is

$$F(df_{\chi^2}, df_{Error}) = \frac{\left[\bar{\chi}^2 / \left(df_{\chi^2}\right)\right] - \left[r_2 (m+1)/(m-1)\right]}{1 + r_2}, \tag{18}$$

where $df_{\chi^2}$ is the degrees of freedom for $\chi^2$, $\bar{\chi}^2$ is the mean of the *m* imputations' $\chi^2$ values, $df_{Error}$ is the error degrees of freedom of the *F*-ratio statistic, and $r^2$ is the sample variance of the square root of $\chi^2$ across all imputations . The equation for $df_{error}$ is

$$df_{Error} = \frac{(m-1)\left(1+r_2^{-1}\right)^2}{\left(df_{\chi^2}\right)^{3/m}}, \tag{19}$$

and the equation for $r^2$ is

$$r_2 = \frac{(1+1/m)}{(m-1)}\left[\sum_{i=1}^{m}\chi_i^2 - \frac{\left(\sum_{i}^{m}\sqrt{\chi_i^2}\right)^2}{m}\right]. \tag{20}$$

## Data Analysis

SAS was used to conduct the Mantel and OLR DIF analyses. SAS was also used to calculate the effect size values for the SMDES and the R-squared effect size measure. The effect size guidelines for the SMDES and the R-squared effect size measure were not used to classify each effect size value because the size categories in which effect size values can be placed may vary. As a result, this study focused only on the effect size values obtained when data were MCAR.

After the DIF analyses were completed, the performance of each DIF detection method was compared across all conditions to determine the Type I error and power of each DIF method under the various study conditions. Type I error was the proportion of times out of 1,000 replications where DIF was falsely identified at the 0.05 level. Power was the proportion of times out of 1,000 replications where DIF was correctly identified

at the 0.05 level. For each DIF condition, the effect size values for every replication was retained and averaged at the end of the replication loop.

# CHAPTER 4

# RESULTS

The purpose of this Monte Carlo simulation study was to investigate two DIF detection methods, the Mantel and OLR, and two methods of handling missing data, WMS and MI, when DIF and missing data occur simultaneously. In addition to investigating the performance of the Mantel and OLR, this simulation study also investigated effect size measures associated with the Mantel and OLR, the SMDES and the R-squared effect size measure. The Type I error, power, and effect size results were reported for item 4 and item 17 because these two items were investigated for DIF in this study.

Tables 2-9 contain the Type I error rates for each DIF detection method when no data were missing and when data were missing and a specific missing data method was used to impute missing values. The Type I error results in each table are reported as a percent. Type I error rates were obtained from conditions in which there were no DIF items (i.e., baseline conditions). The DIF method, the method of handling missing data, the percent of missing data, the focal group ability distribution, the sample size, and the sample size ratio were manipulated in these conditions.

Tables 10-13 contain the power rates for each DIF detection method when no data were missing and when data were missing and a specific method of handling missing data was used to impute missing values. The power results in each table are reported as a percent. Power results were obtained from conditions that contained items with DIF. In

power conditions, the DIF method, the method of handling missing data, the percent of missing data, the DIF magnitude, the sample size, and the sample size ratio were manipulated for each method of DIF detection and method for handling missing data.

## Type I Error Results

### *Item 4*

#### *No Missing Data*

*No impact.* The Type I error results for item 4 when no data were missing is shown in Table 2. The Type I error rates for the Mantel and OLR were at or close to the nominal rate of 5%. Type I error rates for the Mantel ranged from 5.00 to 5.55, whereas the Type I error rates for OLR ranged from 5.60 to 5.83.

*Impact.* The Type I error results for item 4 when no data were missing, but impact was present is shown in Table 3. The Type I error rates for both DIF detection methods were slightly higher when no data were missing and impact was present. Type I error rates for the Mantel ranged from 6.00 to 6.80, whereas the Type I error rates for OLR ranged from 5.84 to 6.45.

Table 2

*Type I Error for Item 4 with No Missing Data and No Impact*

| | Method | |
|---|---|---|
| Sample size ratio | Mantel | OLR |
| 500:500 | 5.00 | 5.80 |
| 700:300 | 5.40 | 5.60 |
| 900:100 | 5.30 | 5.77 |
| 845:355 | 5.55 | 5.83 |
| 1183:317 | 5.46 | 5.78 |

Table 3

*Type I Error for Item 4 with No Missing Data and Impact*

| | Method | |
|---|---|---|
| Sample size ratio | Mantel | OLR |
| 500:500 | 6.80 | 6.10 |
| 700:300 | 6.50 | 6.45 |
| 900:100 | 6.00 | 6.20 |
| 845:355 | 6.20 | 6.00 |
| 1183:317 | 6.10 | 5.84 |

*Missing Data*

*No Impact.* The Type I error results for item 4 when data were missing is shown in Table 4. When WMS was used to impute missing values, the Type I error rates for the Mantel and OLR were close to the nominal rate of 5% although some conditions had Type I error rates slightly below and above the 5% level. The Type I error rates for the Mantel ranged from 4.00 to 6.20, whereas the Type I error rates for OLR ranged from 4.70 to 6.70. The Type I error rates for both DIF detection methods when MI was used to impute missing data were well below the nominal level. With MI, the Type I error rates for the Mantel ranged from 0.00 to 3.00, and the Type I error rates for OLR ranged from 2.00 to 3.10.

*Impact.* The Type I error results for item 4 when data were missing is shown in Table 5. When WMS was used to impute missing values, the Type I error rates for the Mantel ranged from 4.95 to 6.90, whereas Type I error rates for OLR under this same condition ranged from 5.10 to 9.30. MI had Type I error rates well below the nominal level. Type I error rates for the Mantel when MI was used ranged from 0.40 to 4.00, and Type I error rates for OLR when MI was used ranged from 0.23 to 3.10.

Table 4

*Type I Error for Item 4 with Missing Data and No Impact*

| | | Method | | | |
|---|---|---|---|---|---|
| | | Mantel | | OLR | |
| Sample size ratio | % Missing | WMS | MI | WMS | MI |
| 500:500 | 10 | 5.60 | 2.50 | 5.10 | 2.10 |
| | 25 | 5.00 | 1.20 | 5.90 | 1.00 |
| | 40 | 4.80 | 0.30 | 5.60 | 0.20 |
| 700:300 | 10 | 4.80 | 3.00 | 6.00 | 3.10 |
| | 25 | 5.10 | 1.70 | 5.50 | 1.60 |
| | 40 | 5.10 | 0.00 | 5.50 | 0.20 |
| 900:100 | 10 | 4.40 | 2.55 | 5.75 | 2.75 |
| | 25 | 6.20 | 1.40 | 6.70 | 1.55 |
| | 40 | 4.85 | 0.50 | 5.70 | 0.55 |
| 845:355 | 10 | 4.90 | 3.00 | 6.10 | 2.60 |
| | 25 | 4.60 | 1.60 | 5.90 | 1.50 |
| | 40 | 5.00 | 0.20 | 4.70 | 0.20 |
| 1183:317 | 10 | 4.00 | 2.10 | 4.90 | 2.80 |
| | 25 | 5.30 | 1.30 | 6.10 | 1.10 |
| | 40 | 5.50 | 0.20 | 5.50 | 0.30 |

Table 5

*Type I Error for Item 4 with Missing Data and Impact*

| | | Method | | | |
|---|---|---|---|---|---|
| | | Mantel | | OLR | |
| Sample size ratio | % Missing | WMS | MI | WMS | MI |
| 500:500 | 10 | 5.50 | 3.30 | 5.10 | 2.00 |
| | 25 | 5.40 | 1.30 | 7.00 | 1.20 |
| | 40 | 6.80 | 1.00 | 9.30 | 0.40 |
| 700:300 | 10 | 6.05 | 3.15 | 6.00 | 2.05 |
| | 25 | 4.95 | 1.25 | 6.90 | 1.25 |
| | 40 | 5.95 | 0.9 | 8.20 | 0.25 |
| 900:100 | 10 | 5.60 | 2.97 | 6.30 | 2.13 |
| | 25 | 5.27 | 1.27 | 6.50 | 1.20 |
| | 40 | 5.67 | 0.60 | 7.37 | 0.23 |
| 845:355 | 10 | 6.90 | 4.00 | 8.00 | 3.10 |
| | 25 | 5.70 | 1.20 | 8.10 | 0.40 |
| | 40 | 4.90 | 0.40 | 7.20 | 0.50 |
| 1183:317 | 10 | 5.00 | 2.70 | 5.30 | 2.50 |
| | 25 | 5.30 | 1.30 | 7.00 | 0.60 |
| | 40 | 5.80 | 0.60 | 8.40 | 0.30 |

*Item 17*

*No Missing Data*

*No Impact.* The Type I error results for item 17 when no data were missing is shown in Table 6. The Type I error rates for the Mantel and OLR were at or close to the nominal rate of 5% for these baseline conditions. Type I error rates for the Mantel ranged from 4.93 to 5.10, whereas the Type I error rates for OLR ranged from 4.80 to 5.22.

*Impact.* The Type I error results for item 17 when no data were missing and impact was present is shown in Table 7. The Type I error rates for the Mantel and OLR were at or close to the nominal rate of 5%. The Type I error rates for the Mantel ranged from 5.37 to 5.72, and the Type I error rates for OLR ranged from 4.60 to 5.28.

Table 6

*Type I Error for Item 17 with No Missing Data and No Impact*

| | Method | |
|---|---|---|
| Sample size ratio | Mantel | OLR |
| 500:500 | 5.10 | 5.10 |
| 700:300 | 4.85 | 4.80 |
| 900:100 | 4.93 | 5.00 |
| 845:355 | 4.93 | 5.15 |
| 1183:317 | 4.86 | 5.22 |

Table 7

*Type I Error for Item 17 with No Missing Data and Impact*

| | Method | |
|---|---|---|
| Sample size ratio | Mantel | OLR |
| 500:500 | 5.40 | 4.60 |
| 700:300 | 5.40 | 5.00 |
| 900:100 | 5.37 | 5.20 |
| 845:355 | 5.65 | 5.18 |
| 1183:317 | 5.72 | 5.28 |

*Missing Data*

*No Impact.* The Type I error results for item 17 when data were missing is shown in Table 8. When WMS was used to impute missing values, the Type I error rates for the Mantel and OLR were both well below and slightly above the nominal level. The Type I error rates for the Mantel when WMS was used to impute missing data ranged from 3.80 to 6.00. Type I error rates for OLR when WMS was used to impute missing values ranged from 3.40 to 5.80. The Type I error rates for both DIF detection methods when MI was used to impute missing data were well below the nominal level. The Type I error rates for the Mantel ranged from 0.10 to 3.50, and the Type I error rates for OLR ranged from 0.30 to 3.60.

*Impact.* The Type I error results for item 17 when data were missing is shown in Table 9. When WMS was used to impute the missing data, the Type I error for the Mantel ranged from 5.60 to 7.50, whereas the Type I error rates for OLR under this same condition ranged from 4.25 to 6.70. Conditions with MI had Type I error rates well below the nominal level. Type I error rates for the Mantel when MI was used to impute missing values ranged from 0.50 to 3.60 and the Type I error rates for OLR when MI was used to impute missing values ranged from 0.40 to 3.55.

Power

*Item 4*

*No Missing Data*

When no data were missing, the power for the Mantel and OLR increased as the DIF magnitude increased. The power results are displayed in Table 10. The power for the Mantel across all sample size ratios ranged from 53.4 to 68.1 when the DIF magnitude

Table 8

*Type I Error for Item 17 with Missing Data and No Impact*

| | | Method | | | |
|---|---|---|---|---|---|
| | | Mantel | | OLR | |
| Sample size ratio | % Missing | WMS | MI | WMS | MI |
| 500:500 | 10 | 4.80 | 3.00 | 5.50 | 3.60 |
| | 25 | 4.20 | 1.30 | 5.70 | 1.70 |
| | 40 | 4.70 | 0.50 | 5.00 | 0.60 |
| 700:300 | 10 | 6.00 | 3.50 | 5.40 | 3.40 |
| | 25 | 4.70 | 1.20 | 4.60 | 1.30 |
| | 40 | 4.70 | 0.10 | 5.40 | 0.30 |
| 900:100 | 10 | 5.65 | 3.30 | 5.80 | 3.55 |
| | 25 | 4.25 | 1.35 | 4.65 | 1.60 |
| | 40 | 4.75 | 0.40 | 5.10 | 0.40 |
| 845:355 | 10 | 5.60 | 2.60 | 5.30 | 2.60 |
| | 25 | 3.80 | 1.00 | 4.30 | 1.00 |
| | 40 | 4.30 | 0.70 | 4.00 | 1.20 |
| 1183:317 | 10 | 3.80 | 1.90 | 4.80 | 2.70 |
| | 25 | 3.90 | 0.60 | 3.40 | 0.40 |
| | 40 | 5.30 | 0.10 | 5.00 | 0.30 |

Table 9

*Type I Error for Item 17 with Missing Data and Impact*

| | | Method | | | |
|---|---|---|---|---|---|
| | | Mantel | | OLR | |
| Sample size ratio | % Missing | WMS | MI | WMS | MI |
| 500:500 | 10 | 7.10 | 3.60 | 5.40 | 3.20 |
| | 25 | 6.50 | 1.90 | 5.00 | 1.40 |
| | 40 | 6.40 | 0.50 | 4.50 | 0.60 |
| 700:300 | 10 | 6.35 | 3.60 | 5.60 | 3.55 |
| | 25 | 6.20 | 1.75 | 5.00 | 1.35 |
| | 40 | 6.70 | 0.70 | 4.25 | 0.75 |
| 900:100 | 10 | 6.13 | 3.27 | 5.50 | 3.27 |
| | 25 | 6.27 | 1.80 | 4.77 | 1.30 |
| | 40 | 6.4 | 0.73 | 4.27 | 0.67 |
| 845:355 | 10 | 7.40 | 3.60 | 6.70 | 3.30 |
| | 25 | 7.20 | 1.50 | 5.30 | 1.10 |
| | 40 | 7.50 | 0.70 | 5.60 | 0.40 |
| 1183:317 | 10 | 6.50 | 3.30 | 5.50 | 3.10 |
| | 25 | 5.60 | 1.20 | 5.10 | 0.70 |
| | 40 | 7.00 | 1.90 | 4.60 | 1.10 |

Table 10

*Power for Item 4 with No Missing Data*

| | Method | | | | | |
|---|---|---|---|---|---|---|
| | Mantel | | | OLR | | |
| | DIF magnitude | | | DIF magnitude | | |
| Sample size ratio | .25 | .50 | .75 | .25 | .50 | .75 |
| 500:500 | 68.1 | 99.9 | 100.0 | 67.1 | 99.6 | 100.0 |
| 700:300 | 65.0 | 99.6 | 100.0 | 64.3 | 99.4 | 100.0 |
| 900:100 | 53.4 | 92.5 | 99.3 | 53.2 | 92.0 | 99.2 |
| 845:355 | 57.4 | 94.3 | 99.5 | 57.3 | 94.0 | 99.4 |
| 1183:317 | 59.1 | 95.4 | 99.6 | 59.0 | 95.0 | 99.5 |

was .25, 92.5 to 100.0 when the DIF magnitude was .50, and 99.3 to 100.0 when the DIF magnitude was .75. The power for OLR across all sample size ratios ranged from 53.2 to 67.1 when the DIF magnitude was .25, 92.0 to 99.6 when the DIF magnitude was .50, and 99.2 to 100.0 when the DIF magnitude was .75. The conditions with a DIF magnitude of .50 and .75 had higher power than the conditions with a DIF magnitude of .25. Within each DIF magnitude condition, the 900:100 condition had the smallest power values, whereas the 500:500 condition had the largest power values. The sample size ratio conditions with the same harmonic mean did not have similar power. The condition with a sample size ratio of 500:500 had slightly higher power to detect DIF than conditions with a sample size ratio of 845:355 or 1183:317. This result was seen for both DIF detection methods.

*Missing Data*

When data were missing, power for the Mantel and OLR increased as the DIF magnitude increased, but within DIF magnitude conditions, the power decreased as the percent of missing data increased. The power results are displayed in Table 11. Power for the Mantel and OLR when WMS was used to impute missing values were comparable,

Table 11

*Power for Item 4 with Missing Data*

| | | Method | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mantel | | | | | | OLR | | | | | |
| | | WMS | | | MI | | | WMS | | | MI | | |
| Sample | % | DIF magnitude | | | DIF magnitude | | | DIF magnitude | | | DIF magnitude | | |
| size ratio | Missing | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 |
| 500:500 | 10 | 53.9 | 95.8 | 100.0 | 53.8 | 98.2 | 100.0 | 54.2 | 95.7 | 99.9 | 51.9 | 97.0 | 100.0 |
| | 25 | 37.7 | 85.3 | 98.4 | 30.1 | 89.3 | 99.8 | 37.8 | 82.1 | 96.9 | 25.5 | 84.1 | 99.6 |
| | 40 | 24.9 | 63.6 | 86.4 | 12.3 | 60.3 | 94.5 | 24.4 | 60.1 | 85.2 | 10.7 | 48.6 | 87.3 |
| 700:300 | 10 | 44.9 | 92.6 | 100.0 | 42.3 | 96.4 | 100.0 | 44.2 | 92.8 | 99.7 | 41.0 | 94.7 | 99.9 |
| | 25 | 30.5 | 76.8 | 95.4 | 22.2 | 79.2 | 98.9 | 31.6 | 75.1 | 93.9 | 19.0 | 74.7 | 97.2 |
| | 40 | 19.2 | 56.3 | 80.4 | 8.2 | 45.7 | 83.6 | 19.1 | 55.6 | 77.5 | 7.6 | 38.1 | 74.1 |
| 900:100 | 10 | 33.4 | 61.0 | 92.7 | 30.1 | 59.8 | 95.1 | 33.2 | 61.3 | 92.0 | 29.4 | 58.0 | 93.9 |
| | 25 | 23.1 | 41.0 | 81.2 | 14.7 | 33.9 | 83.7 | 24.5 | 41.2 | 80.7 | 13.3 | 32.5 | 80.1 |
| | 40 | 15.8 | 27.2 | 63.0 | 5.6 | 12.1 | 58.8 | 16.2 | 26.6 | 61.1 | 5.25 | 11.0 | 51.6 |
| 845:355 | 10 | 55.0 | 96.0 | 100.0 | 52.4 | 97.8 | 100.0 | 56.6 | 95.0 | 100.0 | 51.8 | 96.2 | 100 |
| | 25 | 34.9 | 83.4 | 98.3 | 26.9 | 89.0 | 99.7 | 34.1 | 81.7 | 96.9 | 24.2 | 83.3 | 98.9 |
| | 40 | 25.3 | 63.1 | 88.2 | 12.0 | 55.9 | 91.8 | 25.6 | 58.9 | 86.2 | 8.0 | 49.7 | 84.5 |
| 1183:317 | 10 | 53.3 | 96.2 | 100.0 | 50.0 | 97.8 | 100.0 | 52.8 | 94.7 | 100.0 | 49.7 | 97.0 | 100.0 |
| | 25 | 33.1 | 91.7 | 98.1 | 24.6 | 89.0 | 99.8 | 34.2 | 90.8 | 97.1 | 23.6 | 83.3 | 99.3 |
| | 40 | 24.7 | 63.4 | 87.3 | 8.8 | 55.4 | 90.8 | 24.2 | 59.8 | 85.5 | 6.9 | 49.0 | 84.9 |

although the power rates for the Mantel when MI was used were slightly higher than the power rates for OLR. Sample size ratio conditions with the same harmonic mean had similar power rates as the percent of missing data increased and as the DIF magnitude increased. These three sample size ratio conditions had higher power rates than the 700:300 and 900:100 conditions. Within each DIF magnitude condition, the 900:100 conditions had power rates that were lower than all the other sample size ratio conditions.

*Item 17*

*No Missing Data*

The power for the Mantel and OLR when no data were missing increased as the DIF magnitude increased. The results for both DIF detection methods were comparable. The power results are displayed in Table 12. The power for the Mantel when no data were missing ranged from 61.7 to 77.0 when the DIF magnitude was .25, 95.6 to 100 when the DIF magnitude was .50, and 99.9 to 100 when the DIF magnitude was .75. The power for OLR when no data were missing ranged from 61.8 to 77.6 when the DIF magnitude was .25, 95.9 to 99.9 when the DIF magnitude was .50, and 99.9 to 100 when the DIF magnitude was .75. Within each DIF magnitude condition, the 900:100 condition had the smallest power values, whereas the 500:500 condition had the largest power values. Sample size conditions with the same harmonic mean did not have similar power rates. The condition with a sample size ratio of 500:500 had slightly higher power to detect DIF than conditions with a sample size ratio of 845:355 or 1183:317. This result was seen for both DIF detection methods.

Table 12

*Power for Item 17 with No Missing Data*

| | Method | | | | | |
|---|---|---|---|---|---|---|
| | Mantel | | | OLR | | |
| | DIF magnitude | | | DIF magnitude | | |
| Sample size ratio | .25 | .50 | .75 | .25 | .50 | .75 |
| 500:500 | 77.0 | 100.0 | 100.0 | 77.6 | 99.9 | 100.0 |
| 700:300 | 74.9 | 100.0 | 100.0 | 74.9 | 99.9 | 100.0 |
| 900:100 | 61.7 | 95.6 | 99.9 | 61.8 | 95.9 | 99.9 |
| 845:355 | 65.5 | 96.7 | 99.9 | 65.8 | 96.9 | 100.0 |
| 1183:317 | 67.8 | 97.3 | 99.9 | 67.7 | 97.5 | 100.0 |

*Missing Data*

The power results for item 17 when data were missing are displayed in Table 13. Power for all sample size ratio conditions increased as the DIF magnitude increased, but within DIF magnitude conditions, the power decreased as the percent of missing data increased. Power for the Mantel and OLR were comparable for WMS, but when MI was used to impute missing values, the Mantel had slightly higher power than OLR. Conditions with the same harmonic mean had similar power rates as the percent of missing data and DIF magnitude increased. Within each DIF magnitude condition, these three sample size conditions had higher power rates than the 700:300 and 900:100 conditions. The conditions with a sample size ratio of 900:100 had the smallest power rates of all the other sample size ratio conditions.

Table 13

*Power for Item 17 with Missing Data*

| | | Method | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mantel | | | | | | OLR | | | | | |
| | | WMS | | | MI | | | WMS | | | MI | | |
| Sample | % | DIF magnitude | | | DIF magnitude | | | DIF magnitude | | | DIF magnitude | | |
| size ratio | Missing | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 |
| 500:500 | 10 | 68.1 | 99.6 | 100.0 | 61.2 | 99.7 | 100.0 | 67.6 | 99.4 | 100.0 | 62.6 | 99.6 | 100.0 |
| | 25 | 52.2 | 97.2 | 100.0 | 38.1 | 95.4 | 100.0 | 51.4 | 97.1 | 100.0 | 40.2 | 95.2 | 100.0 |
| | 40 | 39.9 | 88.7 | 98.7 | 17.0 | 76.1 | 98.6 | 37.7 | 87 | 98.8 | 17.1 | 75.9 | 98.4 |
| 700:300 | 10 | 61.4 | 99.1 | 100.0 | 53.8 | 98.7 | 100.0 | 61.2 | 98.8 | 100.0 | 54.3 | 98.7 | 100.0 |
| | 25 | 42.9 | 94.0 | 100.0 | 30.4 | 89.8 | 99.8 | 44.1 | 92.5 | 100.0 | 31.0 | 89.8 | 99.9 |
| | 40 | 35.2 | 81.4 | 98 | 12.7 | 63.4 | 94.6 | 34.6 | 77.9 | 97.5 | 14.0 | 63.6 | 93.9 |
| 900:100.0 | 10 | 43.8 | 78.6 | 98.9 | 37.7 | 77.1 | 99.0 | 44.5 | 79.4 | 99.0 | 38.2 | 77.1 | 99.0 |
| | 25 | 32.8 | 60.3 | 95.4 | 33.8 | 58.5 | 94.9 | 23.1 | 47.6 | 93.1 | 21.3 | 48.0 | 92.8 |
| | 40 | 25.6 | 47.6 | 87.8 | 8.7 | 18.2 | 73.2 | 24.6 | 44.7 | 85.8 | 9.1 | 21.6 | 74.6 |
| 845:355 | 10 | 66.7 | 99.2 | 100.0 | 62.0 | 99.6 | 100.0 | 67.1 | 99.1 | 100.0 | 63.2 | 99.5 | 100.0 |
| | 25 | 53.7 | 97.7 | 100.0 | 36.9 | 95.5 | 100.0 | 52.8 | 96.6 | 100.0 | 38.3 | 95.9 | 100.0 |
| | 40 | 37.5 | 88.3 | 99.0 | 15.1 | 72.1 | 97.7 | 37.5 | 86.9 | 98.8 | 16.5 | 75.2 | 97.7 |
| 1183:317 | 10 | 67.9 | 99.5 | 100.0 | 63.9 | 99.8 | 100.0 | 67.8 | 99.5 | 100.0 | 64.2 | 99.7 | 100.0 |
| | 25 | 53.3 | 98.9 | 100.0 | 36.1 | 95.5 | 100.0 | 52.8 | 96.6 | 100.0 | 37.5 | 95.9 | 100.0 |
| | 40 | 38.9 | 87.7 | 99.3 | 15.7 | 72.4 | 97.8 | 36.8 | 85.9 | 98.4 | 15.9 | 74.4 | 97.6 |

## Effect Size

### *SMDES for Item 4 and Item 17*

The average SMDES values for item 4 and item 17 are located in Tables 14-17. These effect size values were reported for the power conditions only. In the power conditions, the method of handling missing data, the percent of missing data, the DIF magnitude, the sample size, and the sample size ratio were manipulated.

*No Missing Data*

The average SMDES results for item 4 and item 17 when no data were missing are displayed in Tables 14 and 15. When no data were missing, the difference between the mean item score of the reference group and the mean item score of the focal group increased as the DIF magnitude increased. The SMDES values are the difference between the mean item score of the reference group and the mean item score of the focal group in standard deviation units. In all conditions, the mean item score of the focal group was a specific standard deviation value lower than the mean item score of the reference group. The difference between the two groups was larger for item 17 than for item 4 within each DIF magnitude condition.

The 500:500 sample size ratio conditions for item 4 had the largest difference between the reference and focal groups, but the sample size ratio conditions with the smallest difference varied as the DIF magnitude increased. When the DIF magnitude was .25, the 1183:317 condition had the smallest difference, but when the DIF magnitude was .75, the 845:355 condition had the smallest difference. For item 17, the difference in mean item score varied as the DIF magnitude increased. When the DIF magnitude was .25, the 700:300 condition had the largest difference, but when the DIF magnitude was

Table 14

*Average SMDES for Item 4 with No Missing Data*

| | Method: Mantel DIF magnitude | | |
|---|---|---|---|
| Sample size ratio | .25 | .50 | .75 |
| 500:500 | -0.1195 | -0.2268 | -0.3253 |
| 700:300 | -0.1186 | -0.2186 | -0.3047 |
| 900:100 | -0.1192 | -0.2122 | -0.2851 |
| 845:355 | -0.1184 | -0.2146 | -0.2942 |
| 1183:317 | -0.1170 | -0.2146 | -0.2947 |

Table 15

*Average SMDES for Item 17 with No Missing Data*

| | Method: Mantel DIF magnitude | | |
|---|---|---|---|
| Sample size ratio | .25 | .50 | .75 |
| 500:500 | -0.1352 | -0.2626 | -0.3896 |
| 700:300 | -0.1359 | -0.2594 | -0.3730 |
| 900:100 | -0.1312 | -0.2510 | -0.3576 |
| 845:355 | -0.1326 | -0.2543 | -0.3656 |
| 1183:317 | -0.1323 | -0.2548 | -0.3650 |

.50 and .75, the 500:500 conditions had the largest difference. The 900:100 conditions for item 17 consistently had the smallest difference.

*Missing Data*

The average SMDES results for item 4 and item 17 when data were missing are displayed in Tables 16 and 17. When data were missing, the difference between the mean item score of the reference group and the mean item score of the focal group increased as the DIF magnitude increased. There were larger differences between the reference and focal groups for item 17 than for item 4. Although the size of the difference between the reference group and the focal group increased as the DIF magnitude increased, the

Table 16

*Average SMDES for Item 4 with Missing Data*

| | | Method: Mantel | | | | | |
|---|---|---|---|---|---|---|---|
| | | WMS | | | MI | | |
| Sample | % | DIF magnitude | | | DIF magnitude | | |
| size ratio | Missing | .25 | .50 | .75 | .25 | .50 | .75 |
| 500:500 | 10 | -0.1043 | -0.1934 | -0.2700 | -0.1108 | -0.2088 | -0.2988 |
| | 25 | -0.0823 | -0.1537 | -0.2107 | -0.0928 | -0.1817 | -0.2590 |
| | 40 | -0.0644 | -0.1139 | -0.1556 | -0.0771 | -0.1461 | -0.2137 |
| 700:300 | 10 | -0.1010 | -0.1878 | -0.2598 | -0.1061 | -0.2012 | -0.2787 |
| | 25 | -0.0804 | -0.1483 | -0.2011 | -0.0912 | -0.1716 | -0.2379 |
| | 40 | -0.0646 | -0.1150 | -0.1519 | -0.0755 | -0.1392 | -0.1940 |
| 900:100 | 10 | -0.1019 | -0.1820 | -0.2169 | -0.1065 | -0.1912 | -0.2610 |
| | 25 | -0.0802 | -0.1456 | -0.2024 | -0.0901 | -0.1650 | -0.2260 |
| | 40 | -0.0630 | -0.1125 | -0.1500 | -0.0729 | -0.1336 | -0.1831 |
| 845:355 | 10 | -0.1043 | -0.1876 | -0.2564 | -0.1085 | -0.1996 | -0.2739 |
| | 25 | -0.0790 | -0.1473 | -0.1992 | -0.0909 | -0.1714 | -0.2348 |
| | 40 | -0.0629 | -0.1125 | -0.1535 | -0.0759 | -0.1405 | -0.1938 |
| 1183:317 | 10 | -0.1017 | -0.1830 | -0.2553 | -0.1062 | -0.1953 | -0.2717 |
| | 25 | -0.0796 | -0.2532 | -0.1987 | -0.0893 | -0.2855 | -0.2324 |
| | 40 | -0.0626 | -0.1130 | -0.1513 | -0.0724 | -0.1387 | -0.1910 |

difference between the two groups decreased as the percent of missing data increased. This occurred for both items. When MI was used to impute the missing data for item 4, the difference between the reference group and focal group was larger than when WMS was used to impute the missing data. The opposite result occurred when MI was used to impute the missing data for item 17. When MI was used for item 17, the difference between the reference group and the focal group was smaller than when WMS was used to impute the missing data.

For item 4, the 500:500 sample size ratio conditions had the largest differences for all DIF magnitudes when 10% of the data were missing. The sample size ratio conditions

Table 17

*Average SMDES for Item 17 with Missing Data*

| | | Method: Mantel | | | | | |
|---|---|---|---|---|---|---|---|
| | | WMS | | | MI | | |
| Sample | % | DIF magnitude | | | DIF magnitude | | |
| size ratio | Missing | .25 | .50 | .75 | .25 | .50 | .75 |
| 500:500 | 10 | -0.1211 | -0.2309 | -0.3425 | -0.1012 | -0.2173 | -0.3364 |
| | 25 | -0.1000 | -0.1896 | -0.2761 | -0.0850 | -0.1836 | -0.2815 |
| | 40 | -0.0812 | -0.1531 | -0.2176 | -0.0664 | -0.1490 | -0.2310 |
| 700:300 | 10 | -0.1215 | -0.2305 | -0.3328 | -0.1012 | -0.2150 | -0.3224 |
| | 25 | -0.0957 | -0.1889 | -0.2682 | -0.0817 | -0.1788 | -0.2681 |
| | 40 | -0.0791 | -0.1487 | -0.2113 | -0.0671 | -0.1443 | -0.2143 |
| 900:100 | 10 | -0.1179 | -0.2260 | -0.3240 | -0.0983 | -0.2106 | -0.3094 |
| | 25 | -0.0970 | -0.1801 | -0.2664 | -0.0822 | -0.1705 | -0.2596 |
| | 40 | -0.0772 | -0.1476 | -0.2120 | -0.0650 | -0.1360 | -0.2073 |
| 845:355 | 10 | -0.1196 | -0.2314 | -0.3283 | -0.1016 | -0.2155 | -0.3163 |
| | 25 | -0.1002 | -0.1880 | -0.2675 | -0.0854 | -0.1799 | -0.2657 |
| | 40 | -0.0789 | -0.1511 | -0.2106 | -0.0642 | -0.1435 | -0.2141 |
| 1183:317 | 10 | -0.1189 | -0.2281 | -0.3246 | -0.0994 | -0.2126 | -0.3125 |
| | 25 | -0.0979 | -0.2399 | -0.2672 | -0.0821 | -0.1809 | -0.2639 |
| | 40 | -0.0779 | -0.1467 | -0.2087 | -0.0650 | -0.1415 | -0.2112 |

with the smallest difference varied for each missing data method as the DIF magnitude increased and as the percent of missing data increased. For WMS results located in Table 16, when the DIF magnitude was .25 and 10% of the data were missing, the 700:300 condition had the smallest difference, but when the DIF magnitude was .75 and 10% of the data were missing, the 900:100 condition had the smallest difference. For item 17, the sample size ratio conditions with the largest difference as well as the smallest difference varied for each missing data method as the DIF magnitude increased and as the percent of missing data increased.

*R-Squared Effect Sizes for Item 4 and Item 17*

Average R-squared effect size values for items 4 and 17 are reported for the power conditions only. In the power conditions, the method of handling missing data, the percent of missing data, the DIF magnitude, the sample size, and the sample size ratio were manipulated.

*No Missing Data*

When no data were missing, the average effect size values for each sample size ratio condition increased as the DIF magnitude increased. The R-squared effect size values for item 4 (Table 18) were slightly lower than the R-squared effect size values for item 17 (Table 19). The conditions with a sample size ratio of 500:500 had the largest R-squared values across all DIF magnitudes. The conditions with a sample size ratio of 1183:317 had the smallest R-squared values. When compared to the other sample size ratio conditions, the conditions with a sample size of 1,000 had larger effect size values than conditions with a sample size of 1,200 or 1,500. Even though the conditions with a sample size ratio of 500:500 had the largest R-squared values, all effect size conditions had very small R-squared values.

Table 18

*Average R-Squared Effect Sizes for Item 4 with No Missing Data*

| | Method: OLR DIF magnitude | | |
|---|---|---|---|
| Sample size ratio | .25 | .50 | .75 |
| 500:500 | 0.0048 | 0.0156 | 0.0314 |
| 700:300 | 0.0044 | 0.0142 | 0.0283 |
| 900:100 | 0.0037 | 0.0114 | 0.0223 |
| 845:355 | 0.0040 | 0.0126 | 0.0247 |
| 1183:317 | 0.0035 | 0.0113 | 0.0221 |

Table 19

*Average R-Squared Effect Sizes for Item 17 with No Missing Data*

| | Method | | |
|---|---|---|---|
| | OLR | | |
| | DIF magnitude | | |
| Sample size ratio | .25 | .50 | .75 |
| 500:500 | 0.0054 | 0.0185 | 0.0393 |
| 700:300 | 0.0050 | 0.0171 | 0.0356 |
| 900:100 | 0.0041 | 0.0136 | 0.0282 |
| 845:355 | 0.0044 | 0.0152 | 0.0315 |
| 1183:317 | 0.0039 | 0.0137 | 0.0280 |

*Missing Data*

When data were missing, the average effect size value for each sample size ratio condition increased as the DIF magnitude increased, but within DIF magnitude conditions, R-squared effect size values decreased as the percent of missing data increased. Like the conditions with no missing data, the 500:500 conditions had the largest R-squared effect size values, whereas the 1183:317 conditions had the lower values. The effect size values for item 4 (Table 20) were also slightly lower than the effect size values for item 17 (Table 21) when data were missing. The conditions where MI was used to impute missing values had larger R-squared effect size values than conditions where WMS was used to impute missing values.

Table 20

*Average R-Squared Effect Sizes for Item 4 with Missing Data*

| | | Method: OLR | | | | | |
|---|---|---|---|---|---|---|---|
| | | WMS | | | MI | | |
| Sample size | | DIF magnitude | | | DIF magnitude | | |
| ratio | % Missing | .25 | .50 | .75 | .25 | .50 | .75 |
| 500:500 | 10 | 0.0037 | 0.0112 | 0.0207 | 0.0040 | 0.0127 | 0.0249 |
| | 25 | 0.0026 | 0.0073 | 0.0130 | 0.0030 | 0.0093 | 0.0179 |
| | 40 | 0.0018 | 0.0041 | 0.0071 | 0.0022 | 0.0059 | 0.0117 |
| 700:300 | 10 | 0.0031 | 0.0092 | 0.0173 | 0.0037 | 0.0116 | 0.0224 |
| | 25 | 0.0022 | 0.0060 | 0.0103 | 0.0028 | 0.0083 | 0.0158 |
| | 40 | 0.0015 | 0.0036 | 0.0058 | 0.0021 | 0.0054 | 0.0103 |
| 900:100 | 10 | 0.0018 | 0.0043 | 0.0124 | 0.0018 | 0.0093 | 0.0177 |
| | 25 | 0.0014 | 0.0029 | 0.0077 | 0.0015 | 0.0067 | 0.0126 |
| | 40 | 0.0011 | 0.0019 | 0.0044 | 0.0012 | 0.0044 | 0.0082 |
| 845:355 | 10 | 0.0032 | 0.0090 | 0.0168 | 0.0037 | 0.0091 | 0.0194 |
| | 25 | 0.0020 | 0.0057 | 0.0100 | 0.0024 | 0.0117 | 0.0134 |
| | 40 | 0.0014 | 0.0033 | 0.0059 | 0.0018 | 0.0043 | 0.0089 |
| 1183:317 | 10 | 0.0024 | 0.0070 | 0.0134 | 0.0022 | 0.0079 | 0.0172 |
| | 25 | 0.0016 | 0.0114 | 0.0080 | 0.0017 | 0.0162 | 0.0121 |
| | 40 | 0.0011 | 0.0027 | 0.0045 | 0.0013 | 0.0038 | 0.0079 |

Table 21

*Average R-Squared Effect Sizes for Item 17 with Missing Data*

| | | Method: OLR | | | | | |
|---|---|---|---|---|---|---|---|
| | | WMS | | | MI | | |
| | | DIF magnitude | | | DIF magnitude | | |
| Sample size ratio | % Missing | .25 | .50 | .75 | .25 | .50 | .75 |
| 500:500 | 10 | 0.0044 | 0.0144 | 0.0307 | 0.0045 | 0.0152 | 0.0327 |
| | 25 | 0.0031 | 0.0096 | 0.0198 | 0.0035 | 0.0111 | 0.0235 |
| | 40 | 0.0021 | 0.0061 | 0.0118 | 0.0025 | 0.0078 | 0.0162 |
| 700:300 | 10 | 0.0038 | 0.0124 | 0.0250 | 0.0039 | 0.0129 | 0.0264 |
| | 25 | 0.0026 | 0.0082 | 0.0159 | 0.0029 | 0.0093 | 0.0187 |
| | 40 | 0.0018 | 0.0049 | 0.0093 | 0.0023 | 0.0064 | 0.0124 |
| 900:100 | 10 | 0.0028 | 0.0056 | 0.0180 | 0.0029 | 0.0059 | 0.0189 |
| | 25 | 0.0020 | 0.0035 | 0.0116 | 0.0023 | 0.0041 | 0.0134 |
| | 40 | 0.0014 | 0.0024 | 0.0068 | 0.0018 | 0.0029 | 0.0089 |
| 845:355 | 10 | 0.0036 | 0.0123 | 0.0247 | 0.0038 | 0.0128 | 0.0260 |
| | 25 | 0.0026 | 0.0080 | 0.0156 | 0.0029 | 0.0092 | 0.0185 |
| | 40 | 0.0017 | 0.0049 | 0.0093 | 0.0021 | 0.0062 | 0.0124 |
| 1183:317 | 10 | 0.0029 | 0.0096 | 0.0191 | 0.0030 | 0.0100 | 0.0201 |
| | 25 | 0.0021 | 0.0100 | 0.0125 | 0.0022 | 0.0092 | 0.0145 |
| | 40 | 0.0013 | 0.0037 | 0.0071 | 0.0017 | 0.0049 | 0.0096 |

# CHAPTER 5

# DISCUSSION

The purpose of this study was to investigate the Type I error and power of several DIF detection methods using several methods of handling missing data for polytomous items generated under the PCM. The Type I error and power of the Mantel and OLR was compared using within-person mean substitution and multiple imputation when data were MCAR. In addition to assessing the Type I error and power of DIF detection methods under several methods of handling missing data, this study also assessed the impact of missing data on effect size measures associated with the Mantel and OLR. Several factors in this study were varied to determine their impact on DIF detection and effect size values. The results are discussed below.

## Power

The results of this Monte Carlo simulation study indicated that the performance of the Mantel and OLR depended on the percentage of missing data, the DIF magnitude, the sample size and the sample size ratio, and the method used to handle the missing data.

### *Percentage of Missing Data*

When less data were missing, the Mantel and OLR were able to detect DIF at higher rates. Conversely, as the percent of missing data increased, the power of the Mantel and OLR to detect DIF decreased. This finding was consistent for both item 4 and item 17. These findings indicate that it is more difficult to detect DIF as the percent of

missing data increases. As a result, the conditions with no missing data had higher power than the conditions with missing data. An exception is noted for conditions with a DIF magnitude of .75 where the power for both the no missing condition and the missing condition was 100%.

Item 17 had higher DIF detection rates than item 4 when no data were missing and when data were missing. Based on the average of the original category intersection parameters for item 4 and item 17, item 17 was easier than item 4. The average item difficulty for item 4 was 1.45, while the average item difficulty for item 17 was 1.13 (see Appendix A for all item parameters). Although item difficulty was not investigated in this study, these findings suggest that item difficulty may have an effect on DIF detection rates.

*DIF Magnitude*

For both item 4 and item 17, the power to detect DIF increased as the DIF magnitude increased. This trend occurred when there were no missing data as well as when missing data were present. Conditions with a DIF magnitude of .25 had the poorest power, while conditions with a DIF magnitude of .75 had the highest power. Conditions with a DIF magnitude of .25 typically had power below 70% which is generally considered adequate power. On average, item 17 had slightly higher power values than item 4. This difference may be due to the varying degree of difficulty of item 4 and item 17.

*Sample Size and Sample Size Ratio*

When there were no missing data, all the conditions with equal sample size ratios (the 500:500 conditions) outperformed the conditions with unequal sample sizes. These

findings were similar to the results of Raiford-Ross (2007). When data were missing, the 500:500 conditions had higher power than the 700:300 and the 900:100 conditions even though all three conditions had the same overall sample size. The conditions with a sample size ratio of 500:500, however, had similar power as the conditions with the same harmonic mean. Power for the 500:500, 845:355, and 1183:317 conditions were similar within and across all DIF magnitude conditions. These results suggest that when data are missing, sample sizes with the same harmonic mean will have similar performance regardless of the size of the reference and focal groups.

*Method Used to Handle the Missing Data*

Within each sample size ratio condition for both items, the Mantel and OLR had similar power rates when WMS was used to impute missing values. For item 4, when MI was used to impute missing values, the Mantel had slightly higher power rates than OLR. This result did not occur as frequently for item 17, though several conditions showed the same effect. At this time, it is not known why this effect occurred. In general, when WMS was used to impute missing values, power rates were higher than when MI was used to impute missing values. These results differed from the findings of Furlow et al. (2003) and Furlow et al. (2007), where MI performed well under study conditions.

The studies mentioned previously investigated the accuracy and precision of theta estimates when data were missing. In those studies, missing item responses were imputed using MI, the *m* data sets were averaged to obtain one data set, and then that final data set was calibrated to obtain theta estimates to compare to the theta estimates that were obtained when no data were missing. In the prior studies, there were no outcomes that needed to be statistically combined in order to obtain a final statistical estimate. This

study, however, needed to statistically combine chi-square values taken from DIF analyses in order to obtain a final chi-square estimate to determine whether an item would be flagged for DIF. It seems that MI would perform better than WMS. As a result, there may need to be a more effective procedure available for combining chi-square values in order to improve power results. Schafer (1997) stated that the statistic used for combining these values has its best performance when three multiply-imputed data sets are requested. This statistic was developed using *m*=3 imputations, and when simulations were conducted to determine the statistic's performance using *m*=2 to *m*=10 imputations, simulations with *m*=3 had the best performance results.

## Type I Error

The Type I error rates for the Mantel and OLR seemed to differ based on the percent of missing data, the absence or presence of impact, and the method used to impute the missing data.When no data were missing for item 4, the presence of impact affected the Type I error results. When impact was present, the Type I error rates for item 4 were higher than when there was no impact. These results were comparable to Kristjansson et al. (2005). When no data were missing for item 17, however, the Type I error rates for the conditions with no impact were similar to the Type I error rates for the conditions with impact. For this relatively easy item, the presence of impact did not have a substantial effect on the Type I error rates.

For item 4, conditions using WMS had Type I error rates closer to the nominal rate of 5% when the Mantel was used to detect DIF, but when OLR was used to detect DIF, several conditions had inflated Type I error rates. This result occurred for both DIF detection methods. For item 17, conditions using WMS had Type I error rates closer to

5% when OLR was the DIF detection method, but there were more conditions with Type I error rates well below the nominal level when the Mantel was used to detect DIF. For both items, conditions with MI had Type I error rates well below the nominal level. As mentioned earlier, this result may be a function of the method used to combine the chi-square values for MI.

## Standardized Mean Difference Effect Size

The SMDES results indicated that as the DIF magnitude increased, the difference between the mean item score of the reference group and the mean item score of the focal group increased. This result occurred when no data were missing and when missing data were handled with WMS and MI. When no data were missing, the results for item 17 exhibited larger differences between the two groups. The difference in item difficulty between item 4 and item 17 may have influenced the effect sizes values for both items, but this would need further evaluation.

When no data were missing, the 500:500 sample size ratio conditions for both item 4 and item 17 had the largest differences between the reference group and the focal group. Although the 500:500 sample size ratio condition had the largest difference, the effect size values obtained for each DIF magnitude condition for all sample size conditions, corresponded to the SMDES guidelines in Chapter 2 where .25 was small DIF, .50 was medium DIF, and .75 was large DIF.

When data were missing, the size of the difference between the reference group and the focal group increased as the DIF magnitude increased, but within a DIF magnitude condition, the difference between the two groups decreased as the percent of missing data increased. The SMDES values for each DIF magnitude condition

corresponded to the SMDES guidelines when 10% of the data were missing, but as the percent of missing data increased, the item was often classified in the next lowest size category. When there were missing data, item 17 had larger differences between the reference group and the focal group than item 4. The differences in effect size values for both items suggest that item difficulty may influence effect size values. The 500:500 conditions for item 4 consistently had the largest differences between the reference and focal groups, but for item 17, no sample size ratio condition showed consistent differences.

## R-squared Effect Size Measure

This study extended previous effect size research for polytomous DIF detection and found that effect size values for the R-squared effect size measure increased as the DIF magnitude increased, but the effect size values decreased as the percent of missing data increased. In addition to the percent of missing data and the DIF magnitude, the sample size ratio also seemed to influence the effect size values. When the sample size of the reference group and the focal group was the same, the effect size increased. However, when the sample size of the reference group and the focal group were unequal, the size of the effect decreased.

The R-squared effect size values in this study were very small in relation to the DIF magnitude that was added to the focal group item parameters for item 4 and item 17. Based on the R-squared effect size guidelines provided in chapter 2, most of the effect size values indicated that the practical significance of the DIF would be negligible. This was the case for all DIF magnitude conditions even conditions with a DIF magnitude of .75 which is generally considered to be a large DIF magnitude. It may be necessary to

review the R-squared effect size guidelines to determine if they are appropriate for classifying polytomous items under these conditions.

The findings from this study indicate that practitioners must be cautious of DIF results and R-squared effect size values for polytomous items when the dataset contain missing data. These items may possess DIF, but the ability to detect DIF may be affected by the amount of missing data present in the data set, the method used to handle missing data, the magnitude of DIF, and the sample size ratio used to conduct the DIF analysis.

## Limitations and Future Research

Although this study extended the research on DIF detection for polytomous items and the effect of missing data on DIF detection rates, this study has several limitations. One limitation of the study is that only one polytomous IRT model was investigated. There are numerous polytomous IRT models available to model data, but this study investigated only one of those models. In addition to the polytomous IRT model, other limitations include the use of only one test length, one percentage for the percent of DIF items, one type of DIF, and one pattern of DIF. Another limitation of the study was the number of conditions varied. Due to the limited number of conditions, it may be difficult to generalize the findings of this study to other polytomous models, assessments with more than 20 items, assessments with more than 10% of the items with DIF, and assessments with items that contain nonuniform DIF or other DIF patterns (such as balanced DIF).

There are several directions for future research. First, other types of DIF detection methods could be investigated. This study investigated two observed score DIF detection methods, but there are latent trait DIF methods, such as Poly-SIBTEST and DFIT, that

could be used for DIF detection as well. Second, average item difficulty could be investigated. Although this was not investigated in this study, it seems that the average item difficulty of the DIF items had an effect on DIF detection. As a result, investigating average item difficulty seems appropriate. Third, other missing data mechanisms such as MAR or MNAR could be explored to determine their effect on DIF detection. DIF detection may differ based on the missing data mechanism. Fourth, a more effective method of combining chi-square values can be investigated to improve the performance of MI when used as the method of handling missing data for DIF analysis. In addition to more investigation of MI, other methods of handling missing data could be explored as well. Fifth, there could be more research on R-squared effect size guidelines. It seems that the guidelines need to be reviewed or new guidelines should be developed because effect sizes corresponding with a large DIF magnitude were considered negligible based on the existing R-squared guidelines. Last, recent studies on DIF detection for polytomous items have studied the effect of the pattern of DIF on DIF detection, but to date, no study has determined the influence of missing data when there are different patterns of DIF. In light of these studies, focusing on the pattern of DIF when there are missing data would be appropriate.

## Summary

This study compared the Type I error and power of the Mantel and OLR using WMS and MI when missing data were MCAR. In addition to assessing the Type I error and power of DIF detection methods and methods of handling missing data, this study also assessed the impact of missing data on the effect size measure associated with the Mantel, the SMDES, and OLR, the R-squared effect size measure. Results indicated that

the performance of the Mantel and OLR depended on the percent of missing data in the data set, the method used to handle the missing data, the DIF magnitude, and the sample size ratio. The Type I error for both DIF detection methods varied based on the missing data method used to impute the missing data. Power to detect DIF increased as DIF magnitude increased, but there was a relative decrease in power as the percent of missing data increased. Additional findings indicated that missing data, DIF magnitude, and sample size ratio also influenced the SMDES values and the R-squared effect size measure values.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67-91.

Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: Wiley.

Allison, P. D. (2002). *Missing data*. CA: Sage.

Andrich, D. (1978). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2,* 581-594.

Bolt, D.M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15(2),* 113-141.

Camilli, G., & Shepard L. (1994). *Methods for identifying biased test items*. CA: Sage.

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33,* 333-353.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. London: Erlbaum.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. CA: Wadsworth Group.

DeAyala, R. J. (1993). An introduction to polytomous item response theory models. *Measurement & Evaluation in Counseling & Development, 25(4).*

DeAyala, R. J. (2003). The effect of missing data on estimating a respondent's location using ratings data. *Journal of Applied Measurement, 4(1),* 1-9.

Dodd, B. G., DeAyala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19,* 5-22.

Dodeen, H. (2004). Stability of differential item functioning over a single population in survey data. *Journal of Experimental Education*, *72,* 181-193.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed-response and differential item functioning: A pragmatic approach* (ETS Research Report No. 91-47). Princeton, NJ: Educational Testing Service.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey: Lawrence Erlbaum.

Furlow, C. F., Fouladi, R., Gagné, P., & Whittaker, T. (2007). A Monte Carlo study of the impact of missing data and differential item functioning on theta estimates under two polytomous Rasch family models. *Journal of Applied Measurement, 8(4),* 388-403.

Furlow, C. F., Fouladi, R., & Whittaker, T. (August, 2003). The impact of missing data, and procedures for handling missing data using an item response theory framework. Paper presented at the Joint Statistical Meetings, San Francisco.

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice,* 3-14.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. CA: Sage.

Holland, P.W. & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Huisman, M. (2000). Imputation of missing item responses: Some simple techniques. *Quality & Quantity, 34,* 331-351.

Jodoin, M. G. & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14(4)*, 329-349.

Kamata, A., & Vaughn, B. K. (2004). An introduction to differential item functioning analysis. *Learning Disabilities: A Contemporary Journal, 2(2),* 49-69.

Kim, S.-H., Cohen, K., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44(2),* 93-116.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, *65*(6), 935-953.

Li, K.H., Keng, X.L., Raghunathan, T.E., and Rubin, D.B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, 1, 65-92.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Mapuranga, R., Dorans, N. J., & Middleton, K. (March, 2008). A review of recent developments in differential item functioning. Paper presented at the National Council on Measurement in Education, New York.

Mardell, C., & Goldenberg, D. S. (1972). *DIAL: Developmental Indicators for the Assessment of Learning*. Highland Park, Ill: DIAL Inc.

Mardell, C., & Goldenberg, D. S. (1975). For prekindergarten screening information: DIAL. *Journal of Learning Disabilities, 8,* 140-147.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Masters, G. N. (1988). The analysis of partial credit scoring. *Applied Measurement in Education, 1(4),* 279-297.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. NY: Guilford Press.

Meyer, J. P., Huynh, H., & Seaman, M. A. (2004). Exact small-sample differential item functioning methods for polytomous items with illustration based on an attitude survey. *Journal of Educational Measurement, 41(4)*, 331-344.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297-334.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. CA: Sage.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19(3),* 5-15.

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

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19,* 23-37.

Raiford-Ross, T. (2007). The impact of multidimensionality on the detection of differential bundle functioning using SIBTEST. Unpublished doctoral dissertation, Educational Policy Studies, Georgia State University.

Raju, N. S., van der Linden, W. J., & Fleer, P.F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, 19, 353-368.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

SAS Institute (2005). *SAS/IML software: Changes and enhancements, through release 9.1.* Cary, NC: SAS Institute.

Schaeffer, G. A., Henderson-Montero, D., Julian, M., & Bené, N. H. (2002). A comparison of three scoring methods for tests with selected-response and constructed-response items. *Educational Asssessment, 8(4),* 317-340.

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

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the Art. *Psychological Methods, 7,* 147-177.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6,* 317-329.

SPSS (2005). *Statistical Package for the Social Sciences*. Chicago, IL: SPSS.

Statistical Solutions (2006). *SOLAS for Misssing Data Analysis, Version 3.0*. Saugus, MA: Statistical Solutions.

Su, Y.-H., & Wang, W.-C. (2005). Efficiency of the Mantel, generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education*, *18*(4), 313-350.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27(4)*, 361-370.

Sykes, R. C., & Hou, L. (2003). Weighting constructed-response items in IRT-based exams. *Applied Measurement in Education, 16(4),* 257-275.

Thissen, D., & Wainer, H. (2001). *Test scoring*. New Jersey: Lawrence Erlbaum.

Wang, W.-C., & Su, Y.-H. (2004). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28,* 450-481.

Welch, C., & Miller, H.D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education, 6(1),* 1-19.

Welch, C. J., & Miller, T. R. (1995). Assessing differential item functioning in direct writing assessments: Problems and an example. *Journal of Educational Measurement, 32,* 163-178.

Whittaker, T., Fitzpatrick, S., Dodd, B., & Williams, N. (2003). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models. *Applied Psychological Measurement*, *27*(4), 299-300.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233-251.

Zwick, R. & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21(3),* 187-201.

APPENDIXES

APPENDIX A

ITEM PARAMETERS FOR DATA GENERATION

| Reference Group Item Parameters | | | Focal Group Item Parameters | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | DIF Magnitudes | | | | | | | | |
| | | | .25 | | | .50 | | | .75 | | |
| $b_1$ | $b_2$ | $b_3$ | $b_1$ | $b_2$ | $b_3$ | $b_1$ | $b_2$ | $b_3$ | $b_1$ | $b_2$ | $b_3$ |
| -1.33 | -1.28 | -1.03 | | | | | | | | | |
| -0.91 | 0.98 | 0.21 | | | | | | | | | |
| -2.25 | 1.21 | 3.47 | | | | | | | | | |
| 1.81 | 1.07 | 1.46 | 2.06 | 1.32 | 1.71 | 2.31 | 1.57 | 1.96 | 2.56 | 1.82 | 2.21 |
| 0.32 | 0.86 | 2.21 | | | | | | | | | |
| 0.58 | 0.63 | -0.49 | | | | | | | | | |
| -1.76 | -0.09 | 0.19 | | | | | | | | | |
| -1.49 | -0.83 | 2.66 | | | | | | | | | |
| -2.20 | -1.33 | -0.48 | | | | | | | | | |
| -2.25 | -1.80 | 1.66 | | | | | | | | | |
| -0.54 | -2.11 | 0.74 | | | | | | | | | |
| -0.91 | -0.93 | 1.29 | | | | | | | | | |
| -0.91 | 0.98 | 0.21 | | | | | | | | | |
| -2.25 | 1.21 | 3.47 | | | | | | | | | |
| -1.34 | 1.72 | 3.40 | | | | | | | | | |
| 1.81 | 1.07 | 1.46 | | | | | | | | | |
| 0.32 | 0.86 | 2.21 | 0.57 | 1.11 | 2.46 | 0.82 | 1.36 | 2.71 | 1.07 | 1.61 | 2.96 |
| -1.76 | -0.09 | 0.19 | | | | | | | | | |
| -2.50 | -0.85 | 2.28 | | | | | | | | | |
| -1.49 | -0.83 | 2.66 | | | | | | | | | |

*Note*. Blanks indicate the use of the same item parameters listed for the reference group.

APPENDIX B

SAS Program for DIF Detection and Effect Size Measures

```
%macro ss;
options mprint;
%let miss1=10;
%let miss2=25;
%let miss3=40;
%let difp=10;
%let mag=25;
%let ratio1=5to5;
%let ratio2=7to3;
%let ratio3=9to1;
%let ratio4=317to1183;
%let ratio5=355to845;

%do ratsize=1 %to 1; *runs conditions where ratio is 5to5, 7to3, 9to1,
etc;
%do permiss = 1 %to 3; *runs conditions where missingness is 10, 25,
40;

      %do i=1 %to 1000; *replication loop;

%let rat=&&ratio&ratsize;
proc printto log="F:\LOGSTUFFDIF 10 25 40_&difp._&mag._&rat..txt" new;
PROC PRINTTO PRINT="F:\OUTPUTSTUFFDIF 10 25 40_&difp._&mag._&rat..TXT"
NEW;
*********************************************************************;

*MEAN SUBSTITUTION;

*********************************************************************;
%let pmiss=&&miss&permiss;
%let rat=&&ratio&ratsize;

DATA DD;
inFILE
"F:\&pmiss._&DIFP._&MAG._&rat.\pc_20_MEAN_&pmiss._&DIFP._&MAG._&rat._90
\PC10&I..dat";
      INPUT R1-R20 THETA FOC;
run;

DATA EQUINT; set DD;
tot4=sum(of r1-R16 R18-r20); *making matching var(total score):sum of
all items except the second DIF item;
tot17=sum(of r1-R3 R5-r20); * studied item is included in the total
score;
```

```
IF tot4 >= 19 AND tot4 =< 37 THEN feqint4 = 1; *creating interval
variables for every simulee;
else IF tot4 >= 38 AND tot4 =< 46 THEN feqint4 = 2; *this variable
indicates what frequency interval the simulee's total score will be in;
else IF tot4 >= 47 AND tot4 =< 55 THEN feqint4 = 3;
else IF tot4 >= 56 AND tot4 =< 64 THEN feqint4 = 4;
else IF tot4 >= 65 AND tot4 =< 76 THEN feqint4 = 5;

IF tot17 >= 19 AND tot17 =< 37 THEN feqint17 = 21;
else IF tot17 >= 38 AND tot17 =< 46 THEN feqint17 = 22;
else IF tot17 >= 47 AND tot17 =< 55 THEN feqint17 = 23;
else IF tot17 >= 56 AND tot17 =< 64 THEN feqint17 = 24;
else IF tot17 >= 65 AND tot17 =< 76 THEN feqint17 = 25;

proc sort DATA=EQUINT OUT=D4; by feqint4; *puts all total scores into
its appropriate interval;
proc sort DATA=EQUINT OUT=D17; by feqint17;

ods output crosstabfreqs=Mantel1;
ods output CMH=GMH4 ;
ods listing ;  run;
proc freq DATA=D4;
      tables fEQINT4*foc*r4 /cmh; RUN;
ods trace off;

ods output crosstabfreqs=Mantel2;
ods output CMH=GMH17 ;
ods listing ;  run;
proc freq DATA=D17;
      tables fEQINT17*foc*r17 /cmh; RUN;
ods trace off;

data MEAN1; set gmh4 GMH17; keep table prob althypothesis ;
      if statistic=2; *picks up only the Mantel info from the output;

data allmean; set mean1; keep item prob difmethod;
      if _n_=1 then do;
            item = 4 ; difmethod=1;
      end;

      if _n_=2 then do;
            item = 17; difmethod = 1;
      end;

***********************SMD calculations for item
4*************************************************;
data intervals; set Mantel1; *keeping the total number of reference
group members at each interval;
keep R4 frequency FOC _TABLE_;     *this is done five times for each of
the five intervals;
if R4 ne '.' then delete;
if foc='.' then delete;
run;

proc sort data=intervals out=sortinter;
by foc;
run;
```

```
data combref; set sortinter;
drop R4;
if foc='0';
sumfreqref+frequency;
run;

data combfoc; set sortinter;
drop R4;
if foc='1';
sumfreqfoc+frequency;
run;

data sumref; set combref; *keeping only the final number of ref group
members across all intervals;
keep totalfreqref;
if _N_=5;
totalfreqref=sumfreqref;
run;

data sumref2; set sumref combref;
run;

data sumref3; *adding this final number of ref group members to the set
of ref group members at each interval;
if _N_=1 then set sumref; set combref;
prop=frequency/totalfreqref;
if _N_=1 then type='propref1';
else if _N_=2 then type='propref2';
else if _N_=3 then type='propref3';
else if _N_=4 then type='propref4';
else if _N_=5 then type='propref5';
run;

proc transpose data=sumref3 out=transref; *keeps only the VAR variable;
ID type;* this variable becomes new variable names,and the original
name disappears;
VAR prop; *this one variable becomes five new variables;
run;

data transref; set transref;
drop _NAME_;
try=1;
run;

data sumfoc; set combfoc; *keeping only the final number of focal group
members across all intervals;
keep totalfreqfoc;
if _N_=5;
totalfreqfoc=sumfreqfoc;
run;

data sumfoc2; set sumfoc combfoc;
run;

data sumfoc3;  *adding this final number of focal group members to the
set of focal group members at each interval;
```

```
if _N_=1 then set sumfoc; set combfoc;
prop=frequency/totalfreqfoc;
if _N_=1 then type='propfoc1';
else if _N_=2 then type='propfoc2';
else if _N_=3 then type='propfoc3';
else if _N_=4 then type='propfoc4';
else if _N_=5 then type='propfoc5';
run;

proc transpose data=sumfoc3 out=transfoc;
ID type;* this variable becomes new variable names,and the original
name disappears;
VAR prop; *this one variable becomes five new variables;
run;

data transfoc; set transfoc;
drop _NAME_;
try=1;
run;
***************************item mean calculation for the ref
group***************************;
data itemmean1; set Mantel1;
keep foc R4  _TABLE_ frequency;
run;

data itemmeanref; set itemmean1; *calculating the total number of
points at each score point for each interval(sumscore)for the ref
group;
if FOC='0';
sumscore=R4*frequency;
run;

data itemmeanfoc; set itemmean1; *doing the same for the focal group;
if FOC='1';
sumscore=R4*frequency;
run;

data tottrans; set combref;
if _N_=1 then type='totref1';
else if _N_=2 then type='totref2';
else if _N_=3 then type='totref3';
else if _N_=4 then type='totref4';
else if _N_=5 then type='totref5';
run;

proc transpose data=tottrans out=transm;
ID type;
VAR frequency;
run;


data transm; set transm;
drop _NAME_;
try=1;
run;
```

```
data tableref1; set itemmeanref; *keeping score info from interval 1
for the reference group;
if _TABLE_='1';
sum_sumscore1+sumscore;
try=1;
run;

data reffreq1; set tableref1;
keep sum_sumscore1 try;
if _N_=5;
run;

data tableref2; set itemmeanref; *doing the same for the other
intervals;
if _TABLE_='2';
sum_sumscore2+sumscore;
try=1;
run;

data reffreq2; set tableref2;
keep sum_sumscore2 try;
if _N_=5;
run;

data tableref3; set itemmeanref;
if _TABLE_='3';
sum_sumscore3+sumscore;
try=1;
run;

data reffreq3; set tableref3;
keep sum_sumscore3 try;
if _N_=5;
run;

data tableref4; set itemmeanref;
if _TABLE_='4';
sum_sumscore4+sumscore;
try=1;
run;

data reffreq4; set tableref4;
keep sum_sumscore4 try;
if _N_=5;
run;

data tableref5; set itemmeanref;
if _TABLE_='5';
sum_sumscore5+sumscore;
try=1;
run;

data reffreq5; set tableref5;
keep sum_sumscore5 try;
if _N_=5;
run;
```

```
data meanref; merge transm reffreq1 reffreq2 reffreq3 reffreq4
reffreq5; by try;
mref1=sum_sumscore1/totref1;
mref2=sum_sumscore2/totref2;
mref3=sum_sumscore3/totref3;
mref4=sum_sumscore4/totref4;
mref5=sum_sumscore5/totref5;
run;

data meanall; set meanref;
keep try mref1 mref2 mref3 mref4 mref5;
run;

data propmeanall; merge transref meanall; by try;
run;

data combreffoc; merge propmeanall transfoc; by try; *merging the ref
and focal group prop and item mean score sets together;
run;




*****************calculating the total item mean score for the focal
group*************************;
data foctable; set itemmeanfoc; *summing all the item scores and
frequencies of the item scores across all five intervals;
keep R4 frequency sumscore sum_freq sum_sumscore;
if R4='.' then delete;
sum_freq+frequency;
sum_sumscore+sumscore;
run;

data foctable2; set foctable; *creating the total item mean score for
all the focal group intervals;
keep try totalmeanfoc sum_freq sum_sumscore;
if _N_=20;
totalmeanfoc=sum_sumscore/sum_freq;
try=1;
run;

data SMD; merge combreffoc foctable2; by try; *calculating the
standardized mean difference;
smd=totalmeanfoc-
((mref1*propfoc1)+(mref2*propfoc2)+(mref3*propfoc3)+(mref4*propfoc4)+(m
ref5*propfoc5));
run;

***********************calculating the pooled standard
deviation*********************************;
data SDtable; set equint; *change to equint2 for MI;
keep R4 foc;* keeping item 4 scores and the FOC variable;
run;

proc means data=sdtable noprint; *finding the mean and stand deviation
of all the scores for item 4 for the ref and focal group;
by foc;
```

```
var R4;
output out=R4mean;
run;

data refSTD; set R4mean; *keeping the stand deviation of the ref group
only;
keep try refSTD;
if _STAT_='STD' and Foc='0';
refSTD=R4;
try=1;
run;



data refN; set R4mean;
keep try refN;
if _STAT_='N' and Foc='0';
refN=R4;
try=1;
run;

data focSTD; set R4mean; *keeping the stand deviation of the focal
group only;
keep try focSTD;
if _STAT_='STD' and Foc='1';
focSTD=R4;
try=1;
run;

data focN; set R4mean;
keep try focN;
if _STAT_='N' and Foc='1';
focN=R4;
try=1;
run;

data smdes; merge refSTD refN focSTD focN smd; by try; *calculating the
pooled stand deviation of the ref and focal groups;
keep refSTD refN focSTD focN smd pooledSTD smdes;
pooledSTD=sqrt(((refN-1)*refSTD**2 + (focN-1)*focSTD**2)/(refN+focN-
2));
smdes=smd/pooledSTD;
run;

data onlysmdes4; set smdes; *dataset with SMD effect size value, item#,
missing method, and replication#;
keep item smdes miss rep;
item=4;
miss=1;
rep=&i;
run;
```

```
****************SMD calculations for item
17*****************************************************;
data intervals; set Mantel2; *keeping the total number of reference
group members at each interval;
keep R17 frequency FOC _TABLE_;     *this is done five times for each
of the five intervals;
if R17 ne '.' then delete;
if foc='.' then delete;
run;

proc sort data=intervals out=sortinter;
by foc;
run;

data combref; set sortinter;
drop R17;
if foc='0';
sumfreqref+frequency;
run;

data combfoc; set sortinter;
drop R17;
if foc='1';
sumfreqfoc+frequency;
run;

data sumref; set combref;
keep totalfreqref;
if _N_=5;
totalfreqref=sumfreqref;
run;

data sumref2; set sumref combref;
run;

data sumref3;
if _N_=1 then set sumref; set combref;
prop=frequency/totalfreqref;
if _N_=1 then type='propref1';
else if _N_=2 then type='propref2';
else if _N_=3 then type='propref3';
else if _N_=4 then type='propref4';
else if _N_=5 then type='propref5';
run;

proc transpose data=sumref3 out=transref; *only keeps the VAR variable;
ID type;* this variable becomes new variable names,and the original
name disappears;
VAR prop; *this one variable becomes five new variables;
run;

data transref; set transref;
drop _NAME_;
try=1;
run;

data sumfoc; set combfoc;
```

```
keep totalfreqfoc;
if _N_=5;
totalfreqfoc=sumfreqfoc;
run;

data sumfoc2; set sumfoc combfoc;
run;

data sumfoc3;
if _N_=1 then set sumfoc; set combfoc;
prop=frequency/totalfreqfoc; *combine each of these table proportions
with the itemmean for each table;
if _N_=1 then type='propfoc1';
else if _N_=2 then type='propfoc2';
else if _N_=3 then type='propfoc3';
else if _N_=4 then type='propfoc4';
else if _N_=5 then type='propfoc5';
run;

proc transpose data=sumfoc3 out=transfoc;
ID type;* this variable becomes new variable names,and the original
name disappears;
VAR prop; *this one variable becomes five new variables;
run;

data transfoc; set transfoc;
drop _NAME_;
try=1;
run;
***************************item mean calculation for the ref
group***************************;
data itemmean1; set Mantel2;
keep foc R17  _TABLE_ frequency;
run;

data itemmeanref; set itemmean1;
if FOC='0';
sumscore=R17*frequency;
run;

data itemmeanfoc; set itemmean1;
if FOC='1';
sumscore=R17*frequency;
run;

data tottrans; set combref;
if _N_=1 then type='totref1';
else if _N_=2 then type='totref2';
else if _N_=3 then type='totref3';
else if _N_=4 then type='totref4';
else if _N_=5 then type='totref5';
run;

proc transpose data=tottrans out=transm;
ID type;
VAR frequency;
run;
```

```
data transm; set transm;
drop _NAME_;
try=1;
run;

data tableref1; set itemmeanref; *keeping score info from interval 1
for the reference group;
if _TABLE_='1';
sum_sumscore1+sumscore;
try=1;
run;

data reffreq1; set tableref1;
keep sum_sumscore1 try;
if _N_=5;
run;

data tableref2; set itemmeanref; *doing the same for the other
intervals;
if _TABLE_='2';
sum_sumscore2+sumscore;
try=1;
run;

data reffreq2; set tableref2;
keep sum_sumscore2 try;
if _N_=5;
run;

data tableref3; set itemmeanref;
if _TABLE_='3';
sum_sumscore3+sumscore;
try=1;
run;

data reffreq3; set tableref3;
keep sum_sumscore3 try;
if _N_=5;
run;

data tableref4; set itemmeanref;
if _TABLE_='4';
sum_sumscore4+sumscore;
try=1;
run;

data reffreq4; set tableref4;
keep sum_sumscore4 try;
if _N_=5;
run;

data tableref5; set itemmeanref;
if _TABLE_='5';
sum_sumscore5+sumscore;
try=1;
run;
```

```
data reffreq5; set tableref5;
keep sum_sumscore5 try;
if _N_=5;
run;

data meanref; merge transm reffreq1 reffreq2 reffreq3 reffreq4
reffreq5; by try;
mref1=sum_sumscore1/totref1;
mref2=sum_sumscore2/totref2;
mref3=sum_sumscore3/totref3;
mref4=sum_sumscore4/totref4;
mref5=sum_sumscore5/totref5;
run;

data meanall; set meanref;
keep try mref1 mref2 mref3 mref4 mref5;
run;

data propmeanall; merge transref meanall; by try;
run;

data combreffoc; merge propmeanall transfoc; by try;
run;

data foctable; set itemmeanfoc; *summing all the item scores and
frequencies of the item scores;
keep R17 frequency sumscore sum_freq sum_sumscore;
if R17='.' then delete;
sum_freq+frequency;
sum_sumscore+sumscore;
run;

data foctable2; set foctable; *creating the total item mean score for
all the focal group intervals;
keep try totalmeanfoc sum_freq sum_sumscore;
if _N_=20;
totalmeanfoc=sum_sumscore/sum_freq;
try=1;
run;

data SMD; merge combreffoc foctable2; by try;
smd=totalmeanfoc-
((mref1*propfoc1)+(mref2*propfoc2)+(mref3*propfoc3)+(mref4*propfoc4)+(m
ref5*propfoc5));
run;

data SDtable; set equint; *change to equint2 for MI;
keep R17 foc;
run;

proc means data=sdtable noprint;
by foc;
var R17;
output out=R17mean;
run;
```

```
data refSTD; set R17mean;
keep try refSTD;
if _STAT_='STD' and Foc='0';
refSTD=R17;
try=1;
run;

data refN; set R17mean;
keep try refN;
if _STAT_='N' and Foc='0';
refN=R17;
try=1;
run;

data focSTD; set R17mean;
keep try focSTD;
if _STAT_='STD' and Foc='1';
focSTD=R17;
try=1;
run;

data focN; set R17mean;
keep try focN;
if _STAT_='N' and Foc='1';
focN=R17;
try=1;
run;

data smdes; merge refSTD refN focSTD focN smd; by try;
keep refSTD refN focSTD focN smd pooledSTD smdes;
pooledSTD=sqrt(((refN-1)*refSTD**2 + (focN-1)*focSTD**2)/(refN+focN-
2));
smdes=smd/pooledSTD;
run;

data onlysmdes17; set smdes; *dataset with SMD effect size value,
item#, missing method, and replication#;
keep item smdes miss rep;
item=17;
miss=1;
rep=&i;
run;

data smd_1; set onlysmdes4 onlysmdes17;
run;

proc logistic data=EQUINT; * OLR model 1 to get model 1 rsquare;
model R4=TOT4/rsquare;
ods output Rsquare=Rsquare4_one;
run;

data meanrsq4_one; set rsquare4_one;
keep item rsquare1;
item=4;
rsquare1=cvalue1;
run;
```

```
proc logistic data = EQUINT ; *OLR model 2;
  model R4 =  TOT4 FOC/rsquare;
  ods output Rsquare=Rsquare4;
  ods output  ParameterEstimates = LOGIST4;

data meanrsq4; set rsquare4;
keep item rsquare2;
item=4;
rsquare2=cvalue1;
run;




DATA MEANLOG4; SET LOGIST4; KEEP ProbChiSq ITEM;
      IF VARIABLE = 'FOC';
      ITEM=4;

data meanlog4_comb; merge meanrsq4 meanrsq4_one; by item;
keep item rsq miss rep;
rsq=rsquare2-rsquare1;
miss=1;
rep=&i;
run;

proc logistic data=EQUINT; * OLR model 1 to get model 1 rsquare;
model R17=TOT17/rsquare;
ods output Rsquare=Rsquare17_one;
run;

data meanrsq17_one; set rsquare17_one;
keep item rsquare1;
item=17;
rsquare1=cvalue1;
run;

proc logistic data = EQUINT ; *OLR model 2;
  model R17 =  TOT17 FOC/rsquare;
  ods output Rsquare=Rsquare17;
  ods output  ParameterEstimates = LOGIST17;

data meanrsq17; set rsquare17;
keep item rsquare2;
item=17;
rsquare2=cvalue1;
run;

DATA MEANLOG17; SET LOGIST17; KEEP ProbChiSq ITEM;
      IF VARIABLE = 'FOC';
      ITEM=17;

data meanlog17_comb; merge meanrsq17 meanrsq17_one; by item;
keep item rsq miss rep;
rsq=rsquare2-rsquare1;
miss=1;
rep=&i;
```

```
run;

data rsq_1; set meanlog4_comb meanlog17_comb;
run;


DATA MEANLOG; SET MEANLOG4 MEANLOG17; keep item prob difmethod;
      prob = probchisq;
      DIFmethod=2;
      run;

data meandif; set allmean meanlog;
rep = &i;
miss = 1;   run;
/*miss = 1 for mean substitution, 2 for hotdecking, and 3 for multiple
imputation */

*********************************************************************;

*MULTIPLE IMPUTATION;

*********************************************************************;
%do IMP=1 %to 5;
DATA d2;
inFILE
"F:\&pmiss._&DIFP._&MAG._&rat.\pc_20_MI&IMP._&pmiss._&DIFP._&MAG._&rat.
_90\PC10&I..dat";
      INPUT R1-R20 THETA Foc;
RUN;

DATA EQUINT2; set D2;
tot4=sum(of r1-R16 R18-r20); *pull out all DIF items other than studied
item;
tot17=sum(of r1-R3 R5-r20);

IF tot4 >= 19 AND tot4 =< 37 THEN feqint4 = 1;
else IF tot4 >= 38 AND tot4 =< 46 THEN feqint4 = 2;
else IF tot4 >= 47 AND tot4 =< 55 THEN feqint4 = 3;
else IF tot4 >= 56 AND tot4 =< 64 THEN feqint4 = 4;
else IF tot4 >= 65 AND tot4 =< 76 THEN feqint4 = 5;

IF tot17 >= 19 AND tot17 =< 37 THEN feqint17 = 21;
else IF tot17 >= 38 AND tot17 =< 46 THEN feqint17 = 22;
else IF tot17 >= 47 AND tot17 =< 55 THEN feqint17 = 23;
else IF tot17 >= 56 AND tot17 =< 64 THEN feqint17 = 24;
else IF tot17 >= 65 AND tot17 =< 76 THEN feqint17 = 25;

proc sort DATA=EQUINT2 OUT=D4; by feqint4;
proc sort DATA=EQUINT2 OUT=D17; by feqint17;

ods output crosstabfreqs=Mantel1;
ods output CMH=GMH4 ;
ods listing ;  run;
proc freq DATA=D4;
      tables fEQINT4*foc*r4 /cmh; RUN;

ods output crosstabfreqs=Mantel2;
```

```
ods output CMH=GMH17 ;
ods listing ;  run;
proc freq DATA=D17;
      tables fEQINT17*foc*r17 /cmh; RUN;

data Multimp1; set gmh4 GMH17; keep table VALUE althypothesis ;
      if statistic=2; *picks up only the mantel info from the output;

data allMI; set Multimp1; keep item CHI&IMP difmethod;
      if _n_=1 then do;
            item = 1 ; difmethod=1;
      end;

      if _n_=2 then do;
            item = 2; difmethod = 1;
      end;

       CHI&IMP=VALUE;

************************SMD calculations for item
4*************************************************;
data intervals; set Mantel1; *keeping the total number of reference
group members at each interval;
keep R4 frequency FOC _TABLE_;     *this is done five times for each of
the five intervals;
if R4 ne '.' then delete;
if foc='.' then delete;
run;


proc sort data=intervals out=sortinter;
by foc;
run;

data combref; set sortinter;
drop R4;
if foc='0';
sumfreqref+frequency;
run;

data combfoc; set sortinter;
drop R4;
if foc='1';
sumfreqfoc+frequency;
run;

data sumref; set combref; *keeping only the final number of ref group
members across all intervals;
keep totalfreqref;
if _N_=5;
totalfreqref=sumfreqref;
run;

data sumref2; set sumref combref;
run;
```

```
data sumref3; *adding this final number of ref group members to the set
of ref group members at each interval;
if _N_=1 then set sumref; set combref;
prop=frequency/totalfreqref;
if _N_=1 then type='propref1';
else if _N_=2 then type='propref2';
else if _N_=3 then type='propref3';
else if _N_=4 then type='propref4';
else if _N_=5 then type='propref5';
run;

proc transpose data=sumref3 out=transref; *only keeps the VAR variable;
ID type;* this variable becomes new variable names,and the original
name disappears;
VAR prop; *this one variable becomes five new variables;
run;

data transref; set transref;
drop _NAME_;
try=1;
run;

data sumfoc; set combfoc; *keeping only the final number of focal group
members across all intervals;
keep totalfreqfoc;
if _N_=5;
totalfreqfoc=sumfreqfoc;
run;

data sumfoc2; set sumfoc combfoc;
run;

data sumfoc3;  *adding this final number of focal group members to the
set of ref group members at each interval;
if _N_=1 then set sumfoc; set combfoc;
prop=frequency/totalfreqfoc;
if _N_=1 then type='propfoc1';
else if _N_=2 then type='propfoc2';
else if _N_=3 then type='propfoc3';
else if _N_=4 then type='propfoc4';
else if _N_=5 then type='propfoc5';
run;

proc transpose data=sumfoc3 out=transfoc; *only keeps the VAR variable;
ID type;* this variable becomes new variable names,and the original
name disappears;
VAR prop; *this one variable becomes five new variables;
run;

data transfoc; set transfoc;
drop _NAME_;
try=1;
run;
****************************************item mean calculation for the
ref group***************************;
data itemmean1; set Mantel1;
keep foc R4  _TABLE_ frequency;
```

```
run;

data itemmeanref; set itemmean1; *calculating the total number of
points at each score point for each interval(sumscore)for the ref
group;
if FOC='0';
sumscore=R4*frequency;
run;

data itemmeanfoc; set itemmean1; *doing the same for the focal group;
if FOC='1';
sumscore=R4*frequency;
run;

data tottrans; set combref;
if _N_=1 then type='totref1';
else if _N_=2 then type='totref2';
else if _N_=3 then type='totref3';
else if _N_=4 then type='totref4';
else if _N_=5 then type='totref5';
run;

proc transpose data=tottrans out=transm;
ID type;
VAR frequency;
run;

data transm; set transm;
drop _NAME_;
try=1;
run;

data tableref1; set itemmeanref; *keeping score info from interval 1
for the reference group;
if _TABLE_='1';
sum_sumscore1+sumscore;
try=1;
run;

data reffreq1; set tableref1;
keep sum_sumscore1 try;
if _N_=5;
run;

data tableref2; set itemmeanref; *doing the same for the other
intervals;
if _TABLE_='2';
sum_sumscore2+sumscore;
try=1;
run;

data reffreq2; set tableref2;
keep sum_sumscore2 try;
if _N_=5;
run;

data tableref3; set itemmeanref;
```

```
if _TABLE_='3';
sum_sumscore3+sumscore;
try=1;
run;

data reffreq3; set tableref3;
keep sum_sumscore3 try;
if _N_=5;
run;

data tableref4; set itemmeanref;
if _TABLE_='4';
sum_sumscore4+sumscore;
try=1;
run;

data reffreq4; set tableref4;
keep sum_sumscore4 try;
if _N_=5;
run;

data tableref5; set itemmeanref;
if _TABLE_='5';
sum_sumscore5+sumscore;
try=1;
run;

data reffreq5; set tableref5;
keep sum_sumscore5 try;
if _N_=5;
run;

data meanref; merge transm reffreq1 reffreq2 reffreq3 reffreq4
reffreq5; by try;
mref1=sum_sumscore1/totref1;
mref2=sum_sumscore2/totref2;
mref3=sum_sumscore3/totref3;
mref4=sum_sumscore4/totref4;
mref5=sum_sumscore5/totref5;
run;

data meanall; set meanref;
keep try mref1 mref2 mref3 mref4 mref5;
run;

data propmeanall; merge transref meanall; by try;
run;

data combreffoc; merge propmeanall transfoc; by try; *merging the ref
and focal group prop and item mean score sets together;
run;

*****************calculating the total item mean score for the focal
group************************;
data foctable; set itemmeanfoc; *summing all the item scores and
frequencies of the item scores across all five intervals;
keep R4 frequency sumscore sum_freq sum_sumscore;
```

```
if R4='.' then delete;
sum_freq+frequency;
sum_sumscore+sumscore;
run;

data foctable2; set foctable; *creating the total item mean score for
all the focal group intervals;
keep try totalmeanfoc sum_freq sum_sumscore;
if _N_=20;
totalmeanfoc=sum_sumscore/sum_freq;
try=1;
run;

data SMD; merge combreffoc foctable2; by try; *calculating the
standardized mean difference;
smd=totalmeanfoc-
((mref1*propfoc1)+(mref2*propfoc2)+(mref3*propfoc3)+(mref4*propfoc4)+(m
ref5*propfoc5));
run;

*******************calculating the pooled standard
deviation*********************************;
data SDtable; set equint2; *change to equint2 for MI;
keep R4 foc;* keeping item 4 scores and the FOC variable;
run;

proc means data=sdtable noprint; *finding the mean and stand deviation
of all the scores for item 4 for the ref and focal group;
by foc;
var R4;
output out=R4mean;
run;

data refSTD; set R4mean; *keeping the stand deviation of the ref group
only;
keep try refSTD;
if _STAT_='STD' and Foc='0';
refSTD=R4;
try=1;
run;

data refN; set R4mean;
keep try refN;
if _STAT_='N' and Foc='0';
refN=R4;
try=1;
run;

data focSTD; set R4mean; *keeping the stand deviation of the focal
group only;
keep try focSTD;
if _STAT_='STD' and Foc='1';
focSTD=R4;
try=1;
run;

data focN; set R4mean;
```

```
keep try focN;
if _STAT_='N' and Foc='1';
focN=R4;
try=1;
run;

data smdes; merge refSTD refN focSTD focN smd; by try; *calculating the
pooled stand deviation of the ref and focal groups;
keep refSTD refN focSTD focN smd pooledSTD smdes;
pooledSTD=sqrt(((refN-1)*refSTD**2 + (focN-1)*focSTD**2)/(refN+focN-
2));
smdes=smd/pooledSTD;
run;

data onlysmdes4; set smdes; *dataset with SMD effect size value, item#,
missing method, and replication#;
keep item smdes miss rep;
item=4;
miss=3;
rep=&i;
run;

****************SMD calculations for item
17*******************************************************;
data intervals; set Mantel2; *keeping the total number of reference
group members at each interval;
keep R17 frequency FOC _TABLE_;     *this is done five times for each
of the five intervals;
if R17 ne '.' then delete;
if foc='.' then delete;
run;

proc sort data=intervals out=sortinter;
by foc;
run;

data combref; set sortinter;
drop R17;
if foc='0';
sumfreqref+frequency;
run;

data combfoc; set sortinter;
drop R17;
if foc='1';
sumfreqfoc+frequency;
run;

data sumref; set combref;
keep totalfreqref;
if _N_=5;
totalfreqref=sumfreqref;
run;

data sumref2; set sumref combref;
run;
```

```
data sumref3;
if _N_=1 then set sumref; set combref;
prop=frequency/totalfreqref;
if _N_=1 then type='propref1';
else if _N_=2 then type='propref2';
else if _N_=3 then type='propref3';
else if _N_=4 then type='propref4';
else if _N_=5 then type='propref5';
run;

proc transpose data=sumref3 out=transref; *only keeps the VAR variable;
ID type;* this variable becomes new variable names,and the original
name disappears;
VAR prop; *this one variable becomes five new variables;
run;

data transref; set transref;
drop _NAME_;
try=1;
run;

data sumfoc; set combfoc;
keep totalfreqfoc;
if _N_=5;
totalfreqfoc=sumfreqfoc;
run;

data sumfoc2; set sumfoc combfoc;
run;

data sumfoc3;
if _N_=1 then set sumfoc; set combfoc;
prop=frequency/totalfreqfoc;
if _N_=1 then type='propfoc1';
else if _N_=2 then type='propfoc2';
else if _N_=3 then type='propfoc3';
else if _N_=4 then type='propfoc4';
else if _N_=5 then type='propfoc5';
run;

proc transpose data=proptrans out=transfoc; *only keeps the VAR
variable;
ID type;* this variable becomes new variable names,and the original
name disappears;
VAR prop; *this one variable becomes five new variables;
run;

data transfoc; set transfoc;
drop _NAME_;
try=1;
run;
*****************************************item mean calculation for the
ref group**************************;
data itemmean1; set Mantel2;
keep foc R17  _TABLE_ frequency;
run;
```

```
data itemmeanref; set itemmean1;
if FOC='0';
sumscore=R17*frequency;
run;

data itemmeanfoc; set itemmean1;
if FOC='1';
sumscore=R17*frequency;
run;

data tottrans; set combref;
if _N_=1 then type='totref1';
else if _N_=2 then type='totref2';
else if _N_=3 then type='totref3';
else if _N_=4 then type='totref4';
else if _N_=5 then type='totref5';
run;

proc transpose data=tottrans out=transm;
ID type;
VAR frequency;
run;

data transm; set transm;
drop _NAME_;
try=1;
run;

data tableref1; set itemmeanref; *keeping score info from interval 1
for the reference group;
if _TABLE_='1';
sum_sumscore1+sumscore;
try=1;
run;

data reffreq1; set tableref1;
keep sum_sumscore1 try;
if _N_=5;
run;

data tableref2; set itemmeanref; *doing the same for the other
intervals;
if _TABLE_='2';
sum_sumscore2+sumscore;
try=1;
run;

data reffreq2; set tableref2;
keep sum_sumscore2 try;
if _N_=5;
run;

data tableref3; set itemmeanref;
if _TABLE_='3';
sum_sumscore3+sumscore;
try=1;
run;
```

```
data reffreq3; set tableref3;
keep sum_sumscore3 try;
if _N_=5;
run;

data tableref4; set itemmeanref;
if _TABLE_='4';
sum_sumscore4+sumscore;
try=1;
run;

data reffreq4; set tableref4;
keep sum_sumscore4 try;
if _N_=5;
run;

data tableref5; set itemmeanref;
if _TABLE_='5';
sum_sumscore5+sumscore;
try=1;
run;

data reffreq5; set tableref5;
keep sum_sumscore5 try;
if _N_=5;
run;

data meanref; merge transm reffreq1 reffreq2 reffreq3 reffreq4
reffreq5; by try;
mref1=sum_sumscore1/totref1;
mref2=sum_sumscore2/totref2;
mref3=sum_sumscore3/totref3;
mref4=sum_sumscore4/totref4;
mref5=sum_sumscore5/totref5;
run;

data meanall; set meanref;
keep try mref1 mref2 mref3 mref4 mref5;
run;

data propmeanall; merge transref meanall; by try;
run;

data combreffoc; merge propmeanall transfoc; by try;
run;

data foctable; set itemmeanfoc; *summing all the item scores and
frequencies of the item scores;
keep R17 frequency sumscore sum_freq sum_sumscore;
if R17='.' then delete;
sum_freq+frequency;
sum_sumscore+sumscore;
run;

data foctable2; set foctable; *creating the total item mean score for
all the focal group intervals;
```

```
keep try totalmeanfoc sum_freq sum_sumscore;
if _N_=20;
totalmeanfoc=sum_sumscore/sum_freq;
try=1;
run;

data SMD; merge combreffoc foctable2; by try;
smd=totalmeanfoc-
((mref1*propfoc1)+(mref2*propfoc2)+(mref3*propfoc3)+(mref4*propfoc4)+(m
ref5*propfoc5));
run;

data SDtable; set equint2; *change to equint1(HD)and equint2(MI)for
each missing data method;
keep R17 foc;
run;

proc means data=sdtable noprint;
by foc;
var R17;
output out=R17mean;
run;

data refSTD; set R17mean;
keep try refSTD;
if _STAT_='STD' and Foc='0';
refSTD=R17;
try=1;
run;

data refN; set R17mean;
keep try refN;
if _STAT_='N' and Foc='0';
refN=R17;
try=1;
run;

data focSTD; set R17mean;
keep try focSTD;
if _STAT_='STD' and Foc='1';
focSTD=R17;
try=1;
run;

data focN; set R17mean;
keep try focN;
if _STAT_='N' and Foc='1';
focN=R17;
try=1;
run;

data smdes; merge refSTD refN focSTD focN smd; by try;
keep refSTD refN focSTD focN smd pooledSTD smdes;
pooledSTD=sqrt(((refN-1)*refSTD**2 + (focN-1)*focSTD**2)/(refN+focN-
2));
smdes=smd/pooledSTD;
run;
```

```
data onlysmdes17; set smdes; *dataset with SMD effect size value,
item#, missing method, and replication#;
keep item smdes miss rep;
item=17;
miss=3;
rep=&i;
run;

data both&imp; set onlysmdes4 onlysmdes17; *capturing each smd effect
size from each of the five imputations;
run;

proc logistic data=equint2; *OLR model 1;
model R4=TOT4/rsquare;
ods output Rsquare=Rsquare4_one;
run;

data meanrsq4_one; set rsquare4_one;
keep item rsquare1;
item=4;
rsquare1=cvalue1;
run;

proc logistic data = EQUINT2;
 model R4 =  TOT4 FOC/rsquare;
 ods output Rsquare=Rsquare4;
 ods output  ParameterEstimates = LOGIST4;

data meanrsq4; set rsquare4;
keep item rsquare2;
item=4;
rsquare2=cvalue1;
run;

DATA MILOG4; SET LOGIST4; KEEP WaldChiSq ITEM;
      IF VARIABLE ='Foc'; *leave the style of foc as 'Foc', all caps
did not work;
      ITEM=1;

data meanlog4_comb; merge meanrsq4 meanrsq4_one; by item;
keep item rsq miss rep;
rsq=rsquare2-rsquare1;
miss=3;
rep=&i;
run;

proc logistic data=equint2; *OLR model 1;
model R17=TOT17/rsquare;
ods output Rsquare=Rsquare17_one;
run;

data meanrsq17_one; set rsquare17_one;
keep item rsquare1;
item=17;
rsquare1=cvalue1;
run;
```

```
proc logistic data = EQUINT2;
  model R17 =  TOT17 FOC/rsquare;
  ods output Rsquare=Rsquare17;
  ods output  ParameterEstimates = LOGIST17;

data meanrsq17; set rsquare17;
keep item rsquare2;
item=17;
rsquare2=cvalue1;
run;

DATA MILOG17; SET LOGIST17; KEEP WaldChiSq ITEM;
      IF VARIABLE ='Foc';
      ITEM=2;

data meanlog17_comb; merge meanrsq17 meanrsq17_one; by item;
keep item rsq miss rep;
rsq=rsquare2-rsquare1;
miss=3;
rep=&i;
run;

data combrsq&imp; set meanlog4_comb meanlog17_comb; *capturing each of
the rsquare effect size from each imputation;
run;

DATA MILOG; SET MILOG4 MILOG17; keep item CHI&IMP difmethod;
 CHI&IMP=WALDCHISQ;
      DIFmethod=2;

data meandif&IMP; set allMI MIlog; RUN;
%END;

DATA COMB; MERGE MEANDIF1 MEANDIF2 MEANDIF3 MEANDIF4 MEANDIF5;

data bothsmd; set both1 both2 both3 both4 both5; *combining all five
smd effect sizes;
run;

proc sort data=bothsmd out=sortsmdes; by item;
run;

proc means data=sortsmdes noprint mean; *averaging the five effect
sizes;
by item;
var smdes;
output out=avgallsmdes;

data onlyavgsmd4; set avgallsmdes;
keep item smdes miss;
if _STAT_='MEAN' and item='4';
miss=3;
run;

data onlyavgsmd17; set avgallsmdes;
keep item smdes miss;
```

```
if _STAT_='MEAN' and item='17';
miss=3;
run;

data onlymismd; set onlyavgsmd4 onlyavgsmd17;
miss=3;
rep=&i;
run;


data bothrsq; set combrsq1 combrsq2 combrsq3 combrsq4 combrsq5;
*combining all five rsquare effect sizes;
run;

proc sort data=bothrsq out=sortrsq; by item;
run;

proc means data=sortrsq noprint mean; *averaging the five effect sizes;
by item;
var rsq;
output out=avgallrsq;

data onlyavgrsq4; set avgallrsq;
keep item rsq miss;
if _STAT_='MEAN' and item='4';
miss=3;
run;

data onlyavgrsq17; set avgallrsq;
keep item rsq miss;
if _STAT_='MEAN' and item='17';
miss=3;
run;

data onlymirsq; set onlyavgrsq4 onlyavgrsq17;
miss=3;
rep=&i;
run;


proc iml;
DO METHOD=1 TO 2;
DO ITEMP=1 TO 2;
      use COMB; read all var {CHI1 CHI2 CHI3 CHI4 CHI5}
      WHERE (DIFMETHOD=METHOD & ITEM=ITEMP) into G2;
      DF=1;
  m=ncol(g2);
  g=sqrt(g2);
  mg2=sum(g2)/m;
  r=(1+1/m)*(ssq(g)-(sum(g)**2)/m)/(m-1);
  f=(mg2/df - (m + 1)/(m - 1)*r)/(1+r);
    if f < 0 then f=0;
  ddf=(m-1)*(1+1/r)**2/df**(3/m);
  p=1-probf(f,df,ddf);
      IF METHOD=1 & ITEMP=1 THEN PP=p;
      ELSE PP=PP//p;
END;
```

```
END;
      create FINALPS from PP;
      append from PP;
  quit;

DATA MIDIF; SET FINALPS; KEEP PROB ITEM DIFMETHOD MISS REP;
  PROB=COL1;
  IF _N_=1 THEN DO;
  ITEM=4; DIFMETHOD=1;
  END;
 IF _N_=2 THEN DO;
  ITEM=17; DIFMETHOD=1;
  END;
 IF _N_=3 THEN DO;
  ITEM=4; DIFMETHOD=2;
  END;
 IF _N_=4 THEN DO;
  ITEM=17; DIFMETHOD=2;
  END;
  MISS=3;
  REP=&I;
run;

DATA ALL; SET MEANDIF MIDIF;
permiss=&&miss&permiss;
run;

data all2; set smd_1 onlymismd;
permiss=&&miss&permiss;
run;

data all3; set rsq_1 onlymirsq;
permiss=&&miss&permiss;
run;

PROC APPEND BASE=FINALdiss DATA=ALL; *grabs each set of information for
each rep;
      RUN;

proc append base=finalall2 data=all2;
run;

proc append base=finalall3 data=all3;
run;

%end; *for the rep do loop;
%end; *for the permiss do loop;

DATA RESULTSdiss; SET FINALdiss;
      IF PROB NE . THEN DO;
      IF PROB LT .05 THEN DIF=1;
      ELSE DIF=0; END;
PROC SORT; BY permiss DIFMETHOD MISS ITEM;

ods output "One-Way Frequencies" (MATCH_ALL PERSIST=PROC)=totfreq1;
PROC FREQ; TABLES DIF; BY permiss DIFMETHOD MISS ITEM; RUN;
ods output close;
```

```
*be sure to change this per the number of conditions running at a given
time;
DATA FINALRES&&ratio&ratsize; set totfreq1 totfreq2 totfreq3 totfreq4
totfreq5 totfreq6 totfreq7 totfreq8 totfreq9 totfreq10 totfreq11
totfreq12 totfreq13 totfreq14 totfreq15 totfreq16 totfreq17 totfreq18
totfreq19 totfreq20 totfreq21 totfreq22 totfreq23 totfreq24;

KEEP PERCENT DIFMETHOD ITEM MISS permiss;
IF DIF=1;
RUN;

proc sort data=finalall2 out=sortsmdes2; by item miss permiss; *sorting
the final smd effect sizes from each rep;
run;

data cond1; set sortsmdes2;*separating the data by item and missing
data method;
if item ='4' and miss='1';
run;

data cond2; set sortsmdes2;
if item ='4' and miss='2';
run;

data cond3; set sortsmdes2;
if item ='4' and miss='3';
run;


proc means data=cond1 noprint mean; *averaging the 1000 effect sizes
for each item by each condition (10, 25, 40 percent missing data) ;
by permiss;
var smdes;
output out=avgallsmdes2;

proc means data=cond2 noprint mean;
by permiss;
var smdes;
output out=avgallsmdes2_2;

proc means data=cond3 noprint mean;
by permiss;
var smdes;
output out=avgallsmdes2_3;

data onlyavgsmd4_1; set avgallsmdes2;
keep permiss item miss smdes;
if _STAT_='MEAN';
item=4;
miss=1;
run;

data onlyavgsmd4_2; set avgallsmdes2_2;
keep permiss item miss smdes;
```

```
if _STAT_='MEAN';
item=4;
miss=2;
run;

data onlyavgsmd4_3; set avgallsmdes2_3;
keep permiss item miss smdes;
if _STAT_='MEAN';
item=4;
miss=3;
run;

data cond4; set sortsmdes2;
if item ='17' and miss='1';
run;

data cond5; set sortsmdes2;
if item ='17' and miss='2';
run;

data cond6; set sortsmdes2;
if item ='17' and miss='3';
run;

proc means data=cond4 noprint mean;
by permiss;
var smdes;
output out=avgallsmdes2_4;

proc means data=cond5 noprint mean;
by permiss;
var smdes;
output out=avgallsmdes2_5;

proc means data=cond6 noprint mean;
by permiss;
var smdes;
output out=avgallsmdes2_6;

data onlyavgsmd17_4; set avgallsmdes2_4;
keep permiss item miss smdes;
if _STAT_='MEAN';
item= 17;
miss= 1;
run;

data onlyavgsmd17_5; set avgallsmdes2_5;
keep permiss item miss smdes;
if _STAT_='MEAN';
item= 17;
miss= 2;
run;

data onlyavgsmd17_6; set avgallsmdes2_6;
keep permiss item miss smdes;
if _STAT_='MEAN';
item= 17;
```

```
miss= 3;
run;

data sasuser.finalsmdes&&ratio&ratsize; set onlyavgsmd4_1 onlyavgsmd4_3
onlyavgsmd17_4 onlyavgsmd17_6; *final averaged smd effect size;
run;


proc sort data=finalall3 out=sortrsq2; by item miss permiss; *doing the
same for rsquare;
run;

data cnd1; set sortrsq2;
if item ='4' and miss='1';
run;

data cnd2; set sortrsq2;
if item ='4' and miss='2';
run;

data cnd3; set sortrsq2;
if item ='4' and miss='3';
run;

proc means data=cnd1 noprint mean;
by permiss;
var rsq;
output out=avgallrsq2;

proc means data=cnd2 noprint mean;
by permiss;
var rsq;
output out=avgallrsq2_2;

proc means data=cnd3 noprint mean;
by permiss;
var rsq;
output out=avgallrsq2_3;

data onlyavgrsq4c_1; set avgallrsq2;
keep permiss item miss rsq;
if _STAT_='MEAN';
item=4;
miss=1;
run;

data onlyavgrsq4c_2; set avgallrsq2_2;
keep permiss item miss rsq;
if _STAT_='MEAN';
item=4;
miss=2;
run;

data onlyavgrsq4c_3; set avgallrsq2_3;
keep permiss item miss rsq;
if _STAT_='MEAN';
item=4;
```

```
miss=3;
run;

data cnd4; set sortrsq2;
if item ='17' and miss='1';
run;

data cnd5; set sortrsq2;
if item ='17' and miss='2';
run;

data cnd6; set sortrsq2;
if item ='17' and miss='3';
run;

proc means data=cnd4 noprint mean;
by permiss;
var rsq;
output out=avgallrsq2_4;

proc means data=cnd5 noprint mean;
by permiss;
var rsq;
output out=avgallrsq2_5;

proc means data=cnd6 noprint mean;
by permiss;
var rsq;
output out=avgallrsq2_6;

data onlyavgrsq17c_4; set avgallrsq2_4;
keep permiss item miss rsq;
if _STAT_='MEAN';
item=17;
miss=1;
run;

data onlyavgrsq17c_5; set avgallrsq2_5;
keep permiss item miss rsq;
if _STAT_='MEAN';
item=17;
miss=2;
run;

data onlyavgrsq17c_6; set avgallrsq2_6;
keep permiss item miss rsq;
if _STAT_='MEAN';
item=17;
miss=3;
run;

data sasuser.finalrsq&&ratio&ratsize; set onlyavgrsq4c_1 onlyavgrsq4c_3
onlyavgrsq17c_4 onlyavgrsq17c_6; /*final averaged rsquare effect size
values for this condition*/;
run;

%end; *for the ratsize(ratio size)do loop;
```

```
%mend ss;

%ss;
```