11-16-2015

An Empirical Comparison of the Effect of Missing Data on Type I Error and Statistical Power of the Likelihood Ratio Test for Differential Item Functioning: An Item Response Theory Approach using the Graded Response Model

Patricia Rodriguez De Gil
University of South Florida, prodriguezdegil@verizon.net

An Empirical Comparison of the Effect of Missing Data on Type I Error and Statistical Power of the Likelihood Ratio Test for Differential Item Functioning: An Item Response Theory Approach using the Graded Response Model

by

Patricia Rodríguez de Gil

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
in Curriculum and Instruction with an emphasis in
Measurement and Evaluation and Secondary Social Science Education
Department of Educational and Psychological Studies
Department of Teaching and Learning
College of Education
University of South Florida

Co-Major Professor: Jeffrey D. Kromrey, Ph.D.
Co-Major Professor: Bárbara C. Cruz, Ed.D.
Eun Sook Kim, Ph.D.
James A. Duplass, Ph.D.

Date of Approval:
November 10, 2015

Keywords: Validity, Invariance, Civics education, Attitude assessment, Polytomous items

DEDICATION

To my children, with great love.

Adelina Patti

Ismael Everardo

ACKNOWLEDGMENTS

"Pa' riba and pa' lante!"

Alfred North Whitehead said, "No one who achieves success does so without acknowledging the help of others. The wise and confident acknowledges this help with gratitude". Never this statement has been more genuine than in this occasion in which I take the opportunity to express my great appreciation and love for those who stood by my side all along this journey. First and foremost, my utmost appreciation goes to my husband, Ismael Gil and to my children. My husband has always been my source of strength; I would not have completed my degree without his constant help and support. My children have been always my greatest blessing and my motivation for continuing my studies. I also want to thank my mother, Elpidia. She had a clear vision of the relevance that having a career has for a woman and worked tirelessly for providing her daughters with the means to purse one. My mother always instilled and nurtured my desire to pursue a career. I thank her for all the sacrifices she made so I could go to school and earn a degree. I also want to express my appreciation to Dr. Bárbara Cruz. I honestly believe that I would not be celebrating the completion of my degree without her. Over my years as a student, Dr. Cruz has been literally my most enthusiastic, supportive, and caring professor, and I thank her for inspiring me every day. On the same note, Dr. Jeffrey Kromrey has been my advisor since I started my graduate work and I am very appreciative of all his help and guidance for completing my dissertation work. Both Dr. Cruz and Dr. Kromrey as my co-major professors, they were always willing to read my dissertation drafts and provide their valuable feedback. I also want to thank the faculty of the Social Studies program at USF. The experiences that each professor in the program provided me with made me a better student and a better teacher. I am in debt to Dr. Michael Berson, who always trusted that I could do it. Working under the supervision of Dr. .James Duplass not only helped me develop teaching skills but also importantly, provided me with the collegiate association that I had yearned for so long. I will never forget the thoughtfulness of Dr. Howard Johnston. It was very meaningful to me that always, before starting any conversation, he always asked about my family. Over the years, I have had the fortune of having excellent professors who in a great way molded the person I am today. I would like to acknowledge specially the members of my dissertation committee. I thank them greatly for their sharing with me their expertise and for their guidance completing my dissertation.

My journey, not only academic but in life, would not be the same without my sister Maria del Rosario, my best friend. I truly believe that she is my God's sent gift; her unconditional love has accompanied me always. She embodies what having a sister is meant. I also feel fortunate for having such a great friends and colleagues. Thanh Pham and Diep Nguyen have been always there, sharing and cheering. I thank them for being always supportive and for their encouragement. I look forward to continue having them as colleagues and friends.

Lastly, I thank the Lord, for His many gifts, for His always being with me, for protecting me. To God be the glory and honor. Amen.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In the context of educational research, missing data arise when examinees omit or do not reach an item, which generates an item nonresponse problem. Using a simulation approach, in addition to conducting complete data analyses, this study compared the performance of six methods for treating item nonresponse in the context of differential item functioning (DIF). The effect of missing data on the Type I error and statistical power of the Likelihood Ratio test for DIF detection in small scales was examined in the context of Item Response Theory (IRT-LR), using polytomous, Likert-type data and the graded response model. The effect of ability distribution, sample size, number of items, proportion of missing observations, and proportion of missing items on Type I error rates and empirical power of the IRT-LR DIF test were examined under full information maximum likelihood (FIML), multiple imputation (MI), person mean substitution (PMS), single regression substitution (SRS), relative mean substitution (RMS), and Listwise deletion missing data methods. Type I error rates were very consistent across nominal levels and factors, under each missing data method. Among the missing data methods examined, the FIML and PMS methods had Type I error rates comparable to the rejection rates for complete data. Although MI is considered a "state-of-the-art" missing data method, in this study, MI, as well as SRS were the less effective missing data methods (i.e., both MI and SRS had inflated rejections rates across all conditions). On the same note, Listwise deletion has been described as one of the most ineffective methods; however, under large data, the data loss due to implementing Listwise deletion might not be a problem if in addition other conditions are present, such as a small proportions of missing observations and small number of items or variables. Along with complete data and FIML, the PMS method had an adequate Type I error control under both nominal levels examined. MI and SRS had the smallest proportions of conditions meeting Bradley's criteria for robustness at both levels of significance examined; as a result, when alpha was .01 none of the simulation conditions of these methods met the criteria for robustness and were not included in power analyses at this significance level. Power analyses were entirely consistent across nominal levels, factors and missing data methods. Entirely consistent with theory, sample size and proportion of missing observations were the factors affecting the performance of the IRT-LR test for DIF detection across all missing data methods.

CHAPTER ONE

INTRODUCTION

> "Providing information to test takers and test score users about the abilities of test takers at different score levels has been a persistent problem in educational and psychological measurement." — Sinharay, Haberman, and Lee, 2011, p. 61

Overview

The measurement of individuals' traits, or mental properties such as abilities and attitudes, has been a long-lasting quest that dates back to 1882 with Galton's pioneering work developing rating scales and questionnaires, and Thorndike's contributions to psychometric theory and its application to educational measurement (Ward, Stoker, & Murray-Ward, 1996). This quest continues today (Sijtsma & Junker, 2006). But why do we measure individuals' traits?

Currently, the measurement of students' academic achievement has a prominent position in the No Child Left Behind (NCLB) Act of 2001 (NCLB, 2002; US Department of Education, 2002), influencing not only classroom practices but also testing at state and national levels. For example, in agreement with Tyler's (1951) ideas on the influence of educational measurement in the improvement of instruction, Carey (2001) stated that measurement in educational settings serves several purposes, namely, planning, monitoring, and evaluating instruction. Moreover, achievement data influence educational decision making. That is, in addition to improving teaching and learning, the information that test scores provide greatly impacts the classification, selection, placement, and promotion of test takers (Clauser & Mazor, 1998; Garcia & Pearson, 1994). Therefore, empirical evidence should support the validity of inferences from test scores.

At the core of assessment-driven educational reforms such as the NCLB (2002) is the development of methods for eliminating nonrandom, systematic errors in measurement that arise when students with the same ability or trait but from different groups (e.g., male, female; minority, nonminority) do not have the same probability of answering correctly or endorsing a test item, after the item has been conditioned on ability or trait level (Balsis, Gleason, Woods & Oltmanns, 2007; Embretson & Reise, 2000). Because precision of measurement is required so that it allows for valid interpretations of test scores (Cronbach & Gleser, 1965; Kane, 1996), the focus of extensive research has been the development and improvement of item and test evaluation procedures that ensure the accurate measurement of students' ability and traits and consequently, the validity of the interpretation of test scores (Robitzsch & Rupp, 2009). One of these item evaluation procedures is differential item functioning (DIF), which "has become an essential aspect of the validation of test score interpretations" (Ankenmann, Witt, & Dunbar, 1999, p. 278). The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association & National Council of Measurement in Education, 1999), hereafter the *Standards,* states that:

> When credible research reports that differential item functioning exists across age, gender, racial / ethnic, cultural, disability, or linguistic groups in the population of test takers in the common domain measured by the test, test developers should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias the test score for particular groups. (p. 81)

That is, when the members of a subgroup of students taking a test do not have the same probability of responding correctly to or endorsing an item as the members of another subgroup of students with the same ability or trait, we say that DIF is present. DIF suggests that the internal structure of tests items (i.e., item parameters' properties or characteristics such as item discrimination and item difficulty) is not the same for different groups matched on ability (Woods, 2008).

Regardless of this extensive research on the improvement of item and test evaluation procedures, empirical evidence of the disparities in test performance across subgroups of students continue to be a concern well documented in the literature, e.g., gender differences in language achievement (Mullis, Martin, Gonzalez, & Kennedy, 2003), gender differences in mathematics and science achievement (Mullis, Martin, Gonzalez, & Chrostowski, 2004), and gender differences in civic knowledge (Baldi, Perie, Skidmore, Greenberg, & Hahn, 2001).

This increased interest in subgroups' differences in test scores has resulted in the development of theories that pose the responsibility of the lower test scores of minority students on external factors to the tests (Hulin, Drasgow, & Komocar, 1982). Such attempts to explain the lower mean test scores of minorities students, for example, rule out differences between groups *a priori* (Thorndike, 1971), or pose the responsibility of lower test scores on the minority groups themselves, within their genetic heritage (Jensen, 1969) and home environments (McPhee, Kreutzer, & Fritz, 1994). Other explanations focus primarily on the role of society and schools (Coleman, 1966), and on the range of variables such as racial discrimination, prejudice, and stereotype that can stigmatize and contribute to the alienation of minority students. However, these external factors are not necessarily evidence of DIF.

There are several statistical methods for identifying differentially functioning test items, each with its strengths and limitations. Some methods are based on classical test theory (CTT) and other methods are based on item response theory (IRT) and the decision on which framework and procedure are to be implemented, should be taken within the theoretical and empirical specifics of each research situation. The application of CTT and IRT models for dichotomously scored items for the measurement of student achievement and the evaluation of DIF has dominated since early educational testing (Camilli & Shepard, 1994; Clauser & Mazor, 1998; De Ayala & Sava-Bolesta, 1999; Kim, Cohen, Alagoz, & Kim, 2007; Lane, Stone, Ankenmann, & Liu, 1995). However, it is important to consider also noncognitive assessments because they "influence, in either facilitative or debilitative ways, both student learning and test performance" (Messick, 1984, p. 215). Attitude measurement, for instance, has been a cornerstone in empirical research (Thissen, Steinberg, Pyszcznski, & Greenberg; 1983) and evidence of validity is equally important for such measures. Furthermore, the current popularity of performance assessments has increased the use of polytomous IRT models e.g., partial credit model (PCM; Masters, 1982), generalized partial credit model (GPCM; Muraki, 1992), and graded response model (GRM; Samejima, 1969, 2010).

Ankenmann et al. (1999) stated that "the detection of differential item functioning (DIF) in polytomously scored, constructed-response items that constitute most performance assessments has become an essential aspect of the validation of test scores interpretations" (p. 278). However, the use of polytomously scored items does not preclude the potential for differentially functioning items. In addition, the detection of DIF might be complicated by the presence of a pervasive problem in empirical research, that of missing data. In educational testing, missing data occur when a student either does not respond to an item or question (i.e., item nonresponse) or does not respond to any question at all (i.e., unit nonresponse). That is, data are missing for some test items, and / or for some students.

When students do not answer items in a test because they do not know the answer, do not have time to respond to all questions, or omit the questions they are not comfortable with (such as in the case of attitudinal measurement), the item nonresponse generates a missing data problem (e.g., the variable of interest and the omitted response are not independent) which cannot be ignored (i.e., leaving data untreated, doing nothing about it). Because the statistical methods used to analyze item responses so that items can be evaluated for DIF might not be robust to missing data (e.g., failure to converge; Drasgow, Levine, Tsien, Williams, & Mead, 1995), data should be treated by applying missing data methods (MDM) that impute plausible values and replace missing data. Then, analyses can be conducted on complete data using standard statistical methods for evaluating DIF.

The crucial question is then, should we care about item nonresponse while doing a DIF analysis? The answer is yes if there is the risk of potential statistical bias associated with valid inferences of test scores and their use.

Statement of the Problem

The development of test validation procedures has led to the study of DIF. However, DIF analyses are potentially subject to spurious interpretations due to the presence of missing data. Depending on the missingness mechanism (i.e., types of missing data), the magnitude of the missing data (i.e., percentage of missing responses for a person and by item), sample size (i.e., number of students taking the test), number of test items (i.e., test length), and the magnitude of DIF (i.e., negligible, moderate, and high), the MDM used to treat the existing missing data due to item nonresponse might result in data handling complications such as 1) reduced sample size (Davey, Savla & Luo, 2005; O'Rouke, 2003; Zhang, 2003), 2) reduction of analytical power (Yenduri & Iyengar, 1994), and 3) seriously statistically biased results due to the systematic differences between the observed and non-observed data (Little, 1988). In addition to the extent to which item nonresponse might impact the accuracy and precision of the point estimates (i.e., item difficulty and item discrimination parameter estimation), the implementation of a MDM might also impact the performance of the methods used to detect DIF items.

Purpose of the Study

The validity of test score interpretations is closely tied to how the test scores are used (e.g., selection of students). Thus, it is important to state the scope or focus of this study, which is on the internal properties of test items and how these properties might be impacted by the presence of missing data.

Missing data have been broadly explored in the context of statistical methodology such as structural equation modeling (SEM; Gold & Bentler, 2000), and multiple regression (Brockmeier, Kromrey, & Hogarty, 2003; Kromrey & Hines, 1994). However, there is relatively less research on the effect of missing data on DIF analysis for polytomous data within the IRT framework, compared to research on achievement and binary data. Because much less is known about noncognitive tests such as attitude measurement and because both item nonresponse and MDM play an important role in the performance of a given DIF detection method, the purpose of this study was, within the context of IRT, to empirically compare the effects of six MDM (a maximum likelihood based method, single regression substitution (SRS), relative mean substitution (RMS), person mean substitution (PMS), multiple imputation(MI), and Listwise deletion) on the Type I error rates and statistical power of the Likelihood Ratio (IRT-LR) test for DIF detection in attitude measurement, using the graded response model (GRM) for polytomous items.

Research Questions

1. What is the effect of missing data (i.e., item nonresponse) and their treatment on the Type I error rate of the Likelihood Ratio test for Differential Item Functioning detection?
   a. To what extent is the effect consistent across levels of significance?
   b. To what extent is the effect consistent across MDM?
      i. To what extent is the effect consistent across sample size?
      ii. To what extent is the effect consistent across percentage of missing data by persons and items?
      iii. To what extent is the effect consistent across the magnitude of DIF?
      iv. To what extent is the effect consistent across population distributions?
2. What is the effect of missing data and their treatments on the statistical power of the Likelihood Ratio test for Differential Item Functioning detection?
   a. To what extent is the effect consistent across levels of significance?
   b. To what extent is the effect consistent across MDM?
      i. To what extent is the effect consistent across sample size?
      ii. To what extent is the effect consistent across percentage of missing data by persons and items?
      iii. To what extent is the effect consistent across the magnitude of DIF?
      iv. To what extent is the effect consistent across population distributions?

Overview of the Study Design

The research questions were addressed using a simulation approach in which a crossed factorial design was used to investigate the effect that missing data or item nonresponse and MDM had on the effectiveness of the IRT-LR test for detecting DIF in the GRM, in terms of Type I error and statistical power. The factors manipulated in the simulation study included population distributions for the reference and focal groups $\theta_R \sim N(0,1) : \theta_F \sim N(0,1)$, and $\theta_R \sim N(0,1) : \theta_F \sim \mathrm{N}(-.5,1)$, total sample size ($N_1$=500 and N_2=1000) in the following ratios: 1:1, and 3:2 for the reference and focal groups respectively (n_R=250 : n_F=250, n_R=500 : n_F=500, and n_R=300 : n_F=200, n_R=600 : n_F=400), number of items or test length (4-, 5-, and 6-items scales, with 4 response Likert-type categories in each item), proportion of missing observations or persons (10%, 20%, 40%) , proportion of missing items (~20% and ~40% or 1 and 2 items respectively), and magnitude of DIF (.25, .50 and .75). Nominal levels of alpha for the test of the null hypothesis were .01 and .05. For each combination of conditions, 1000 samples or number of replications (*n*R) were generated. The use of 1000 samples or replications provided a maximum standard error of .015 and a 95% confidence interval width ± .03 around observed rates of Type I error (Robey & Barcikowski, 1992). The proposed MDM were applied and their effect on the Type I error rates and statistical power of the IRT-LR test for DIF were estimated. Bradley's liberal criterion (1978) was used for the evaluation of robustness of Type I error control at α = .01 and α = .05; power analyses were conducted for those conditions evidencing adequate Type I error control (Ankenmann et al., 1999). In addition, complete data analyses were conducted for comparison purposes.

Data Source: The Civics Education Study of 1999

Data for this study was generated using item parameters' values estimated from three subscales included in the survey section of the Civics Education Study of 1999 (U.S. Department of Education, National Center for Education Statistics, 1999). Considering the broad range of restricted data and publicly available data, why selecting data from the social studies framework?

To achieve quality education in American schools, a better understanding of how classroom instruction work is needed (Stodolsky, 1988). Social studies as a school subject matters; Yeager and Davis (2005) stated, for example, that "social studies is a potentially powerful, engaging, and relevant curriculum area" (p. 2). However, as federal and state-mandated assessment has elevated the status of mathematics, sciences, and reading literacy core subjects, the study of the social sciences, also a core subject, has been relegated as an "enrichment" subject matter with a limited allocation of instructional time in elementary level (Brophy & VanSlendright, 1997), and reduced to the irrelevant teaching of facts in middle and secondary levels (Vogler & Virtue, 2007). Has placing the social studies in the "back burner" (Vogler, Lintner, Lipscomb, Knopf, Heafner & Rock, 2007) made middle school students passive recipients of current global realities? What do young people think about democracy? Do they understand how democratic institutions work? Do they expect to vote and to take part in other civic activities as adults? These are the questions that motivated the Civics Education study of 1999 and if we are to understand how schools and social studies classroom instruction prepares our young students for participating in our democratic institutions, promoting civic knowledge, attitudes, and involvement, it is important that the instruments used to measure students' attitudes toward those institutions are valid. This is basically what the subscales selected for this study address and what make them worth to study.

Thus, this study was conducted using item parameters estimated from three subscales of the Civics Education Study measuring students' degree of adherence to common values and attitudes toward women's rights, immigration, and political activism. In addition to the advantage of conducting the simulation study by generating data that emulate the conditions under study, using real data provided the additional advantage of carrying the study under normally occurring violation of assumptions (e. g., normality of distributions; Kromrey and Hines, 1994).As this international study was conducted in 28 countries; the item parameters for this study were estimated from the United States public data sample for the standard population (9th grade students; *N*=2811). The study was conducted using SAS 9.4 and the IRTGEN SAS macro (Whittaker, Fitzpatrick, Williams, & Dodd, 2003) was used to generate data to conform the GRM. The item parameters were calibrated using the marginal maximum likelihood estimation method, as implemented in MULTILOG 7.03 (Thissen, 2003).

Significance of Research

Previous to the legislation of the NCLB (2001), any mandated testing did not have any consequences for poorly performing students and schools. But with the implementation of the NCLB, not only testing dramatically increased (e.g., the NCLB mandates the administration of 17 tests (personal communication with Dr. Bárbara Cruz, May 4, 2015)) but also, testing steered toward accountability for those not meeting the goals of the NCLB. How students and schools are held accountable is observed in the mounting role of standardized testing (e.g., American students are tested far more than students in other countries) and in the sanctioning of low performing schools. However, as John Merrow (2001) asserted, tests are not evil. Andrich (2002) said that, "Assessment should be valid, educative, explicit, fair, and comprehensive" (p. 105).

Thus, the quality of measurement instruments has drawn the attention not only of test developers but also of scholars from many disciplines, policy makers, and administrators (Cizek, 2012). However, the interpretation of test scores must be placed in the overall test development procedure, not just on the total score. As Linn (1990) stated, “the most important question regarding any measure concerns the validity of the uses and interpretation of the scores” (p. 115). Despite the continued development of test validation procedures (e.g., DIF studies), the presence of systematic measurement errors (measurement bias) that arise when factors other than the underlying construct are measured is a threat to the validity of the inferences from test scores (Zumbo, 1999). The relevance of validity in test score interpretations is captured in its references as being “one of the major deities in the pantheon of the psychometrician” (Ebel, 1961, p. 640) and as “the foundation for virtually all of our measurement work” (Frisbie, 2005, p. 21). Yet, as Cizek (2012) stated, “all is not well with validity” (p. 31). DIF continues to be a threat to validity and while the psychometric basis of tests has changed dramatically (Embretson & Reise, 2000; Hambleton & Slater, 1997; Linn, 1990) and several school reforms have been implemented in response to these problems, missing data and DIF are ubiquitous in empirical research and both pose a serious threat to the fairness in test use and validity of the interpretation of test scores.

Su and Wang (2005) stated, “The detection of differential item functioning (DIF) in polytomous items has attracted much attention in recent years” (p. 313). However, most IRT work has been based on dichotomous models (De Ayala & Sava-Bolesta, 1999) and currently, there is still a predominance of achievement testing using binary item formats. Testing relying on dichotomously scored items can be a disadvantage (Dodd, Koch, De Ayala; 1989); thus, a better option is to use a model that assesses information across all item categories (De Ayala & Sava-Bolesta, 1999) as is the case in polytomous models.

Because of the dominance of achievement testing and the application of objective tests, much less is known about other types of measurement using noncognitive data, as is the case of attitude assessment that is also part of the process through which students construct knowledge and develop abilities. Polytomous data, collected in the form of graded responses can be used to address and provide insights on attitudinal aspects that affect student academic achievement. Thus, evidence of the validity of these measures is equally important (Ankenmann et al., 1999).

Limitations and Summary

A limitation of simulated data studies is that the data generation process and methods might apply only to the conditions under study (e.g., data generation methods might favor the IRT method used for bias detection). Also, the study will use self-reported measures of attitudes which can be problematic due to, for example, providing a socially desirable response.

In this chapter, the purpose of the study was introduced along with the problem in making valid inferences from test scores, and the problems that missing data generate in the detection of differentially functioning test items. Issues of differential item functioning in measurement were introduced and its impact in students' selection and classification was discussed. The implementation of a simulation study was elaborated and the use of item response theory (IRT) and specifically, the GRM, was justified. The next chapter provides a literature review on missing data in Likert-type scales, and on IRT and its implementation in the detection of differentially functioning items or DIF. In addition, the analytical plan was developed in chapter III.

Definition of Terms

Ability: It refers to the ability or trait being measured and within item response theory is represented by the Greek small letter Theta (θ). Thus, the ability for examinee *i*, is represented as θ_i. In educational research, the value of ability or trait is assumed to be unknown and hence, estimated (Harwell, Baker, & Zwarts, 1988). Although the terms are not synonymous, ability can also be referred or used interchangeably as *proficiency* (Osterlind & Everson, 2009).

Accuracy: Degree of closeness of measurements of a quantity to the actual or true value. In this context, *Bias* is a statistical index of the accuracy of measurement (Mellenbergh, 1989).

Bias: It refers to statistical bias or the difference between the average value of the estimated parameter across simulation replications and its true value (DeMars, 2003; Stone, 1992; Wang & Chen, 2005).

Construct (psychological): Postulated attribute of people, assumed to be reflected in test performance (Cronbach & Meehl, 1955). Thus, a construct is an unobserved, latent variable underlying behavior and "imperfectly measured by a test or questionnaire" (Embretson & Reise, 2000; Schafer & Graham, 2002).

Dichotomous item: An item is a dichotomous item if it is scored with two response categories such as yes/no, correct/incorrect, or agree/disagree (Cohen, Kim, & Baker, 1993; Clauser & Mazor; 1998).

Differential Item Functioning (DIF): Differences in the functioning of an item among groups that are matched on the attribute measured by the item (Clauser & Mazor, 1998; Cohen et al., 1993; Paek & Guo, 2011). When a test item favors one group over another, the item exhibits DIF.

Spurious DIF: Identification of DIF in an item due to the method used. Andrich and Hagquist (2012) also termed this type of DIF as *Artificial DIF*.

Unbiased item: Item for which the probability of a correct response is the same for all persons of a given ability, regardless of their ethnic, cultural, sex, or group membership (Cohen et al., 1993). On the other hand, a *biased item or item bias* is one that unfairly favors one group over another (Clauser & Mazor, 1998). In IRT, an item is considered biased "when it differs in difficulty between subjects of identical ability from different groups" (Mellenbergh, 1989, p. 128).

Uniform DIF: When examinees of a subgroup taking a test consistently have higher probability of answering correctly an item, we say that the item presents uniform (i.e., constant difference across all levels of ability measured by the test) differential functioning (Mellenbergh, 1982).

Item difficulty (β)*:* Item technical property or descriptor. In CTT, item difficulty is the proportion of examinees of the total group that responded correctly to an item. In IRT and binary items, item difficulty or location parameter, specifies the point in the ability scale at which the probability of an examinee of responding correctly or selecting an item response is .50. Because β indicates where an item functions on the ability scale, within IRT it is a location index (Baker, 1977; Embretson and Reise, 2000).

Item discrimination (α): Item technical property or descriptor. Determines how well the item differentiates between examinees whose ability is below the item location and those having ability above the item location.

Item impact: Refers to the differences in the performance of groups of examinees on specific items as a result of actual or "real" differences in the groups' ability to respond to the item; that is, it is the true difference in performance of groups of examinees with different abilities on specific items (Clauser & Mazor; 1998; Robitzsch & Rupp, 2009).

No DIF item: The expectation for valid test score comparisons is that items' structural properties are the same among test takers having the same standing on the trait being measured. Thus, a No DIF item is one that is invariant across groups so that "the expected value of a response to the item from persons from different identifiable groups is the same" (Andrich & Hagquits, 2012, p. 387).

False Negative (FN): Failure of detection in the presence of the condition being tested for (also known as Type II error or β, the probability of failing to reject the null hypothesis when the null hypothesis is false). In DIF studies, identifying an item as being free of DIF (NO DIF item) when the item is a DIF item (Andrich & Hagquits, 2012).

True Negative (TN): Failure of detection in the absence of the condition being tested for (i.e., failure to reject the null hypothesis when the null hypothesis is true). In DIF studies, identifying a DIF free item as a DIF free (Andrich & Hagquits, 2012).

False Positive (FP): Incorrect detection in the absence of the condition being tested for (also known as Type I error or α, the probability of rejecting the null hypothesis when the null hypothesis is true). In DIF studies, FP means identifying an item as a DIF item when the item does not show DIF (Andrich & Hagquits, 2012).

True Positive (TP): Correct detection in the presence of the condition being tested for (also known as power; rejection of the null hypothesis when the null is false). In DIF studies, a DIF item is correctly identified as a DIF item (Andrich & Hagquits, 2012).

Trait: Messick (1989) defines a trait as "a relative stable characteristic of a person—an attribute, enduring process, or disposition—which is consistently manifested to some degree when relevant, despite considerable variation in the range of settings and circumstances" (p. 15).

Likert-type scale: psychometric scale commonly involved in research that employs questionnaires. It is the most widely used approach to scaling responses in survey research.

Likelihood Ratio test (LR): Statistical test used to compare the fit of two models, one of which (the *null model*) is a special case of the other model (the *alternative model*). The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than for the other model.

Measurement: Procedure with which a number is assigned to an object of measurement to represent the value of some attribute for that object of measure (Kane, 1996).

p: The proportion of correct responses to the total number of responses of people scoring within that range. At the item level, Fan (1998) defines the *p* as the index for the item difficulty, with a higher value indicating an easier item; that is, *p* is the success rate of examinees on an item (assuming that it is scored dichotomously).

Parameter: In the IRT context, it refers to both population item and person parameters within a specific IRT model, whose values are estimated with a random sampling design (Harwell, Baker, and Zwarts, 1988).

Parameter recovery: In the context of a simulation study, it refers to the ability of the software or computer program to generate non-significantly different item parameters (Wang & Cheng, 2005).

Polytomous item: An item is a polytomous item if it is scored in more than two categories (DeMars, 2003; Zwick, Thayer, & Mazzeo, 1997).

Power: The probability of rejecting of the null hypothesis when the null hypothesis is false (true negative); 1 – β when β represents Type II error probability (Andrich & Hagquits, 2012).

Precision: Much broader and more fundamental than the concept of reliability. Measurements are said to be precise to the extent that they are consistent across different observations on the same object of measurement (Kane, 1996).

Root Mean Squared Error (RMSE): The square root of the average squared difference between estimated parameter values and the parameters used to generate the data or true parameters (DeMars, 2003; Stone, 1992).

Type I error: The rejection of the null hypothesis (e.g., null hypothesis of equal item parameters) when the null hypothesis is true. That is, an item is identified as displaying DIF when there is no between groups performance difference on the item (Clauser and Mazor, 1998).

Validity: Extent to which evidence and theory support the interpretations of test scores (Osterlind & Everson, 2009; Kane, 2013)

Validation: As formulated and elucidated by Kane, validation is, fundamentally, a simply-stated two-step enterprise: one that specify the claims inherent in a particular interpretation and/or use of test scores; and another that provides an evaluation of the claims based on empirical evidence and logical arguments (Kane, 2013).

CHAPTER TWO

LITERATURE REVIEW

“When people are evaluated, they want to be evaluated fairly” — Dorans, 2004, p. 45

Overview

This review of the literature addresses the following topics. First, this review of the literature addresses issues of test validation in the test construction process and the application of IRT to the analysis of noncognitive data derived from the measurement of attitudes. Next, the ubiquitous issue of missing data is introduced and its threats to test validity, and addresses the specific case of missing data in Likert-type scales and the methods that have been proposed for imputing missing values in such cases. Lastly, IRT is briefly introduced, the property of item invariance, and its application in the study of DIF using polytomously scored items. Finally, the application of missing data methods in DIF studies was reviewed.

Test Validation

Reforms in education have stressed the role of tests, the information they provide, and their intended purposes (e.g., improvement of education). While the views on the role of tests have been diverse (see, for example, Linn, 2000), the primary goal of validity has not changed: that of the intended interpretation and uses of test scores. Because of the impact that test scores have on examinee’ outcomes, statements such as that in Stone and Lane (2003), of the importance of ensuring that “the information provided by such programs is valid” (p. 1), and

Linn's (2000) suggestion, "Don't put all the weight on a single test" (p. 15) as a way to enhance the validity, credibility, and positive impact of assessment and accountability systems come right into the definition of validity provided by the *Standards:*

> A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. (p. 17)

Moss, Girard, and Haniford (2006) referred to validity as "the soundness of interpretations, decisions, or actions" (p. 109). That it, the validity of inferences from tests scores should seek not only the implementation of psychometric validation procedures of achievement tests but also the integration of other indicators such as those from the application of survey assessments (e.g., student attitudes toward school subjects). Stone and Lane (2003), for instance, conducted an examination of the relationship between changes in test scores and students' and teacher's attitudes toward a standardized test, finding that a "greater external validity was imparted to the interpretations" (p. 4). Kline (2000) asked, "how can we tell whether a test is valid or not?" (p. 17). As can be inferred from the previous paragraphs, within educational measurement, validity procedures are developed around the use of tests and have focused on "the evaluation of intended interpretations and uses of test scores" (i.e., score meaning), rather than on the test itself, "to inform decisions and actions" (i.e., consequences of test use) (Moss et al., 2006, p. 112). This approach calls for the types of evidence that should enable not only the "sound interpretations and uses" of tests (p. 115) but also for "expanding conceptions of assessment" (p. 122). But as Linn (1990) rightly noted, it is important not only to consider the validity of the use and interpretation of the scores, but it is also important to evaluate validity within a context or specific measurement issue. As such, in this study, validity is addressed concretely in the application of missing data methods (MDM) and their effect on the detection of DIF.

Missing Data

While missing data are not usually the focus of any given study (Schafer & Graham, 2002), missing data are a pervasive problem that researchers frequently encounter when conducting empirical research (Kromrey & Hines, 1994; Rubin, 1976; Schafer & Graham, 2002); that is, it is unlikely that researchers will have complete information for all cases and for all variables in their studies (Allison, 2001; Kim & Curry, 1977). Before employing data analysis methods, researchers must determine how missing data will be treated because the majority of the statistical techniques are not robust to missing data (Allison, 2001; Rubin, 1987; Schafer & Graham, 2002). If left untreated (that is, letting the software defaults proceed), the issues that arise due to missing data are very common, namely, reduced sample size (Davey et. al, 2005; O'Rouke, 2003; Zhang, 2003), and consequently, reduction of analytical power (Yenduri & Iyengar, 1994). In the context of large surveys, for example, Little (1988) stated that seriously biased results are due to the systematic differences between the observed data and the missing data. In sum, missing data may significantly affect the study outcome(s) due to the loss of information, thus complicating the interpretation of data analyses (Brockmeier, Kromrey, & Hogarty, 2003). The seminal work of Rubin (1976) and Little and Rubin (1987) on missing data provides one of the most accepted theoretical frameworks for its study, which is briefly presented next.

Rubin’s Missing Data Taxonomy

In the context of item response data, data are arranged in matrix form in which the rows correspond to observations (i.e., examinees *i*) and columns correspond to the variables (i.e., items responses *j*). The following notation (Zhang, 2003) is used to explain Rubin’s missing data taxonomy. Let *Y* be the *n* x *p* data response matrix where $y_\mathrm{i} = (y_{i1}, y_{i2}, \ldots y_\mathrm{n})^T$ and $y_\mathrm{j} = (y_{\mathrm{j}1}, y_{\mathrm{j}2}, \ldots y_{\mathrm{jp}})$ is a random sample from the probability distribution $P(Y \mid \theta)$. Further, let *R* be the missingness indicator variable where $r_{\mathrm{ij}} = 0$ if y_{ij} is missing (y_{miss}) and $r_{\mathrm{ij}} = 1$ if y_{ij} is observed (y_{obs}). Thus, *R* is under the conditional distribution of missingness $P(R \mid Y, \psi)$. Thus, for data arranged in matrix form, a model for the data would specify a probability distribution for the data $P(Y \mid \theta)$ for *Y* indexed by the unknown parameter θ and a probability distribution for the missing data $P(R \mid Y, \psi)$ for *R* given *Y*, indexed by the unknown parameter ψ. The joint probability distribution of the response variables and the missingness indicator can be expressed as,

$$P(Y, R \mid \theta, \psi) = P(Y \mid \theta)\, P(R \mid Y, \psi)$$

Thus, correct inferences on the parameter of interest θ will depend on how the probability model for missingness is defined. Rubin (1976) explained the reasons why data are missing and defined them as probabilistic mechanisms or processes that cause missing data. Rubin’s missing data mechanisms are missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR).

Data Missing at Random (MAR)

If the missingness of the data does not depend on the missing values (y_{miss}) but might depend on observed values in the data set (y_{obs}), then data are missing at random (MAR); that is,

$$Pr(r \mid y_{\mathrm{obs}}, y_{\mathrm{miss},}\, \psi) = Pr(y_{\mathrm{obs}} \mid \psi) \text{ for all } y_{\mathrm{miss}}\,,$$

The MAR mechanism “allows the probabilities of missingness to depend on observed data but not on missing data” (Schafer & Graham, 2002, p. 151).

Data Missing Completely at Random (MCAR)

Data are missing completely at random (MCAR) when the reason why data values are missing is unrelated to the variable itself as well as to other measured variables. Thus, if y = (y_{obs}, y_{miss}), where y_{obs} represents the observed values of Y and y_{miss} represents the missing values, data are missing completely at random (MCAR) if the missingness is independent from both observed and missing responses; that is, missingness is unrelated to the data,

$$Pr\ (r \mid y_{\text{obs}}, y_{\text{miss},}\ \psi) = P\ (r \mid \psi\ )\text{ for all } y_{\text{obs}}, y_{\text{miss}}$$

Under these two missing data mechanisms (MAR and MCAR), missingness is *ignorable* for likelihood based inferences (Rubin, 1976) because the observed data points represent a random sample of the hypothetically complete data set or it can be said that data missing at random and data missing completely missing at random are a random sub-sample of the original sample (O’Rourke, 2003).

Data Missing Not at Random (MNAR)

On the other hand, data missing due to y_{miss} are considered missing not at random (MNAR). That is, the distribution of missingness depends on y_{miss} and is thus considered nonignorable.

$$Pr\ (r \mid y_{\text{obs}}, y_{\text{miss},}\ \psi) \neq P\ (r \mid \psi\ )\text{ for all } y_{\text{obs}}, y_{\text{miss}}$$

An example by Schafer and Graham (2002) on blood pressure measurements helps exemplify MAR, MCAR, and MNAR missingness mechanisms. Systolic blood pressure for 30 patients was recorded in January (the complete data). Some patients have a second recording in February but not all. A scenario could be that from the complete January data, some patients were randomly selected for a blood pressure recording in February. For those not selected, missing the blood pressure reading in February is MCAR; that is, missingness is not due to the measured variable (blood pressure) or to any other variable in the study. In a different scenario, other patients returned for a blood pressure recording in February because their January recording showed hypertension. Thus, for those patients missing the reading in February is related to the reading in January or MAR; that is missingness is not related to the February reading but is related to the reading in January. As for the MNAR, a scenario could be if all patients returned for the recording in February but it was decided to record the measure if it showed to be in the range for hypertension. In this scenario, for those missing the recording in February, the missingness is not at random since it is related to the value of the variable. In addition to the relevance of the missing data mechanism, the application of a given MDM selected by the researcher can also have an impact on the study outcome(s), which might be reflected in biased parameter estimation (Robitzsch & Rupp, 2009), and in the ability of the statistical method to detect an effect (statistical power) if one is present. Missing data and MDM in Likert-type scales are not the exception.

Missing Data in Likert-Type Scales

Research using Likert-type scales tend to have missing data for several reasons. When respondents omit sensitive questions like income level or certain behaviors, such as sexual behavior, this type of missing data is called item nonresponse (Buhi, Goodson, & Neilands, 2008; Downey & King, 1998; O’Rourke, 2003). Item nonresponse could be also due to an examinee not reaching an item (i.e., examinee did not respond to the last item or items due to time constrains). Although there will be some research situations when the application of simple deletion procedures for treating missing data, such as Listwise and Pairwise, would be appropriate (e.g., large sample size, low percentage of missing data), previous work on missing data in Likert-type scales have reinforced the idea of the inadequacy of treating item nonresponse using these deletion procedures (Beale & Little, 1975), mostly if the assumption of data missing at random does not hold.

Like in the test theory and statistical model selection, the selection of an appropriate MDM depends on the factors of each research situation. As such, the type of data plays an important role in the method selected for treating missing data. In the case of Likert-type data, the items comprising a scale measure the same trait; consequently, scale items will correlate to a certain degree among them and with the total score of all items (Crocker and Algina, 1896). Thus, methods that consider item correlation can be appropriate for the treatment of missing data in Likert-type data (i.e., SRS, RMS, and PMS). Sample size and the magnitude of missing data are also relevant in MDM selection (Roth, 1994). Thus, in theory, methods that reduce the sample size by eliminating missing data (e.g., observations or items) can substantially impact statistical analysis; however, survey data are normally large and if in addition scales are short, which could lessen the overall loss of data (Raymond, 1987), Listwise deletion is a MDM to consider because the complete data generated will offer the advantage of consistent correlation matrices (Kim & Curry, 1977). In addition to two "state-of-the-art" MDM such as MI and FIML, complete data analyses were conducted for comparison.

Full-Information Maximum Likelihood

Full-information maximum likelihood (FIML) is the maximum likelihood estimation when there are missing values in the data. FIML does not impute missing values but derives parameter estimates and standard errors directly from the maximum likelihood (ML) estimation using available (observed) data. This estimation method has been improved over the years and reformulated (Bock & Lieberman, 1970; Bock & Aitkin, 1981), improving each time in regards to the computational demands. Basically, as its name implies, ML maximizes the likelihood (probability) of the estimated values as being what would have been observed if true. The formula for this likelihood or probability is

$$L(\theta) = \prod_{i=1}^{n} f(y_i \mid \theta) ,$$

where

θ = the parameter to be estimated, and $f(y \mid \theta)$ = the probability of observing y given θ

That is, ML is the probability of observing the data as a function of both the data and the missing or unknown parameter (i.e., likelihood of observing Y given some value of θ). Within this approach, the Bock and Aitkin (1981) reformulation of the ML estimates item parameters is the marginal distribution of ability. This estimation method is also known as marginal maximum likelihood (MML). When applied under some conditions, the MML is a case of the Expectation Maximization (EM) algorithm.

It is reported that FIML yields unbiased and asymptotically efficient estimates. It is also found that FIML performs as well as multiple imputation (Allison, 2001), but has advantages over multiple imputation in implementation because multiple imputation generally requires multiple steps from data imputation to generate multiple datasets to summary of results from multiple data analyses. However, FIML assumes that data are missing at random.

Multiple Imputation

Multiple imputation (Rubin, 1976, 1987; Little & Rubin 1987) is one of the most accepted methods for the treatment of missing data and (see for example, reviews by Graham & Hofer, 2000; Schafer & Graham, 2002; Schafer & Olsen, 1998). The availability of several statistical software packages for the implementation of MI makes important to delimit this section to MI as it is implemented in SAS. In addition, this section addresses only the imputation of categorical data, including the rounding needed so the imputed values fit the 4-point Likert scoring used in this study (i.e., strongly disagree (1), disagree (2), agree (3), and completely agree (4). Steps for applying multiple imputation (Rubin, 1987; Little & Rubin, 1987):

1. Missing values are replaced with a number of *m* plausible values determined by the researcher, creating $m \geq 2$ datasets with identical observed values across data sets but the imputed values will vary. This variability allows the MI procedure to consider the uncertainty of the missing data (Buhi et al., 2008; Patrician, 2002).

2. *M* completed datasets are then analyzed using standard procedures.

3. Results across *m* analyses are then combined into a single inference.

As explained before, the type of data is relevant to the selection of an appropriate MDM as it also important the missingness mechanism. In the case Likert-type data, MI is an appropriate method for imputing ordinal values but it is important to handle the imputation process so that imputed values are in the ranges in which the items' categories were scored.

Single Regression Substitution

In single regression substitution (SRS), for each missing variable, an observed variable most highly correlated to the missing variable is used to predict the missing value. That is, for a respondent presenting valid responses to items 1 through $a - 1$, but missing data for item a, the item that correlates most highly with item a is used to predict the missing item response.

Relative Mean Substitution

The relative mean substitution (RMS), designed specifically to estimate missing values for Likert-type scale items, estimates missing data using three sources of information: the person mean of the *k*th respondent for all valid (nonmissing) item scores, the grand mean of all valid item scores of all respondents, and the mean of all valid scores on the a^{th} item, excluding person *k* (Raaijmakers, 1999). More accurate imputations could be obtained if the mean of groups of similar records are used (Schulte, 1998). Thus, Raiijmakers' formula was adjusted for a multigroup scenario as,

$$X_{ak}^{(g)} = \left(\frac{\frac{\sum_{i=1}^{n} X_{ik}^{(g)}}{n}}{\frac{\sum_{j=1}^{N} \sum_{i=1}^{n} X_{ij}^{(g)}}{Nn}} \right) \left(\frac{\sum_{j=1}^{N} X_{aj}^{(g)}}{N} \right) ; (j \neq k)$$

where

g = group membership of examinee *k* (that is, g = R, reference group and g = F, focal group)

X_{ak} = the estimated value for missing item *a* for examinee *k* in group g

i = the valid responses to items 1 to *n* of examinee *k* in group g, and

j = the valid *N* cases of the sample g with no missing data excluding examinee *k*

For Raaijmakers (1999), important factors to consider when implementing a missing data method is that of the availability of adequate methods for treating missing data for the research problem at hand (e.g., sample size, proportion of missing data, missing data distribution, type of variables). In his investigation of the effectiveness of the RMS for estimating missing values in Likert-type items, Raaijmakers (in agreement with Downey & King,1998) stipulated the relevance of the correlation among items and scale reliability for the efficient performance of the RMS, which relies on this psychometric property of Likert-type items. The proportion of missing data in Raaijmakers' study was applied under five missing data combinations: 1) random missing items (30%), 2) random missing items (10%), 3) 20% of higher scorers (thus nonrandom) with 30% of missing items while 10% of missing items assigned to the other respondents, 4) two items with the most divergent sample means (thus nonrandom) were assigned 30% of missingness contrasted to 10% of the other items, and 5) proportion of missing items assigned according to the value of the item scores (thus nonrandom) so that 5% of random missing values were assigned to value 1, 10% to value 2, 15% to value 3, 20% to value 4 and 25% to value 5. Among Raaijmaker's results, the random results for the scales with 4, 5, and 6 items were of interest for this study, which consider scales of these lengths. The outcome (mean *d* differences on R^2 and β between true parameters and those from the MDM) for these scales showed that increases in mean differences were observed with increases on the proportion of missing items. Sample size was not an issue when the proportion of missing data was small (10%) and of course, the inverse was true: with higher proportions of missing data, mean differences were the largest.

Person Mean Substitution

The person mean substitution approach substitutes the mean of the nonmissing items for person *k* for person *k*'s missing items. That is,

$$X_{ak} = \frac{\sum_{i=1}^{n} X_{ik}}{n}$$

where

X_{ik} = the valid score *i* for person *k*

As explained previously, items in attitude scales are correlated to a certain degree. Thus, Downey and King (1998) stated that the mean of the items responded to by a person "seems a reasonable estimate for a missing item for that person" (p. 177). In Downey and King's application of the PMS to two scales using summed Likert scores (15-item and 20-item scales; *N*=834), both persons and items had missing data at varying proportions and each proportion of persons having missing data (from 5% to 35% in increases of 5%) were crossed with seven levels of missing items (10% to 70% in increases of 10%). Thus, the performance of the PMS was evaluated for each combination of the proportion of persons having missing data with each proportion of missing items (e.g., 5% of the persons in the sample had 10% of missing items; 5% of the persons in the sample had 20% of missing items). Analyses to evaluate PMS included running a correlation of original scores with generated scores after PMS and by comparing the reliability of the original scale with the reliability of the scale after PMS. Results indicated that for the 15-item scale, the correlations between the original scores and the substituted values declined when the proportion of missing items exceeded 30% and when the proportion of persons having missing data exceeded 20%. For the 20-item scale, correlations decreased when the proportion of missing items exceeded 50%, regardless the proportion of persons with missing items. In both cases (analyses for the 15- and 20-item scales), reliability was overestimated. Overall, the PSM worked well (i.e., high correlation between original scores and little variation in reliability) when the proportion of persons having missing data did not exceed 20%. The combination of the proportion of persons having missing items or less with the proportion of missing items being 20% or less provided stable outcomes. While the increase of either proportion beyond 20% resulted in less effective substitute values (smaller correlations and inflated reliability), the proportion of missing items had a more relevant impact on the outcomes.

Listwise Deletion

Listwise deletion treats missing data by deleting any observation with missing values. Also called complete case analysis, Listwise deletion has been frequently addressed not only as one of the most commonly used missing data methods but also as one of the most ineffective methods for treating missing data and its application has been strongly discouraged (see for example, Wilkinson & Task Force on Statistical Inference, 1999). Its implementation, it is argued, can lead to a great amount of data loss. However, before rejecting the idea of implementing Listwise deletion or any method, it is important to recognize it is not always clear how large the sample size needs to be or how much missing data is too much. As mentioned in the previous section, Listwise deletion has advantages that apply to the type of data and factors of this study and thus was selected for implementation.

Missing data are best studied within a particular context, such as DIF. Valid interpretations and fair uses of test scores require that the properties of test items (e.g., item discrimination and item difficulty parameters) are invariant among groups of examinees taking the test; that is, that test items are free of DIF. Some DIF methods are based on CTT and other methods are based on IRT; thus, in addition to evaluating the impact of MDM developed for Likert-type data within DIF analysis, the decision on which framework and procedure for conducting for conducting the statistical analyses should be taken within the theoretical and empirical specifics of each research situation.

Overview of CTT and IRT

As stated by Clauser and Mazor (1998), in the specific case of DIF analysis, the framework and DIF method cannot be selected in a cookbook fashion. Accordingly, the analysis of attitude measures should be conducted within the test theory framework, CTT or IRT, which best allows for valid score interpretations. Differences between CTT and IRT as well as basic terminology are introduced next. The advantage of using a polytomous IRT model for the analysis of attitude measurement and the detection of differential functioning was established.

Differences of CTT's Test Statistics and IRT's Item Statistics

Several authors have addressed the differences and the advantage of IRT methods over the methods used in CTT for estimating examinees' ability or trait in cognitive and noncognitive assessments (Embretson, 2004; Fan, 1998; Hambleton, 2004; Hambleton & Slater, 1997; Lin, 2008; Mislevy, 1989; Progar & Sočan; 2008; Wiberg, 2004). These test theories estimate ability or trait using different approaches or measurement models. A measurement or test model is a mathematical model in which "independent variables are combined numerically to optimally predict a dependent variable" (Embretson & Reise, 2000, p. 41). The models specify a scale for the dependent variable (e.g., a test score) and a design for how the independent variables (e.g., item responses) are combined to predict or explain the dependent variable. For binary data where observed responses are coded or scored 1 if the examinee responded correctly to the item or coded or scored 0 if the response to the item was incorrect, both CTT and IRT have specific mathematical measurement models for item analysis. CTT binary item statistics are presented first.

Classical Test Theory Item Analysis

For binary data (observed responses are coded 1 if the item was responded correctly or coded 0 if the item response was incorrect), the CTT mathematical measurement model is simply:

$$X_{\mathrm{ij}} = T_{\mathrm{ij}} + E_{\mathrm{ij}}$$

where

X_{ij}= observed test score *j* for examinee *i*

T_{ij} = true test score *j* of examinee *i*

E_{ij} = random error of measurement of test *j* of examinee *i*

That is, in the CTT model, estimates of examinee's observed score are assumed to consist of a true score and an error score (Downing, 2003; Hambleton & Slater, 1997). Table 1 shows binary data for 20 examinees and 6 binary items; x_i is the item score and total scores are arranged in descending order to conduct an item analysis for demonstration purposes (in reality, a larger sample size would be needed. See for example Crocker and Algina (1986) for a discussion on suggested sample size for conducting an item analysis).

Table 1

Classical Test Theory Item Analysis

	Items						
Examinee	1	2	3	4	5	6	Total Score
2	1	1	1	1	1	1	6
3	1	1	1	1	1	1	6
6	1	1	1	1	1	1	6
8	1	1	1	1	1	1	6
1	1	1	1	0	1	1	5
4	1	1	1	0	1	1	5
5	1	1	1	0	1	1	5
7	1	1	1	1	0	1	5
9	1	1	1	0	1	1	5
11	1	1	1	1	1	0	5
10	1	1	0	0	1	1	4
12	1	1	1	0	0	1	4
13	1	1	1	0	1	0	4
14	1	1	1	0	1	0	4
15	1	1	1	0	0	1	4
17	1	1	0	1	1	0	4
18	1	1	1	0	1	0	4
20	1	1	0	0	1	1	4
16	1	1	0	0	0	1	3
19	0	1	1	0	1	0	3
Difficulty (p_j)	.95	1.00	.80	.35	.80	.70	
Diff Upper	1.00	1.00	1.00	.60	.90	.90	
Diff Lower	.90	1.00	.60	.10	.70	.50	
Discrimination (d_j)	.10	.00	.40	.50	.20	.40	

Note. Data were created using random numbers, following an example provided by McDonald (1999) and further modified to illustrate the items' discrimination and difficulty properties.

The total score for examinee *i* can be obtained as a sum of the items responded correctly. In Table 1, the total score for examinee *i* was computed as the total sum of item *j* scores using McDonald's (1999) formula,

$$y_\text{i} = \sum_{j=1}^{m} x_\text{ji}$$

Where y_i is the observed correct total score by the *i*th examinee.

As observed in Table 1, within CTT, an item analysis is conducted by using indices of item difficulty (*p*) and item discrimination (*d*). For binary data, item difficulty p_{j} is defined as the proportion of examinees responding correctly the *j*th item in sample *n*.

$$p_{\mathrm{j}} = n_{\mathrm{j}}/n$$

Using the sample *n* data in Table 1, the value of the item difficulty p_{j} for items 1, 2, 4, 6 are computed as follows

$$p_1 = \frac{1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1}{20} = \frac{19}{20} = .95$$

$$p_2 = \frac{1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1}{20} = \frac{20}{20} = 1.00$$

$$p_4 = \frac{1+1+1+1+1+1+1}{20} = \frac{7}{20} = .35$$

$$p_6 = \frac{1+1+1+1+1+1+1+1+1+1+1+1+1+1}{20} = \frac{14}{20} = .70$$

It should be easy to appreciate that item 2 is a very easy item, that all examinees could respond correctly (*p* = 1.00). Item 4 on the other hand, is a very difficult item because the proportion of examinees responding the item correctly was only .35 (See Carey (2001) for a complete description of item difficulty levels). The sample item difficulty index *p* is also an indicator of population item difficulty parameter or the probability π_j of scoring item *j* correctly (McDonald, 1999). Thus, $p_{\mathrm{j}} = \pi_{\mathrm{j}} = \mu_{\mathrm{j}}$.

Item discrimination (*d*) compares the difference in performance between high scoring students and low scoring students (e.g., top 10 students and lower 10 students in Table 1 respectively). Item discrimination values around 0 indicate that the item does not discriminate between high and low test performers.

While $T_{\mathrm{ij}} + E_{\mathrm{ij}}$ are considered fixed values, they will vary for examinee *i* on different testing occasions (Allen & Yuen, 2001). This is precisely where the limitation of CTT is more obvious because the item statistics indices (i.e., item difficulty and item discrimination) will vary depending on the ability of the examinees taking the test and the test item difficulty. This dependence of CTT's item and test statistics on the sample (termed "circular dependency" by Fan, 1998, and group-dependent by Hambleton, 1989) is a major disadvantage because it means that the comparison of examinees and the comparison of items across examinations is not feasible due to group variability and more importantly, the interpretation of test scores through these item and test statistics are valid only for the sample from which they are obtained (Embretson & Reise, 2000). Also, Fan (1998) mentioned that this dependency makes it difficult to apply of CTT to other testing applications such as test equating and computerized adaptive testing.

When CTT is used to analyze test data, both student ability and test item difficulty confound or depend on each other, regardless of test format and scoring. Bock, Mislevy, and Woodson (1982) for example, argued on the dependence of CTT statistics on the sample, pointing out that within CTT, the standard errors for number-right (number of items answered correctly) test scores are always random and independent or uncorrelated. Moreover, Downing (2003) stated that the effect of uncontrolled conditions on the observable score (e.g., poorly constructed test items, inadequate testing conditions, examinee's internal conditions such as inattention, illness or fatigue) make up for the randomness of the error of measurement in CTT analyses, interfering with the "precise and accurate measurement of the examinee true ability or proficiency" (p. 740). This is not to say that CTT is never appropriate (Kromrey & Bacon, 1992; Mislevy, 1989); CTT might work well in some situations or settings (e.g., locally developed tests such as classroom tests). Moreover, the less complicated computations in CTT both at the item and test level, and its weak assumptions ease the applicability, understanding, and interpreting of item and test scores in educational and psychological measurement instruments, which could be considered an advantage (Fan, 1998). Lin (2008) argued that in those cases in which available data do not allow the implementation of an IRT method, CTT can be used to develop parallel test forms as effectively as when using IRT. However, for such cases in which the validity of the inferences made of test scores is relevant, IRT offers important advantages over CTT, the most important being that of sample independence of item parameters estimates and item independence of person parameter estimates.

While the advantage of these IRT person-free item statistics and item-free ability estimates (as termed by Wright, 1968) has been explored in achievement measurement, less research has been conducted on the application of IRT to noncognitive assessments and on polytomous items. Thus, the use of the IRT framework using a graded model in the present study is justified. A brief background on IRT is presented next to introduce basic terminology.

Item Response Theory Item Analysis

The previous section explained that IRT models involve a higher mathematical level than that required in CTT methods. Fortunately, some scholars' discussions on IRT advances and applications are accessible, easy to understand. In this section, these contributions are summarized to introduce important concepts in IRT and delve into the description of IRT item analysis.

Baker (1977) and Embretson and Reise (2000) explained the basics of IRT by defining traits as unobservable characteristics that people possess and that account for their behavior, for example, intelligence and verbal ability. Baker noted that while traits, denoted in IRT using the lower-case Greek letter Theta (θ), have a “considerable intuitive meaning” (p. 299), they cannot be observed (or measured) directly and as such, traits are considered as being latent. However, within IRT traits can be estimated or inferred from responses to test items intended to measure a given construct, the item response being an indicator of the examinee’s standing on the latent variable (Embretson & Reise, 2000). The next section provides a working definition of the IRT measuring Theta scale also denoted θ, followed by the explanation of its explicit meaning and relationship to the examinee’s trait and to item difficulty and item discrimination parameters using basic IRT models.

Trait (θ)

Kane (1996) defined measurement as a procedure with which a number is assigned to an object to represent the value of some attribute for that object of measurement. But the interpretation of the measurement of an object requires a specific comparison or standard to which the measurement (e.g., a test score) is compared and a scale or numerical basis for the comparison (Embretson & Reise, 2000). When measuring student achievement using CTT, for example, test scores can be easily interpreted as the sum of correct responses or can be interpreted as McDonald (1999) suggested, by selecting a metric on the basis of the distribution of scores and interpreting the test scores using norm-referenced criteria in which an examinee’s performance is compared against the performance of other examinees by placing the test scores in the normal distribution scale.

In IRT, on the other hand, examine trait interpretation derives from modeling an examinee’s responses to test items (Linn, 1990). When IRT is used to measure an examinee’s latent trait θ, which “is hypothesized to be behind observable items” (Samejima, 2010, p. 78), the specific comparison that is needed for score interpretation is carried out by placing both trait level θ and items in a common scale. Before turning to the description of IRT item difficulty and discrimination parameters, the scale of measurement in IRT is introduced.

Measuring Scale (θ)

To apply Kane’s (1996) definition of measurement as a procedure with which a number is assigned to an object to represent the value of some attribute for that object of measurement, a rule for the assignment of those numbers is needed. Posed with the challenge of coming up with a scale of measurement for subjective magnitudes, as meaningful as those scales used to measure, for example, height and weight, Stevens (1946) defined measurement as “the assignment of numerals to objects or events according to rules” (p. 677). His understanding that different rules lead to different scales helped to make explicit rules for the assignment of numbers to quantify a given observed attribute and the construction of appropriate scaled values. That is, the choice of numerals for measurement is the choice of a metric or scale (i.e., interval, ratio, ordinal, and nominal).

Which scale is appropriate for best interpretation of IRT person and item parameters? As it has been stressed in this review of the literature, the decision of which method, procedure, or as in this case, the scale of measurement is to be selected should be based within the theoretical and empirical specifics of each research situation. To illustrate first how the IRT common scale is constructed, Table 2 extends Table 1 by including the proportion of examinee’s correct responses. Table 2 illustrates the proportion of correct items by each examinee, ranging from π_i .50 to π_i 1.00 for the lowest to highest scoring examinees respectively.

Table 2

Classical Test Theory Item Analysis

	Items							
Examinee	1	2	3	4	5	6	Total Score	Proportion
2	1	1	1	1	1	1	6	1.00
3	1	1	1	1	1	1	6	1.00
6	1	1	1	1	1	1	6	1.00
8	1	1	1	1	1	1	6	1.00
1	1	1	1	0	1	1	5	.83
4	1	1	1	0	1	1	5	.83
5	1	1	1	0	1	1	5	.83
7	1	1	1	1	0	1	5	.83
9	1	1	1	0	1	1	5	.83
11	1	1	1	1	1	0	5	.83
10	1	1	0	0	1	1	4	.67
12	1	1	1	0	0	1	4	.67
13	1	1	1	0	1	0	4	.67
14	1	1	1	0	1	0	4	.67
15	1	1	1	0	0	1	4	.67
17	1	1	0	1	1	0	4	.67
18	1	1	1	0	1	0	4	.67
20	1	1	0	0	1	1	4	.67
16	1	1	0	0	0	1	3	.50
19	0	1	1	0	1	0	3	.50
Difficulty (p_j)	.95	1.00	.80	.35	.80	.70		
Diff Upper	1.00	1.00	1.00	.60	.90	.90		
Diff Lower	.90	1.00	.60	.10	.70	.50		
Discrimination (d_j)	.10	.00	.40	.50	.20	.40		

While the proportion of correct responses provides a CTT index of student proficiency or standing at the construct measured, this proportion is a tentative indication of the level of examinee ability (McDonald, 1999); the proportion of correct responses by each examinee, computed by averaging the correct or endorsed responses, cannot be used to state that examinees with the same proportion value of correct responses or endorsed have the same level of ability or standing at the construct measured.

In addition, note for example that while examinees 1, 4, 5, 7, 9 and 11 obtained the same total score (5) and same proportion of correct responses (π = .83), there is a difference in which items these examinees scored correctly or endorsed. For example, all of these examinees responded correctly or endorsed “easy” items (*p* = .95, 1.00, .80) but only examinees 9 and 11 responded correctly or endorsed item 4, which is the most difficult in Table 2 (*p* = .35).

Thus, it is important to consider the item difficulty for estimating the examinee’s ability when estimating the probability of responding correctly to a given item. Based on the examinee ability (θ) and the item difficulty, the probability of selecting the correct response or endorsing an item is related to the difference between examinee’s trait level θ and the item difficulty or location β (Embretson & Reise, 2000). Within IRT, an examinee’s trait level is estimated through modeling the relationship between examinee’s responses to the test items and test item parameters,

$$P_{\mathrm{j}}(\theta) = \frac{1}{1 + \exp[-D(\theta - \beta_{\mathrm{j}})]}$$

where

$P_{\mathrm{i}}(\theta)$ = probability of correct or positive response to the *j*th item among individuals with a score θ on the underlying trait.

β_{j} = item difficulty.

D = scaling factor, usually 1.702 (see Camilli, 1994, on the origin of this constant)

exp(*x*) = exponential function that raises the mathematical constant *e* (natural log base 2.718) to the power of *x*

IRT Item Parameters

Basic IRT models (e.g., one-, two-, and three-parameter models) are characterized by the number of item parameters they include. The IRT one-parameter model (1PL) is the simplest IRT model and as its name implies, has one parameter, the item difficulty parameter denoted *β*.

The mathematical model for estimating the probability of selecting the correct response or endorsing an item in binary data is

$$P\left(X_{\mathrm{ij}} = 1 \middle| \theta_{\mathrm{i}}, \beta_{\mathrm{j}}\right) = \frac{\exp(\theta_{\mathrm{i}} - \beta_{\mathrm{j}})}{1 + \exp(\theta_{\mathrm{i}} - \beta_{\mathrm{j}})}$$

where

X_{ij} = response of examinee *i* to item *j*

θ_{i} = trait level for examinee *i*

β_{j} = difficulty of item *j*

Thus, the probability of examinee *i* responding correctly or endorsing an item (x_{ij}= 1) is a function of the examinee's ability or trait $P(\theta_{\mathrm{i}})$ and the item difficulty parameter (β_{j}). One of the most important features of IRT models is the Item Characteristic Function (ICF) which represents the probability of selecting the correct response for a given item or endorsing the item as a function of the examinee's ability or trait (θ) and the item difficulty (β) in the scale continuum; that is, $P(X_{\mathrm{ij}} = 1)$. After the item parameters for an IRT model are estimated, the item parameters' information can be used to model the response patterns of a given item across different levels of ability. The graphical representation of this probability of correct response or endorsing of an item is called the Item Characteristic Curve (ICC), denoted as $P_j(\theta)$. Figure 1 shows the ICC for a dichotomous item (e.g., scored 1 if correct and scored 0 if response is incorrect as in achievement measures).

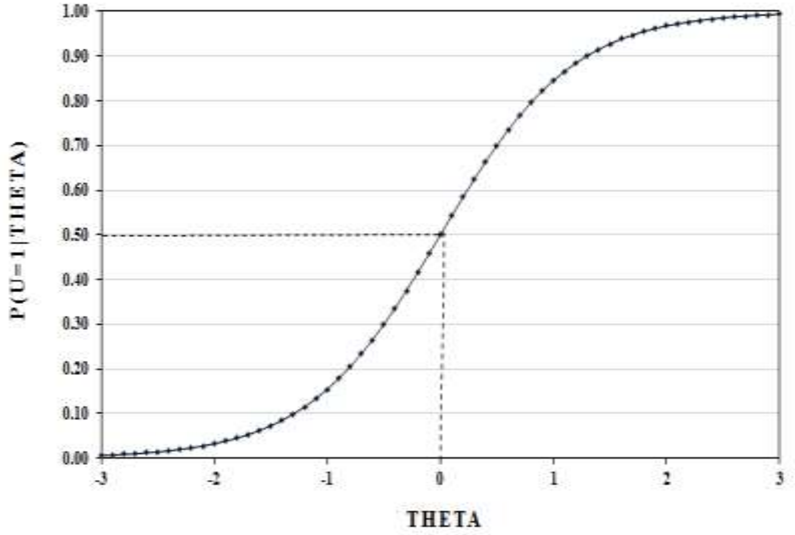


*Figure 1***.** Item characteristic curve or ICC for a dichotomous item, at which the likelihood of a correct response is .5 or 50%.

When an item is dichotomously scored, the probability of responding correctly to a given item, as in the case of achievement measurement for example, can be plotted so that the item difficulty and the ability scale are expressed in the horizontal plane. In other words, item difficulty and examinee ability are located in a common scale. Figure 1 shows the location in the Theta scale of the ICC of a dichotomous item with *b*=0, as a monotonically, s-shaped increasing curve at which an examinee with an ability or trait θ = 0 has a probability of .50 of responding correctly the item. The probability of answering this item correctly increases with increases in ability; that is, examinees with higher ability levels (i.e., θ > 0) have higher probabilities of responding the item correctly. It is an advantage of IRT, that the metric of θ or Theta levels corresponds to the location of β. ; see for example Figure 2 in which the ICC of three dichotomous items having different item difficulty parameter values are plotted.

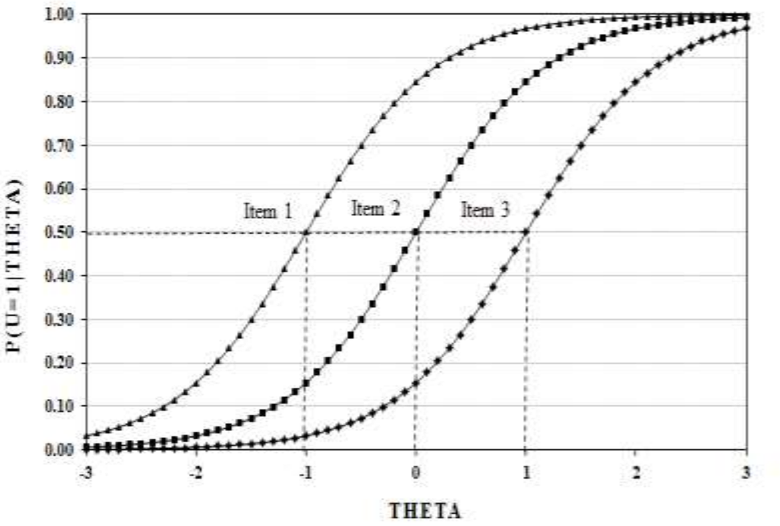


Figure 2. Item Characteristic Curve (ICC) for three items differing in difficulty parameter (*b*=-1, *b*=0, and *b*=1).

Besides observing in Figure 2 the monotonic relationship between the probability of correct response to an item and ability level, the ICC for each of the three items also shows the items location on the ability scale. The mean of the item score distribution serve as the threshold of the item on the ability scale at which the probability of correct response or item endorsement is equal to .50 (i.e., *b* = -1, *b* = 0, and *b* = 1). That is, each ICC plotted in Figure 2 differs in location (*b*) at ability levels (θ) as it is shown by the reference lines from the ability levels to the inflection point of each ICC. Thus, item 3 (*b* = -1) is easier than items 1 and 2 (*b* = 1 and *b* = 0 respectively). Item 2 is easier than item 1 but more difficult than item 3. This relationship of the Theta level and the item location in the common scale indicates the probability at which an examinee is likely to pass or endorse an item. For instance, if

$$P_{\mathrm{j}}(\theta) = \frac{1}{1 + \exp[-D(\theta - \beta_{\mathrm{j}})]}$$

For an examinee θ = 1.0 and item difficulty *b* = 1.0

$$P_{\mathrm{j}}(\theta) = \frac{1}{1 + \exp[-1.7(1 - 1)]}$$

$$P_{\mathrm{j}}(\theta) = \frac{1}{1 + \exp[-1.7(1 - 1)]}$$

$$P_{\mathrm{j}}(\theta) = \frac{1}{1 + 1}$$

$$P_{\mathrm{j}}(\theta) = .50$$

If examinee θ = 1.0 and item difficulty *b* = 0.0

$$P_{\mathrm{j}}(\theta) = \frac{1}{1 + \exp[-1.7(1 - 0)]}$$

$$P_{\mathrm{j}}(\theta) = \frac{1}{1 + 0.18}$$

$$P_j(\theta) = .85$$

If examinee θ = 1 and item difficulty *b* = -1.00

$$P_{\mathrm{j}}(\theta) = \frac{1}{1 + \exp[-1.7(1 - (-1))]}$$

$$P_{\mathrm{j}}(\theta) = \frac{1}{1 + 0.03}$$

$$P_{\mathrm{j}}(\theta) = 0.97$$

As observed in Figure 3, an examinee having an ability θ = 1 has a .50 probability of responding correctly item 3 (*b* = 1) and a much higher probability of responding correctly or endorsing items 2 (*b* = 0; *π* =.85) and 1 (*b* = -1; *π* =.97).

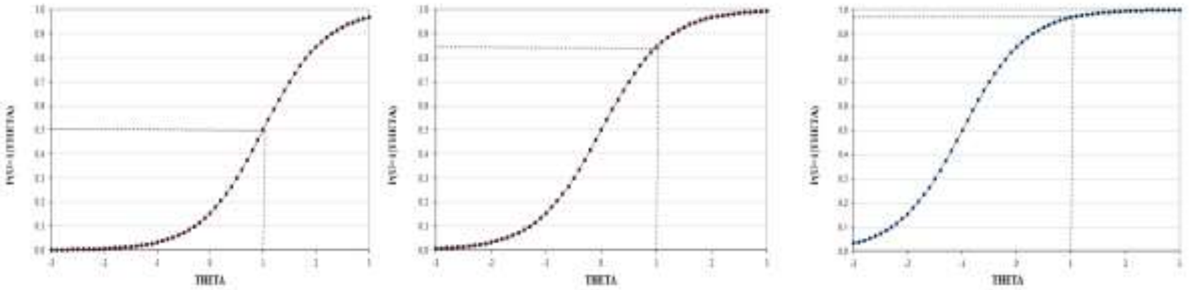


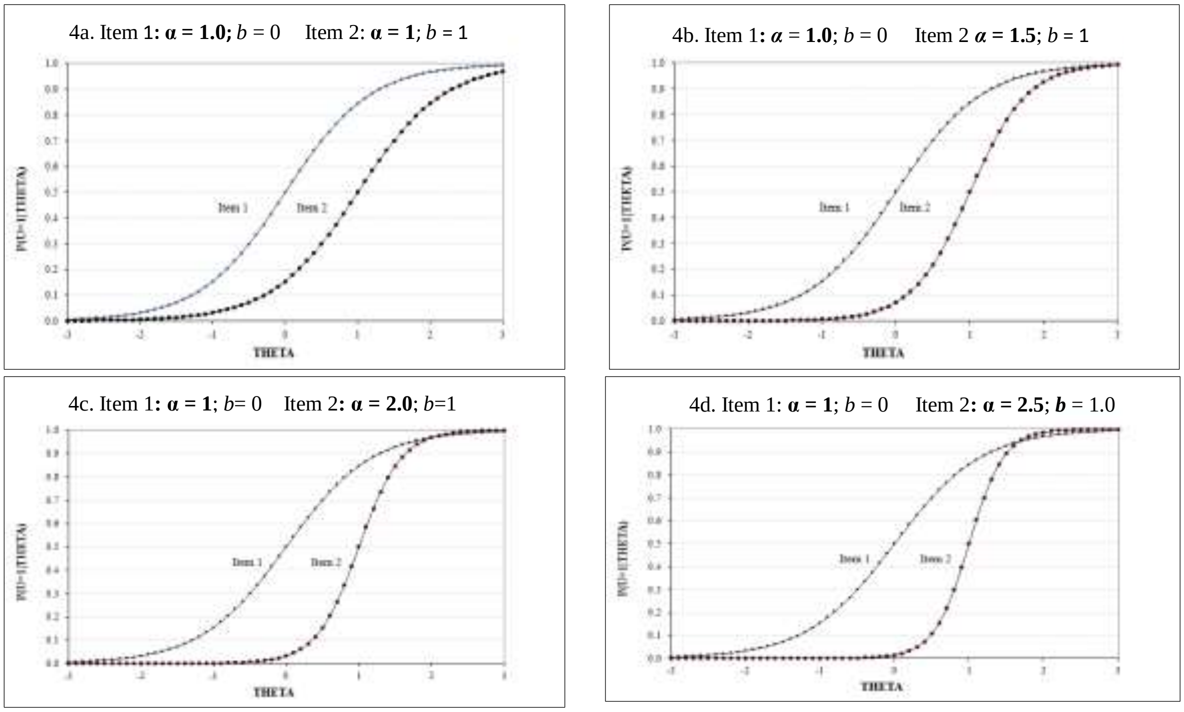
Figure 3. Probability of correct response at different levels of item location *b*

Another basic IRT model is the two-parameter model (2PL) for which in addition to the difficulty parameter β, a discrimination parameter (α) is included in the model. This two-parameter model or 2PL estimates the probability of correct response given θ, as a function of *α* and β as,

$$P\left(X_{\mathrm{ji}} = 1 \middle| \theta_{\mathrm{i}}, \beta_{\mathrm{ji}}\right) = \frac{\exp[\alpha_{\mathrm{j}}\left(\theta_{\mathrm{i}} - \beta_{\mathrm{j}}\right)]}{1 + \exp[\alpha_{\mathrm{j}}\left(\theta_{\mathrm{i}} - \beta_{\mathrm{j}}\right)]}$$

Whether the value for the item discrimination parameter α is freely estimated or for example, is set constant to 1.0, the implication of adding the discrimination parameter to an IRT model is evident in the ICCs for the items in Figure 4.

The item discrimination and difficulty parameters for two items are shown in Figure 4. The item parameters α and *b* for item 1 are constant (i.e., same parameters' values) across the four graphs and the item parameter *b* is also held constant in item 2 across all graphs. When comparing the ICCs for both items in each graph, the effect of increasing *α* by .5 in item 2 is observed. The ICC for item 2 becomes "steeper" as the value of *a* increases.


4a. Item 1: α = **1.0;** *b* = 0   Item 2: α = **1**; *b* = 1

4b. Item 1: α = **1.0**; *b* = 0   Item 2 α = **1.5**; *b* = 1

4c. Item 1: α = **1**; *b*= 0   Item 2: α = **2.0**; *b*=1

4d. Item 1: α = **1**; *b* = 0   Item 2: α = **2.5**; ***b*** = 1.0

Figure 4. Item discrimination parameter varying by .5 across graphs steeper.

The item discrimination parameter for item 2 across four graphs shows how rapidly the probabilities of correct response change with trait level. The probability change is much slower for item 2 when *a* = 1; thus item 2 in graph 4a is less discriminating than item 2 in the last graph 4d when *a* = 2.5.

The previous examples of the graphical display of the items' ICC help understand how in the context of IRT examinee θ and item *b* are located in the same logistic metric. The location of both θ and *b* in the Theta interval scale allows for the interpretation of *b* as the point in the scale at which the likelihood of an examinee for responding correctly or endorsing the item would be .50. Because in IRT the difficulty parameter *b* indicates the item location in the Theta scale, *b* will be addressed as the location parameter hereafter to distinguish it from the difficulty parameter *p* in CTT. As described, the mathematical form of the ICC varies among IRT models depending on the item's characteristics. The differences observed in the ICC's trace lines help understanding how the item' parameters work within a model and the key properties or assumptions that IRT models involve, such as dimensionality and local independence. The dimensionality and local independence assumptions are presented next before the introducing the GRM, the selected IRT model for conducting the DIF analyses of this study.

Dimensionality

One of the major underlying assumptions of most commonly used IRT models is the unidimensionality assumption. That is, only a single factor (latent construct) underlies or explains the relationship of a set of items and consequently, a single trait level represents or characterizes person differences (Embretson & Reise, 2000). Most research on IRT models has been done on unidimensional models (McDonald, 1999) and this study will apply an unidimensional model, but it is worth to mention that there are times when test items measure more than one ability or trait and more complex IRT models are needed, such as those for multi-trait or multidimensional IRT (MIRT) models, in which item responses depend on, or are explained by two or more latent traits simultaneously.

Local Independence

Within CTT, test items are expected hold together in a common factor (i.e., their common attribute) that explains the covariance of a pair of binary item scores. In a population with a fixed value of the factor score, the covariance equals zero. This is the pairwise conditional independence (McDonald, 1999). But IRT makes a stronger assumption, which is called local independence. Controlling for the latent trait or ability, the responses to any item are assumed to be independent; for a given examinee, the response to an item should not be affected by the response to another item. Local independence can be mathematically defined as

$$P(X_1, X_2, X_3, \ldots, X_{\mathrm{m}}|\theta) = P(X_1|\theta)P(X_2|\theta)P(X_3|\theta) \ldots P(X_{\mathrm{m}}|\theta)$$

where X_1, X_2, …, X_{m} are the items in a test and θ represents the trait or ability measured by the items. In other words, the assumption of local independence means that pairs of items are uncorrelated. When an appropriate dimensionality (e.g., unidimensionality or multidimensionality) is not specified, local independence is likely to be violated (Embretson & Reise, 2000).

Model Selection for Attitude Measurement

Both CTT and IRT have dichotomous and polytomous models for the analysis of attitude measurement. But as it has been already discussed, more is known about CTT methods and on the application of statistical models for dichotomously scored items and they will not be discussed here. Rather, attention is given to the use of polytomous items for the measurement of attitudes within the IRT framework.

Data collected in attitude measurements are often from polytomous items and the information that the IRT methods provide is considered more interpretable and less ambiguous than that from CTT methods. The next section provides an overview of attitude measurement and introduces the GRM model for graded responses.

Attitude Measurement

An expanded view of test validity is one that poses student learning in context. As it is used in schools and classrooms, assessments inform, for example, decisions on curriculum, targeting topics in need of improvement, setting new goals, and developing specific instructional plans to meet those goals (Moss et al., 2006). Fulfilling these purposes “requires multiples types of evidence” (p. 123), for example, that from surveys which are considered efficient methods for obtaining information about attitudes. Lane, Liu, Ankenmann, and Stone (1996) also stated that “the relationship between scores on an assessment and other measures can provide additional validity evidence” (p. 72).

According to Alwin (1992), an attitude is a latent, unobserved tendency to behave positively or negatively; that is, attitudes are often assumed to have direction and intensity (e.g., approve or disapprove, agree or disagree). The purpose of the measurement of attitudes is to obtain a response along the response scale. There is a variety of response scales to use in the measurement of attitudes, from the use of scales using dichotomous items (e.g., agree/disagree) to the use of polytomous scales having three or more response categories (e.g., strongly disagree/disagree/agree/strongly agree). While it has been argued that the former are easy for respondents (McKennell, 1974), it has also been argued that three or more response categories yield the desired level of precision in social measurement (Benson, 1971).

Likert scales are widely used for measuring attitudes and typical response modes in items constructed in the Likert tradition require examinees to indicate the degree of intensity with which they agree or disagree with statements, to indicate the degree of importance they place on statements, and or to indicate how often they behave in certain ways (always, often, sometimes, seldom, never). For these cases, when item responses consist of three or more ordered options or categories, the GRM is an appropriate, reasonable model for Likert-type data (Koch, 1983).

The Graded Response Model (GRM)

In attitude assessments, polytomous, Likert-type items with *m* ordered response categories require examinees to indicate, for example, the degree of intensity with which they agree or disagree with statements. This degree of intensity is underlined by the k^{th} item category, whose label (e.g., Strongly Disagree, Disagree, Agree, Strongly Agree) not only underlines the degree of intensity of the response but also the direction or order of the categories. Samejima's model for graded responses, the graded response model (GRM; Samejima, 1969, 2010) is an appropriate model for this type of items (Steinberg, 2001). The next section introduces this model and provides an overview of its assumptions.

Model Definition

For a dichotomously scored item, there are two IRFs. One IRF is for the correct response $P_{\mathrm{j}}(\theta)$ and the other IRF is for the incorrect response $Q_{\mathrm{j}}(\theta) = 1 - P_{\mathrm{j}}(\theta)$. But for each polytomously scored item, there are m_{j} option response functions (ORFs) representing the item response process. Samejima's (1969, 2010) GRM is estimated as,

$$P(k) = \frac{1}{1 + \exp[-a(\theta - b_{\mathrm{k}-1})]} - \frac{1}{1 + \exp[-a(\theta - b_{\mathrm{k}})]} = P^{*}(k) - P^{*}(k+1).$$

where

a = discrimination parameter or the slope of the ORF quantifying the relationship of the item to the latent variable

$b_{\mathrm{k-1}}$ = the k^{th} threshold parameter representing the response endorsement, or the point in in the scale at which the probability of responding to category *k* passes .5

$P^*(k)$ = probability of selecting category *k*

Thus, as Steinberg (2001) explains, *P**(*k*) or the probability of selecting category *k* for a given item is equal to the probability of selecting *k* minus the probability of selecting category k +1 or higher, *P**(*k*+1). That is, in the GRM, for an item with m_j ordered response categories, the cumulative operating characteristic COC or $Pj_k(\theta)$, is the conditional probability of an examinee's response falling in category *k* to item *j* or higher as a function of θ. In other words, $Pj_{\mathrm{k}}(\theta)$ represents the relation of the ability scale and the cumulative probability over the *m* ordered response options for a given polytomous item. Baker and Kim (2004) and Steinberg (2001) observed that at each probability level,

$$\sum_{k=1}^{m} P_k(\theta) = 1\ ,$$

Given the homogeneous case of Samejima's GRM logistic model,

$$P^*_{jk}(\theta) = \{1 + \exp[-\alpha_{\mathrm{j}}(\theta - \beta_{\mathrm{jk}})]\}^{-1},$$

in which each item *j* includes a discrimination parameter α_{j}, and location parameters β_{jk} with two or more *k* categories,

$$\begin{array}{ll} 1 - P^*_{\mathrm{j}1}(\theta) & \text{when } k = 1 \\ P^*_{\mathrm{j}(\mathrm{k}-1)}(\theta) - P^*_{\mathrm{jk}}(\theta) & \text{when } k = 2,\ldots(K_{\mathrm{j}} - 1) \\ P^*_{\mathrm{j}(\mathrm{k}_{\mathrm{j}-1})}(\theta) & \text{when } k = K_{\mathrm{j}} \end{array}$$

Accordingly, the graphical representation of the ORFs of a polytomous item will differ from the IRF in the dichotomous case. Figure 5 shows the ORFs for a 4-option graded item.

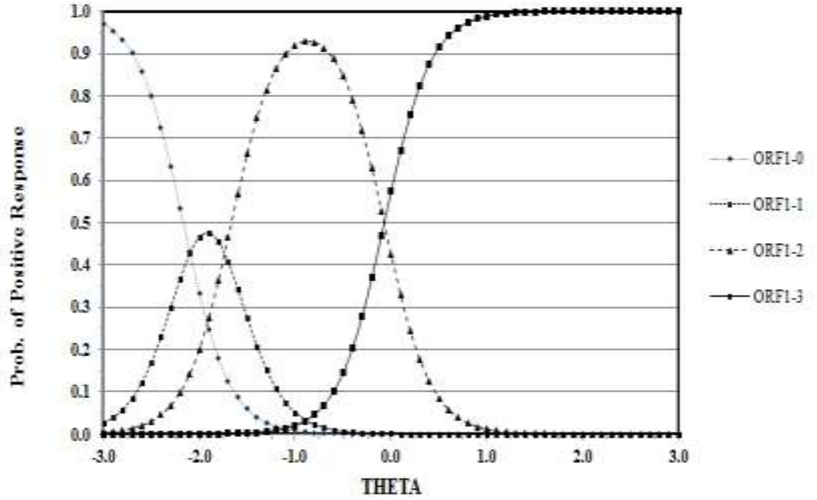


Figure 5. Option Response Function for a graded item with four categories

For a polytomously scored item with *m* graded categories, the ORFs do not have the same form or trace line. It can be observed in Figure 5 that the ORF for the lowest category *k*=0 decreases as θ increases while the ORF for the highest category *k*=3 increases as θ increases. The two intermediate ORFs, for categories *k*=1 and *k*=2, increase and then decrease. Because the ORFs in the GRM do not have the same form, Samejima (1969) defined the between category threshold or boundary response function (BRF). For the item in Figure 5 there are 3 thresholds (*m*-1) which will be located in the latent trait or Theta continuum. Figure 6 shows the thresholds or BRFs for a 4-option graded item.

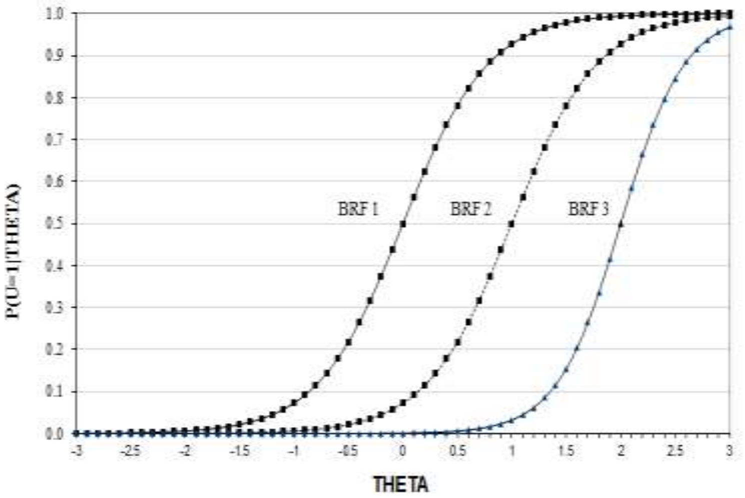


Figure 6. Boundary response functions for a 4-option item

Thus, the threshold parameters for the GRM in Figure 5 are given in Figure 6 as BRF. Once the BRFs are computed, the ORFs can be computed as the difference between successive BRFs. That is,

(0 vs. 1, 2, 3)

(0, 1 vs. 2, 3)

(0, 1, 2 vs. 3)

Based on the assumption that the calibration sample is drawn randomly from a population in which person ability (θ) is normally distributed ~*N*(0,1), and if data fit the logistic model for the ORF, IRT raises the possibility of making testing more efficient by allowing equally valid inferences across different populations of examinees. However, to do so, item parameter invariance should hold.

Parameter Invariance

Parameter invariance is the fundamental and most important property of measurement for IRT models (Rupp & Zumbo; 2004), “the cornerstone of IRT” (Hambleton et al., 1991, p. 18). Parameter in IRT refers to both item and person parameters within a specific IRT model, whose values are estimated with a random sampling design (Rupp & Zumbo, 2004). Within IRT, examinees of the same ability (θ) have the same probability of answering correctly or endorsing an item of given characteristics; thus, item parameter estimates result in invariant parameter values across groups (Hambleton, 1989; Lane et al., 1995). That is, when the item parameter invariance property holds, item parameter values are invariant across groups and a given IRT model is likely to fit the data across populations. Many applications of IRT capitalize on this property of parameter invariance, namely, test design, test equating, item banking, computer-adaptive testing, and DIF analysis (De Ayala & Sava-Bolesta, 1999; Hambleton & Slater, 1997). Because important inferences are made from test scores, the stability of the item parameter estimates has been studied under varied conditions. Several studies have investigated the recovery of item parameters under different models and factors and are summarized in the next section.

Item Parameter Recovery

The property of item parameter invariance has been used to investigate the effect of sample size, latent trait distribution, number of items, and number of response options on the recovery of item parameters. In general, polytomous IRT models require large sample sizes for accurate estimation of item parameters (De Ayala & Sava-Bolesta, 1999; DeMars, 2003) because polytomously scored items have more item parameters to be estimated. Several authors have conducted research to investigate the relationship of sample size and item parameter estimation and have offered different criteria for selecting the sample size needed for this purpose.

De Ayala and Sava-Bolesta (1999), and DeMars (2003) suggested that a sample size ratio (SSR) to the total number of item parameters to be estimated was an important factor for the accuracy of the estimates. Using a simulation approach, De Ayala and Sava-Bolesta conducted a parameter recovery study using the nominal response model (NRM; Bock, 1972) in which SSR was a manipulated factor; other factors included latent distribution (LD) and total number of parameters to be estimated (i.e., the total number of items is multiplied by number of parameters in the model. In this case, two item parameters are estimated (discrimination and location) and the product is multiplied for the number of item options). Each dataset generated in their study had 28 items with either 3 or 4 response options. Thus, the four different SSR used (2.5:1, 5:1, 10:1, and 20:1) corresponded to samples of 420, 840, 1680, 3360, and 560, 1120, 2240, 4480 for the 3-option items and 4-option items, respectively.

As expected, increases in SSR produced higher correlations between item parameters and their estimates and for a given interaction between SSR and the latent distributions factor (LD; normal, skewed, and uniform), consistently larger correlations were observed for the 3-option items and uniform LD. DeMars (2003) expanded De Ayala and Sava-Bolesta findings by also conducting a simulation study on the NRM to evaluate the effect of sample size on item parameter estimation, in addition to the effect of number of items and number of item categories. Whereas De Ayala and Saba-Bolesta varied the number of parameters to be estimated by manipulating the number of item categories per item, DeMars manipulated the number of parameters to be estimated by increasing the total number of items. Generated datasets crossed two tests lengths of 20 and 40 items with three categories per item (six parameters per item, 120 and 240 total item parameters respectively) and two test lengths of 10 and 20 items with 6 categories per item (12 parameters per item, 120 and 240 total item parameters). Thus, sample size 2400 (10:1 and 20:1 SSR) and sample size 600 (2.5:1 and 5:1 SSR) were studied. Results showed high correlations and nearly unbiased item parameter estimates across all study conditions, but correlations were consistently higher and bias was consistently lower when the number of parameters per item was six.

Although the Ayala and Sava-Bolesta and the DeMars studies evaluated the effect of sample size and test length on parameter estimation, the results seem to indicate that sample size and test length depend on the IRT model and the purpose of the investigation. To that effect, DeMars cautions readers that her findings “may not extend to unexplored factors” (p. 287). Of concern is the limited number of replications conducted in the studies because the variability due sampling error is not controlled with 25 or even with 100 replications. However, the estimates were consistent; that is, larger samples provided more stable and accurate parameter estimates.

In DeAyala and Saba-Bolesta’s study, a total of 25 replications were generated for each of the 72 conditions (4 SSRs, x 3 LDs, 3 $I_{\max}$s (items’ maximum amount of information) x 2 *mi* s). Item parameters were estimated using MULTILOG 5.1 using the default program parameters. Number of iterations and number of cycles were set at 999, which in the authors’ opinion, was “unrealistically high” (p. 6). Convergence was typically achieved in less than 25 cycles. As expected, as the SSR increased, the correlation between the item parameters and their estimates increased when the number item options were three. Under the different LDs, the difference between the number of item options (3 and 4) decreased as SSR increased for the positively skewed LD, and remained constant for normal and uniform LDs. In addition, LD accounted for 43.5% of the variability in RMSE and SSR accounted for 29.5%. The source of variability in

RMSE introduced by the number of options, 3 and 4 was similar. The estimation of item difficulty parameters was not affected by the LD (there are similar mean RMSE across the levels of LD). One of the most important conclusions of the research revised is that “it remains difficult to propose a heuristic rule of what sample size to category ratio is adequate” (DeMars, 2003; p. 287).

But as mentioned before, many applications capitalize on the parameter invariance property. Considering the centrality of the differences in test performance among subgroups in educational research, and the implications of the use of test scores, an important application of the parameter invariance property is that of differential item functioning or DIF. The following section provides a brief introduction to DIF, from a working definition, to the description of the DIF procedure selected for application to attitudinal assessment using the GRM.

Differential Item Functioning (DIF)

The concern on disparities in test performance across subgroups of examinees is not new. Clearly and Hilton (1966) and Clearly (1968) investigated whether a test presented a differential difficulty for students in terms of their racial and socioeconomic backgrounds and proposed a definition of test and item bias. Clearly and Hilton defined a biased item as one that produced an uncommon discrepancy (e.g., differences in average scores) in the performance of one group, compared with the performance of other group or groups taking the test; that is, bias was conceptualized as an item-group interaction. Rather than looking at the item level, Cleary conceptualized the whole test as a possible cause of differences in test scores. While these early approaches to the study of test bias or differences in test performance did not find biased items, it was acknowledged that the term “bias” being used gave the process an unintended connotation of unfairness. Over time, the interest switched to the study of differences in test scores due to the differential performance of problematic items (i.e., those measuring a construct in different ways for two or more groups), and differential item functioning or DIF replaced the early nomenclature of item bias. Some methods for detecting DIF items are explained next.

Methods for Detecting Differential Item Functioning

Basically, there are two types of DIF detection methods. The parametric approach, as its name implies, assumes a specific item response model; however, unless taking the risk of making untenable assumptions, using nonparametric methods is another option. Besides categorizing DIF methods as parametric and nonparametric, DIF methods can be categorized as those for which the criterion is an observed score and those for which the criterion is a latent trait (Millsap & Everson, 1993). DIF methods can also be classified according to the number of response options, that is, there are methods for detecting DIF in binary data and methods for detecting DIF in polytomous data. Methods for detecting DIF in polytomous data can be further categorized, depending on whether the item category options are ordinal or nominal. For ordinal data, e.g., data from attitude measurement analyzed using a model for graded responses, the IRT Because less is known about the detection of DIF in graded responses, the next section introduces how DIF analyses are conducted when using the GRM.

Differential Item Functioning in the Graded Response Model

Within IRT, a DIF study for a set of items requires the estimation of item parameters for subgroups having taken the test (e.g., boys and girls). One group is considered as the group of interest (i.e., the focal group, F); the second group (i.e., the reference group, R) is the group against which the performance on the studied item is compared. If the property of item parameter invariance holds for the studied item, such item is considered a nonDIF item; but if the item parameter invariance property does not hold, that is an indication that the item true score function is not equal for the reference and focal groups; thus, the item is a DIF item (Cohen et al., 1998) . Thus, this property can be used to test the hypothesis that the item parameters for the reference and focal groups are invariant. Considering that in the evaluation of DIF in the IRT framework the matching variable is a latent variable, potential differences in item parameters between the focal and reference groups in the GRM with m_j categories are measured as,

$$\hat{\xi}_{\mathrm{jR}} = [\hat{\alpha}_{\mathrm{jR}}, \hat{\beta}_{\mathrm{j1R}}, \ldots \hat{\beta}_{\mathrm{j(m_j-1)R}}]'$$

$$\hat{\xi}_{\mathrm{jF}} = [\hat{\alpha}_{\mathrm{jF}}, \hat{\beta}_{\mathrm{j1F}}, \ldots \hat{\beta}_{\mathrm{j(m_j-1)F}}]'$$

where

$\hat{\xi}_{\mathrm{j}}$ = vector of differences between parameter estimates for item *j* $[\hat{\alpha}_{\mathrm{jF}} - \hat{\alpha}_{\mathrm{jR}}, \hat{\beta}_{\mathrm{jF}} - \hat{\beta}_{\mathrm{jR}}]'$

That is, DIF in the GRM (Cohen, Kim, & Baker, 1993) is present if:

$a_{\mathrm{jR}} \neq a_{\mathrm{jF}}$ and $b_{\mathrm{jkR}} = b_{\mathrm{jkF,}}$

$a_{\mathrm{jR}} = a_{\mathrm{jF}}$ and $b_{\mathrm{jkR}} \neq b_{\mathrm{jkF,}}$

or

$a_{\mathrm{jR}} \neq a_{\mathrm{jF}}$ and $b_{\mathrm{jkR}} \neq b_{\mathrm{jkF}}$

This study will address the case in which DIF is present in the location parameters only and the shift (Δ) in *b* is the same across all item's categories *k* (uniform DIF).

Overview of the Likelihood Ratio Test for Detecting DIF

When comparing two models, the likelihood ratio (LR) goodness-of-fit statistic is an appropriate technique for this purpose. Within IRT, one method for detecting DIF in polytomous items with ordered categories is the Likelihood Ratio test (IRT-LR; Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988; Thissen, Steinberg, & Wainer, 1993). The IRT-LR compares the likelihood functions estimated from the reference and focal groups to investigate DIF. For each item *j*, a discrimination parameter α, and a set of category boundaries β are estimated. For uniform DIF (i.e., $\alpha_{jF} = \alpha_{jR}$), the null hypothesis and the alternative are as follows

$$H_0: \beta_{jkF} = \beta_{jkR} \text{ for all } k = 1, 2, 3, \ldots, m\text{-}1$$

and the alternative

$$H_1: \beta_{jkF} \neq \beta_{jkR} \text{ for at least one } k \text{ for item } j$$

where

F= focal group

R = reference group, and

k= 1, 2, … , *m* – 1 for *k* response categories boundaries

The IRT-LR DIF estimation involves the comparison of two item response models with respect to a critical value (e.g., χ^2 or Bonferroni) at a specified nominal α (e.g., .01 or .05). In the IRT-LR free-baseline approach, a baseline or compact model (also called constrained model) is fitted by constraining an anchor or referent item (i.e., a no-DIF item) so that its parameters are equally calibrated for both reference and focal groups and the baseline's –2 times the log likelihood fit index is obtained. Next, the rest of items in the scale are studied or evaluated for DIF one at a time by creating a series of augmented models, one for each studied item, in which the studied item is added to the referent item and is also constrained, while the rest of the item parameters are freely estimated. To evaluate the studied item for DIF applying the free-baseline IRT-LR test approach, the –2 times the log likelihood statistic of the augmented model is compared with that of the baseline model. If the G^2 statistic or absolute value of the chi-square difference of the augmented model to the compact model exceeds the critical value, the IRT-LR test is significant at the corresponding *p*-value, the null hypothesis is rejected, and the studied item is categorized or flagged as a DIF item (Stark, Chernyshenko, & Drasgow, 2006).

Therefore, the IRT-LR tests whether additional parameters in the augmented model are significantly different from zero, which would indicate DIF in the studied item. This process is continued until all items are evaluated for DIF. The LR statistic (Ankenmann et al., 1999; Cohen, Kim, & Wollack, 1998; Stone & Lane, 2003, and Thissen et al., 1993) is thus defined as

$$G^2(df) = -2\log\left[\frac{Likelihood\ (\mathrm{C})}{Likelihood\ (\mathrm{A})}\right]$$

where,

df = degrees of freedom, estimated as the difference between the number of parameters in the augmented model and the number of parameters in the compact model,

G^2 = absolute difference between the -2 times the log likelihood for the compact model [C] and -2 times the log likelihood for the augmented model [A].

Under the null hypothesis of no DIF, the value of G^2 is assumed to be distributed as χ^2 (*df*). Thus, if the value of G^2 is large (i.e., an unlikely value), the null hypothesis and the compact model are rejected (Thissen et al., 1993, p. 73).

A Numerical Example

Data for a 5-item survey have items graded using a 5 category response format (from strongly disagree to strongly agree). Item 5 is the DIF item and the other 4 items are nonDIF items. Table 3 contains the critical chi-squares at a level of significance equal to .05, for 1 to 7 degrees of freedom (*df*). Because the DIF method evaluates one item at the time, it implies that multiple tests of significance are conducted (one for each studied item). Thus, a Bonferroni adjustment, based on 5-items, is applied to set an upper bound on the family wise error rate, to ensure that with each test of significance, the probability of rejecting or falling to reject the hypothesis, is not greater than the nominal alpha = .05.

Table 3

Critical Chi-Square and Critical Bonferroni Values at Significance Level .05 for 4 df

Upper tail chi-sqr (*p* =.05)		# items	5
df	ChiCrit (.05)	Bonf *p*	ChiCrit(Bonf)
5	**11.07**	**0.01250**	**14.54**

The estimated chi-square statistic or G^2 for the baseline model and for each consequent studied item are displayed in table 4. In this case, the G^2 value of the baseline model is equal to -2294.1. For item 2, the constrained model estimate of G^2 is equal to -2256.80. The absolute difference between the G^2 values of the baseline and constrained models is equal to 37.30. That is, -2294.1 – 2256.80 = 37.30. Since this difference (37.3) exceeds the critical values (11.07 and 14.54) shown in Table 3, item 2 is flagged as DIF item (yes) as shown in table 4.

Table 4

IRT-LR DIF Test using Graded Responses with 5 Response Categories

	Free Baseline Model	Constrained or Comparison Models			
	Reference	Item 2	Item 3	Item 4	Item 5
Chi-square (G^2)	-2294.1	-2256.80	-2291.00	-2288.10	-2290.80
Difference		37.30	3.10	6.00	3.30
DIF		**Yes**	no	no	**no**

Because item 2 was not a pre-designated DIF item, yet the Likelihood Ratio (LR) test determined that it was, a Type I error was made (FP; false positive). On the other hand, item 5 was the pre-designated DIF item but LR failed to detect it, which is a Type II error (FN; false negative). Items 3 and 4 were correctly classified as nonDIF items (TN; true negatives).

Missing Data and Differential Item Functioning

Whenever an examinee does not respond to an item, the item nonresponse generates patterns of missing data that can impact adversely the inferences from test scores as well as affect the effectiveness of the methods used to evaluate DIF. While research on the effect of MDM on the effectiveness of DIF methods has been limited, “state-of-the-art” missing data methods (e.g., multiple imputation and likelihood-based methods) have been implemented to treat missing data before conducting a DIF study. This research for handling missing data within DIF is introduced next.

Previous Studies of the Effect of Missing Data on Differential Item Functioning

Both Emenogu (2006) and Garrett (2009) conducted dissertations studies addressing the effect of missing data on the performance of procedures for DIF detection. Using a total of 41 multiple-choice, dichotomously scored items from the national data (Ontario, Canada) of the 1995 Trends in International Mathematics and Science Study (TIMSS) and 75 items from the School Achievement Indicators Program (SAIP) 2001 Mathematics Assessment, Emenogu (2006) investigated the effect of three MDM (Listwise deletion, analysiswise deletion, and scoring missing data as incorrect) on the performance of the Mantel-Haenszel non-parametric DIF detection method. For his applied study, Emenogu sampled students in the last year of secondary schooling (Population 3, ~ 18 years old) that took the English version (*n*=756) and French version (*n* = 318) of the TIMSS comprised the reference and focal groups respectively. As for the SAIP test, students that took the English-language version (*n* = 452) and students that took the French version (*n* = 304) comprised the reference and focal groups respectively. In addition to manipulating the MDM factor, the study varied the matching criterion too (i.e., total score, TS, and proportional score, PS). This 3 x 2 completely crossed factorial design had thus 6 experimental conditions and kept the level of significance constant at .01. The proportion of missing data and the magnitude of DIF were not manipulated by the researcher; the levels of both factors were studied as they were present in the data. Both patterns of missing data and differential item performance between the two groups for some items were presented using line graphs. The statistical results of the effect of the MDT on the MH-DIF procedure were presented in tabled form in separate sections: Patterns of performance and nonresponse, MH DIF and missing data treatments, patterns of missingness and item difficulty were addressed but there is no specific quantification (i.e., percentage of students missing items or percentage of items missing responses). Existing differences or no differences between items attempted and items responded correctly are mentioned but the percentage is not provided in any case. The same omission is observed when DIF is discussed. Per the author description, results showed that the application of the PS matching criterion to the TIMSS data resulted in more items showing significant DIF values and that MDT performed similarly in 80% of the items in terms of identifying the same magnitude and direction for 33% of the DIF items. As was noted, this study was an applied research study (or case study) in which neither missing data nor DIF were manipulated. Instead, actual patterns and rates of missingness and DIF magnitude were evaluated on each data set as they were present in the data.

Garrett (2009) on the other hand, conducted her dissertation as a simulation study using polytomous items. In addition to complete data analyses, the effect of missing data on the performance of the Mantel test and the ordinal logistic regression method for DIF detection was studied implementing the within-person mean substitution and multiple imputation missing data methods, when data was missing completely at random (MCAR, i.e., missingness was spread randomly across all items). Data for 20 items, 2 of which were the studied items, were generated for the partial credit model using the IRTGEN macro (Whittaker et al., 2003) and factors of the study included 1) Balanced sample size (*n* = 500 / *n* = 500) and unbalanced samples (*n* = 700 / *n* = 300, *n* = 900 / *n* = 100, *n* = 845 / *n* 355, and *n* = 1183 / *n* 313 for referent and focal groups respectively): 2) Percentage of missing data (10%, 25%, and 40%); 3) DIF magnitude (.25, .50, and .75); 4) Impact (0, 1 for both reference and focal groups respectively, and 0, 1 / -0.5, 1 for reference and groups respectively). Outcomes of the study were Type I error and statistical power. Type I error rates were very similar for both studied items. When data were complete and there was no impact simulated, Type I error rates were close to the nominal level ($\alpha = .05$);

however, higher rejection rates were observed for unbalanced samples. The OLR showed higher rejection rates than the Mantel method. When impact was present, Type I error rates were above the nominal level and higher rejection rates were observed in the balanced sample and overall, for the Mantel tests. In analyses with missing data and no impact, MI showed rejections rates below the nominal level across DIF methods and across sample size and proportions of missing data. It was observed that in this simulation condition, rejection rates decreased as missing data increased. When impact was present, Type I error rejection rates for MI were higher than when impact was not present but rejection rates for the within-person mean substitution method were about the same than when impact was not present. For the missing data conditions, Type I error rates were very similar than the complete data rejection rates for both studied items. Power conditions when data was complete showed lower power rates for the smaller DIF condition (.25) across all sample sizes for both DIF methods; however, smaller power estimates were observed for the unbalanced sample sizes. When missing data were present, power increased as DIF magnitude increased and power decreased as the percentage of missing data increased. This pattern was very consistent for both studied items.

Robitzsch and Rupp (2009) conducted a factorial, completely crossed simulation study to investigate the impact of five MDM (Listwise deletion, zero imputation, two-way imputation, adjusted two-way imputation, and MICE) on two methods, Mantel-Haenszel (MH) and logistic regression (LR), for uniform DIF detection. Both DIF methods are observed-variable matching methods selected on the basis of being widely applied in practice, easy to implement, and requiring small sample size for parameter estimation. Additional factors of the simulation included missing data mechanisms (MAR I, MAR II, MCAR, and MNAR), impact *N*(0, 1:0, 1), *N*(0, 1:-.5, 1), and *N*(0, 1:.5, 1) for focal and reference group respectively, balanced sample sizes

(250:250, 1000:1000, and 4000:4000) , number of items (20, 40), percentage of missing data (10%, 30%), and DIF magnitude (0, .2, .4, .6). Missing data and DIF were manipulated only for the first item (the studied item). The Rasch model was used to generate unidimensional data using the specified population distributions. Outcomes of the study were bias, RMSE and the rejection rate of the null hypothesis at .05 significance level, over 1000 replications. The bias and RMSE analyses were conducted to investigate the accuracy and precision of the estimated parameters ($\hat{\Delta}_r$= $\log\left(\hat{\alpha}_{\mathrm{MH,r}}\right)$ for the MH procedure, $\hat{\Delta} = \hat{\beta}_{\mathrm{2r}}$ for the LR procedure and $\hat{\Delta}$= $\hat{\beta}_{\mathrm{focal}} - \hat{\beta}_{\mathrm{ref}}$ for the IRT-based procedure), in terms of the amount of DIF manipulated. An analysis of variance (ANOVA) for BIAS for main effects and one-, two-, and three-way interactions showed that the effect size of the first-order interactions of missing data mechanism and missing data rate, missing data mechanism and MDT, and missing data rate and MDT were significant (η^2 =.12, .07, and .05 respectively). Both bias and RMSE analyses were discussed as a possible cause for the DIF rejection rates. For example, rejection rates under the baseline conditions (i.e., no missing data were present) were at the nominal level when DIF and impact was not present. But when DIF was infused, the parameters of interest were biased, even if minimal (e.g. bias = .04), with higher rejection rates as DIF effect increased. Mean bias for the MCAR and MAR missing data mechanisms were relatively similarly low for all methods except when the missing data was treated as wrong or incorrect (i.e., missing data coded 0) and negative bias were observed when the missingness mechanism was MNAR. Mean RMSE was reported to mirror the bias results. However, rejection rates were reported being high overall, even for the no DIF cases. While some conditions were reported as showing significantly high Type I error rates, even up to 100%, it is not disclosed which conditions showed these high Type I error rates. Power was reported to be very low for some conditions and there is mention to some conditions having no power whatsoever. However, there is no information as to which conditions had low to no power. Robitzsch and Rupp (2009) argued that interpreting these high rejection rates as Type I error rates or power was not straightforward due to significant biased results in some conditions. (i.e., high rejection rates and low power could not be traced back to the simulation factors). DIF was imposed to the IRT-based conditions but analyses of Type I error rates and power were not conducted or reported. For this part of the study, which had fewer conditions than the MH and LR methods (no imputation was implemented), only bias and RMSE were analyzed. The interaction of sample size and number of items showed a significant effect on bias ($\eta^2 = .38$), and the impact of the interaction of the missing data mechanism and missing data rate on bias was also large ($\eta^2 =.13$). It was concluded that the choice of DIF method was not important; however, the results do not show empirical evidence of that (e.g., DIF method is not included in the factorial ANOVA analyses).

Finch published in 2011 two studies that explored the effect of Multiple Imputation (MI), Listwise deletion (LD), and treating missing as incorrect on the Mantel-Haenzsel (MH), logistic regression (LR), and SIBTEST methods for detecting uniform and nonuniform DIF in binary data, in terms of their Type I error rates and power across 100 replications. These DIF detection methods were selected on the basis of being reported in the literature as performing similarly well. As for the missing data methods, selected based on their use in previous research, the application of multiple imputation (MI) to DIF analyses was of particular interest to the author. Additional factors in the simulation included balanced sample sizes (250:250, 500:500, 1000:1000), impact $N(0, 1: 0, 1)$ and $N(0, 1:-.5)$, percentages of missing data (5% and 15%), DIF magnitude (0.03, 0.06), levels of item difficulty (1, 0, -1), and type of missing data (MCAR, MAR, MNAR). Nominal alpha was constant at .05. Complete data analyses were conducted as a reference or baseline comparisons and the 3-PL logistic model was used for data generation. An analysis of variance (ANOVA) was conducted to investigate factors main effects and their interactions. Finch reported that the analysis of the impact of the study factors and their interactions showed similar results across methods; thus, only results for LR were reported. For the analysis of Type I error, large effect sizes were observed. As done by Robitzsch and Rupp (2009), Finch reported the effect size of 4-, 3-, and 2-order interactionss. The interaction effects of sample size by impact by missing data method by type of missing data showed a large effect size ($\eta^2 = 0.416$) at .05 significance level, and the third-order interaction effect for method by item difficulty by type of missing data was $\eta^2 = 0.535$. Second-order interactions of method, by item difficulty, by type of missing data had also a large effect size ($\eta^2 = 0.703$). These results were similar to the results for the no DIF conditions, especially for the MCAR data. But the impact of missing data on the rejection rates for MAR where higher when missing data were treated as incorrect.

As for Finch (2011) article on the effect of MDM on the detection of nonuniform DIF, the IRTLR DIF detection method was included, eliminating the MH procedure, and the SIBTEST procedure was substituted for the crossing SIBTEST method (CSIB). Thus, in addition to MI, Finch’s investigation on nonuniform DIF included the stochastic regression imputation method (SRI), omitted as incorrect (i.e., zero imputation or ZI), and Listwise deletion (LD). The outcomes of interest were the Type I error and power of the DIF methods. Sample size levels and impact were the same as the investigation of uniform DIF. The percentage of missing data were manipulated to infuse 10%, 20%, and 30% missingness, in addition to the complete data analysis (i.e., no missing data) across MAR, MCAR, and MNAR missing data mechanisms. DIF to simulate nonuniform DIF (differences in the discrimination parameter) imposed to favor the reference group (.0, 0.4, 0.8, and 1.0). Data were generated using the 3-PL model for 20 and 40 items, designating only one item as the studied item. MCAR missingness were imposed to both the reference and focal groups on the studied item; MAR on the other hand, was implemented 1) to associate missingness to group membership by imposing missingness only to the studied item in focal group and 2) imposing MAR randomly to the studied item in both the reference and focal groups. For MNAR, missing data were imposed on those cases from the reference and focal groups having an incorrect response in the studied item. Results for the Type I error rates showed that the following interactions had large effect sizes: the interaction of the missing data method by the percentage of missing data by the type of missing (η^2 =.388; p = .004); the interaction of method by impact (η^2 =.3; p < .001) and the interaction of method by sample size (η^2 = .424; p < .001). LR and CSIB had similar Type I error rates across all conditions but the rejection rates for the IRTLR were higher than the rejection rates for the other methods. For the MCAR results, ZI and LD showed similar results across levels of missing data for the LR, CSIB, and IRTLR DIF methods. The rejection rates for MI declined slightly when missing data went from 10% to 30% while the rejection rates for CSIB increased within the same levels of missing data. As expected, complete data analyses showed lower rejection rates than those for all the other methods. When missing data were 20%, the rejection rates were comparable with those of complete data. When 10% of missing data were imposed only in one group, it resulted in severe inflated Type I error rates for ZI and SRI, above the nominal alpha .05. and up to .70 when 30% of missing data were imposed. Surprisingly, the rates for MI were higher than those for LD. When MAR missingness was simulated in both groups, LD maintained the rejection rates at the nominal alpha .05 while MI did not. The rejection rates of ZI and SRI were at .05 when the percentage of missing data was 10% but above the rejection rates of complete data analysis. For

MNAR, results showed that for the LD, Type I error results were not elevated when the percentage of missing data was 30%; For MI, results showed elevated rejection rates across all levels of missing data but tended to decline as the level of missing data increased. When mean differences were the same for the reference and focal groups, DIF methods observed good but when impact was imposed, Type I error was inflated. MI showed inflated error rates across DIF methods even when mean differences between groups were the same. The IRTLR DIF method rejection rates were lower with the larger sample size and no missing data was imposed. Otherwise, the rejection rates of the IRTLR were similar to those observed in the other methods. As expected, power was higher for the largest DIF effects (i.e., large DIF was easy to detect). Also as expected, power was higher for larger sample sizes across MDM and levels of missing data, but decreased as missing data increased. For the small sample size condition (250/250), the IRTLR had a slightly lower power than LR or CSIB. Differences in power were observed across missing data mechanisms when impact was also manipulated. When no impact was imposed (i.e., group means were the same), power for LD was slightly lower than the power for the complete data analysis but dropped substantially with MNAR as the percentage of missing data increased. Also when impact was not simulated, the power of MI and LD was comparable, except when MAR was imposed only in the focal group.

Summary

Because missing data and the methods used to treat it can have a significant impact on the DIF detection methods, their selection requires careful consideration. For valid inferences of tests scores, missing data should not be overlooked and when conducting a DIF analysis, factors such as sample size, proportion of missing data, and magnitude of DIF should help in the selection of an appropriate MDM as demonstrated in results sections of the articles. Because only binary, achievement data were explored in the articles summarized, the study of the effect of missing data on the performance of DIF methods using noncognitive, polytomous data seems appropriate. In contrast to the large test lengths generated in the revised studies and the mostly nonparametric models used, this dissertation will approach the problem of missing data in short scales to investigate the effect of missing data in Likert-type scales on the Type I error rates and power of the IRT likelihood ratio (IRT-LR) test for detecting DIF. MDM developed for the treatment of missing data in Likert-type scales will be implemented, which will allow addressing the following research questions: 1) What is the effect of missing data (i.e., item nonresponse) and their treatment on the Type I error rate of the Likelihood Ratio test for Differential Item Functioning detection? To what extent is the effect consistent across MDM? To what extent is the effect consistent across sample size? To what extent is the effect consistent across percentage of missing data by persons and items? To what extent is the effect consistent across the magnitude of DIF? And 2) What is the effect of missing data and their treatments on the statistical power of the Likelihood Ratio test for Differential Item Functioning detection? To what extent is the effect consistent across MDM? To what extent is the effect consistent across sample size? To what extent is the effect consistent across percentage of persons and items with missing data? To what extent is the effect consistent across the magnitude of DIF?

CHAPTER THREE

METHOD

Design of the Simulation Study

Quantitative researchers make inferences about populations of interest through the use of methods that will work. That is, methods that will estimate item parameters efficiently (i.e., accurately and precisely) when the factors under investigation meet the selected method's specific conditions or assumptions. When the selected research factors have not been studied extensively or have not been studied at all under certain conditions, assumptions, or methods, simulation studies provide "an excellent method for evaluating estimators and goodness-of-fit statistics under a variety of conditions, including sample size, nonnormality, dichotomous or ordinal variables, model complexity, and model specification" (Paxton, Curran, Bollen, Kirby, & Chen, 2001; p. 288).

Metropolis and Ulam (1949) explained that there are two methods by which data are generated in a simulation study. One method is a stochastic or random process that draws independent new sets of data; the other method is a deterministic method under which the value of the simulated datasets "are strictly determined by the value of other parameters" (p. 338). Based on the literature, Kromrey and Hines (1994) strongly recommended using actual field data to generate the simulated data so that given study factors and / or methods are applied to realistic situations, which in turn provide a better index of effectiveness than those obtained from random-generated processes. Considering that simulation studies within an IRT framework can be conducted to evaluate the validity of their models' performance in less-than-ideal conditions (Harwell, Stone, Hsu, & Kirisci 1996), in addition to the advantage of conducting a simulation study by generating data that emulate the conditions under study, using real data will provide the additional advantage of carrying out the study under normally occurring violations of assumptions (e.g., normality of distributions; Kromrey & Hines, 1994).

This study will conduct a Monte Carlo simulation using item parameter estimates from the Civic Education study (U. S. Department of Education, National Center for Educational Statistics, 1999), hereafter CivEd, conducted by the International Association for the Evaluation of Educational Achievement (IEA), to generate item response data that conform the GRM. A crossed mixed factorial design will be used to evaluate the effect of six MDM on the Type I error rates and statistical power of the IRT-LR test for detecting DIF. A complete cases analysis will be conducted for comparison purposes.

Overview of the IEA Civics Education Study

The International Association for the Evaluation of Educational Achievement (IEA) conducted the second Civics Education Study (CivEd, 1999) as a two-phased study, aiming to assess the content and process of civic education in the 28 participating countries, which included the United States. Using information collected in phase 1 (1996-1997), phase 2 of the study (1998-2000) consisted of a test that assessed students' knowledge in the domain of civic education as well as a survey of their attitudes toward concepts of democracy, citizenship, and government. Table 5 shows the content domain of each of the survey subscales that assessed students' attitudes relating to citizenship, democracy, government, civic issues, and expected political participation.

Table 5

Survey Scales of the Civics Education (CivEd) Study and Their Goals.

Section	Subscales	Core Goals
Civic concepts	A. Democracy B. Good citizens C. Government D. Trust in institutions	Section covers students' understanding of the concepts of democracy, citizenship, and government.
Attitudes	E. Our country F. Opportunities 1 G. Opportunities 2 H. Immigrants I. The political system J. School K. School curriculum	Section covers the degree of adherence to common values and attitudes, which along with knowledge of rights and responsibilities is required for creating sustaining democratic institutions.
Behavior (action)	L. Political action 1 M. Political action 2 N. Classrooms	Political participation is a central characteristic of a democracy. Thus, section covers political interest and exposure to political news, as well as expected participation in political activities.

Source: Schulz, W., & Sibberns, H. (Eds). (2004). *IEA Civic Education Study technical report*. Amsterdam, The Netherlands, International Association for the Evaluation of Educational Achievement.

Selection of Subscales

Attitudes are part of the process through which students construct knowledge and develop abilities. Because students' attitudes can have a predictive and explanatory value, researchers are interested in studying them in educational settings (i.e., CivEd, 1999; TIMSS, 2003). The attitude subscales from the CivEd survey included items to measure students' degree of adherence to common values and attitudes toward women's rights (G), immigration (H), and political activism (J),

- Subscale G: Attitudes toward Women's Political and Economic Rights
- Subscale H: Positive Attitudes Toward Immigrants
- Subscale J: Confidence in School Participation

Subscale G

The items in this measure reflect the attitudes of students toward rights for women, minorities, and anti-democratic groups. The Civics Education Study technical report provided in page 110 the standardized maximum likelihood estimates for the international sample (RMSEA = .052, AFGI = .96, NNFI = .93, and CFI = .94). From these three dimensions or factors, only the factor measuring desired rights or opportunities for women, consisting of six items, was retained. Scale reliability estimate for the U.S. sample was α = .82.

Subscale H

The items in this measure reflect the attitudes of student towards immigration. This factor showed items having poor fit, and some items were discarded due to poor item reliability. This factor retained 5 items for scaling. The Civics Education Study technical report provided in page 112 a scale reliability estimate for the U.S. sample as α = .85.

Subscale J

The items on this subscale measured school participation. The Civics Education Study technical report stated that not all the scale original items contributed to one factor dimensionality satisfactorily, as indicated by the poor fit (RMSEA = .117, AGFI = .89, NNFI = .77, CFI = .84) in the international sample. A two-factor solution showed better fit but because only three items loaded onto one factor, and one item loading to two factors, only one dimension was retained with 4 items. The scale reliability estimated reported in page 114 of the technical report, for the U.S. sample was α = .85.

Sample and Data Generation

Data for the simulation study was generated from the item parameters of scales G, H, and J of the Civics Education Study, which was administered to a standard population of 2811 students from 124 schools in the United States. The standard population was operationalized as that of full time 9th grade students, the grade in which most 14-year olds were at the time of testing. The items of these three subscales used a Likert-type response format using a four-point scale ranging from *strongly disagree, disagree, agree, and strongly agree*. To create the subscales' samples from which the generating item parameters were estimated, observations with items coded 8 (unit nonresponse) and 9 (item nonresponse) were deleted. Additionally, observations with items scored 0, a "don't know" option included in each item, were eliminated from the analysis. Some items had a negative construction and were reverse-scored. Tables C1-C3 in the appendix C show the frequencies of each of these options per scale.

Simulated normally distributed item response data *N*(0, 1) for the DIF analyses were generated using the IRTGEN macro, to conform the GRM for three test lengths (4, 5, and 6 items) with four Likert-type response categories. This macro generated the study's response data by reading the item parameter values (Table 6); item parameter values were calibrated using MULTILOG 7.03 (Thissen, 2003).

Because the quality of a study is as good as the quality of the data, it was of interest to evaluate / validate that the data generated using the IRTGEN macro fit the GRM. Theoretically, as explained in previous sections, the GRM within the IRT framework is an appropriate calibration and scoring procedure for items with ordered, Likert-type categories. However, empirical evidence that the GRM fits the CivEd data from the subscales selected was also required to ensure that the advantages of selecting the IRT GRM model were realized. To this end, graphical and statistical model-data fit analyses were conducted. Appendix A presents the graphical and χ^2 results of such model-data fit, which indicated sufficient model-data fit. That is, the comparison of the empirical and predicted IRF's and χ^2 values provided evidence of both the appropriateness of the GRM as a calibration and scaling procedure for the generated data and the satisfactory evidence of dimensionality.

Table 6

True Item Parameter Estimates ~$N(0, 1)$

Item	Subscales	α	β_1	β_2	β_3
	Confidence in school participation *N*=2164				
J1		1.9659	-2.1029	-1.3023	0.6874
J2		**2.6903**	-2.1195	-1.3569	0.4638
J3		2.4593	-2.2623	-1.3512	0.6071
J5		1.8265	-2.4991	-1.6811	0.2184
	Positive attitudes toward immigrants *N* = 2125				
H1		2.0993	-1.8050	-1.0576	0.7632
H2		**3.0527**	-2.0724	-1.4945	0.0331
H3		2.1249	-2.0999	-1.1814	0.5905
H4		2.6479	-1.9730	-1.1813	0.4257
H5		2.8120	-1.9703	-1.1506	0.2442
	Support for women rights *N* = 2104				
G1		2.4350	-2.1676	-1.6674	-0.0712
G4		**2.5039**	-2.2722	-1.6366	-0.5027
G6		2.7513	-1.9685	-1.5006	-0.4952
G9		1.5371	-2.3467	-1.2342	0.0229
G11		1.7993	-2.6009	-1.8004	-0.3803
G13		2.4786	-1.8894	-1.2225	-0.1894

Note: Items are scored on a four-point Likert-type scale: 1 = strongly disagree; 2 = disagree; 3 = agree; 4 = strongly agree. The most discriminating items in each scale are bolded and were used as anchor items for the free-baseline approach to the IRT-LR DIF method.

The process of extracting the response data from the selected subscales is presented. The CivEd student data file contains students' responses to all scales' items that compose the survey section. Response data from the selected subscales were extracted and a final SAS data file was created, containing all the response data from the selected subscales. Once the response data from the selected subscales were extracted, item parameters were calibrated or estimated using MULTILOG 7.03 (Thissen, 2003) to fit the GRM. Syntax for item-parameter estimation (*MLG) is provided in Appendix B.

Once model-data fit was confirmed, the item parameter estimates were used to generate simulated item response data with the IRTGEN SAS® macro (Whittaker et al., 2003). IRTGEN simulates item response data for the polytomous two-parameter graded response model using Dodd, De Ayala, and Koch (1995) equation for computing the probabilities of responding in a response category through a two steps process,

1. Simulees are randomly assigned to a known theta (θ) value which will act as simulees' trait level. Then, theta (θ) and user provided item parameters are used to compute the probability of a simulee responding in each GRM response category (Dodd et al., 1995),

$$P_{\mathrm{ix}}^{*}(\theta) = \frac{\exp[a_{\mathrm{i}}(\theta - b_{\mathrm{ix}})]}{1 + \exp[a_{\mathrm{i}}(\theta - b_{\mathrm{ix}})]},$$

2. Random numbers are drawn from a uniform distribution so random error is introduced into simulee's response by comparing these random numbers to the sum of cumulative probabilities computed in step 1.

$$P_{\mathrm{ix}} = P_{\mathrm{ix}}^{*}(\theta) - P_{\mathrm{i,x+1}}^{*}(\theta).$$

A response category is endorsed if the random number is below or at the cumulative probability for a certain category.

Because it was of interest to evaluate the data generated by IRTGEN, in terms of accuracy and precision, a preliminary item parameter recovery study was conducted (see appendix C). The obtained item parameter estimates yielded by calibrating the item response data generated with IRTGEN were compared with the baseline or starting item parameter estimates observed by calibrating the original data from the CivEd (1999) study.

Evaluation of the Data Generation Process

The design of an IRT simulation study should include factors relevant to the task at hand. One critical factor when an IRT response model is used is that of the sample size (de la Torre, & Hong, 2010; Kieftenbeld & Natesan, 2012). As De Ayala and Sava-Bolesta (1999) and DeMars (2003) stated, it remains difficult to propose a sample size for efficient parameter estimation; the number of item categories when a polytomous IRT model is used makes the selection of an adequate sample size for efficient parameter estimation complicated too. Moreover, the power of the IRT Likelihood Ratio (IRT-LR) test for evaluating test items for DIF might be affected by the effectiveness of the data generation program and/or by the effectiveness of the imputation methods in preserving the items' parameter invariance property. Kieftenbeld and Natesan (2012) stated that "accurate recovery of model parameters from response data is central in item response theory (IRT)" (p. 399); thus, a preliminary simulation study was conducted with the purpose of determining the effectiveness of the program used to generate the item response datasets (i.e., IRTGEN), as well as to determine the sample size, the number of items, the effectiveness of the missing data methods, and the number of replications needed to minimize sampling error. Bias and root mean squared error (RMSE) were the outcomes for estimating the accuracy and precision of the GRM item parameter estimates across these factors. The results of this preliminary study are located in Appendix C. Figure 7 summarizes all the process from accessing the CivEd data to the imposing of missing data.

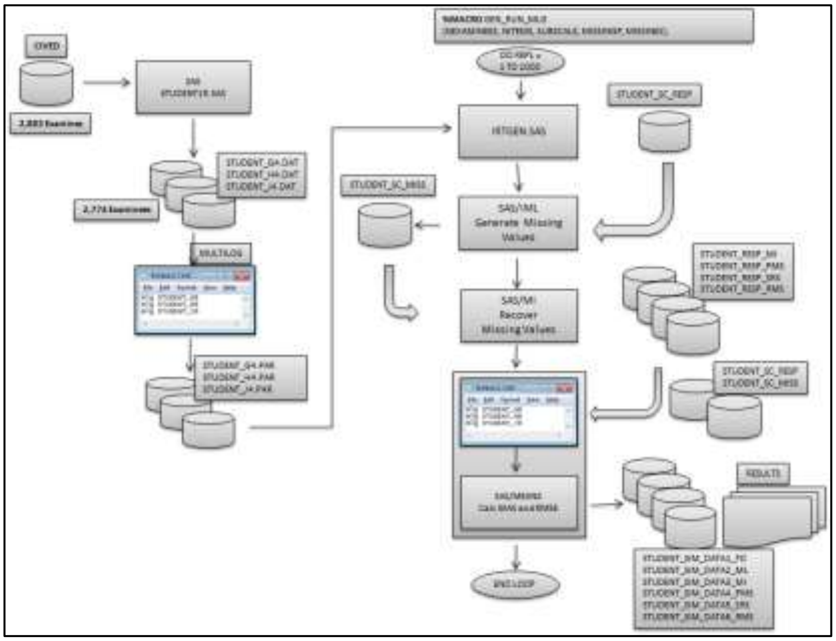


Figure 7. Flow process from accessing the CivEd data for the national sample, extraction of the subscales for analysis to the process of calibrating item parameters using MULTILOG and generating the response data with IRTGEN. The manipulation of missing persons and missing items is shown.

DIF Simulation Study

Eight factors were manipulated in this study: 1) missing data methods, 2) sample size, 3) magnitude of DIF, 4) proportion of missing observations (mP), 5) proportion of missing items (mI), 6) scale length, 8) level of significance or alpha (α), and 9) ability group distributions. Table 7 summarizes the study design factors.

Table 7

Missing Data Simulation Matrix: $\theta_R \sim N(0, 1)$: $\theta_F \sim N(0, 1)$ and $\theta_R \sim N(0, 1)$: $\theta_F \sim N(-.5, 1)$ *at* α = .01 *and* .05 *Across Six Missing Data Methods and a Complete Data Case.*

			DIF Magnitude							
			Δb = 0		Δb = 0.25		Δb = 0.50		Δb = 0.75	
			MissingI		MissingI		MissingI		MissingI	
$n_{F,}$ n_R	Test Length	MissingP	~ .20 (1 item)	~ .40 (2 items)	~ .20 (1 item)	~ .40 (2 items)	~ .20 (1 item)	~ .40 (2 items)	~ .20 (1 item)	~ .40 (2 items)
250/250 500/500	4	.10								
		.20								
		.40								
	5	.10								
		.20								
		.40								
	6	.10								
		.20								
		.40								
200/300 400/600	4	.10								
		.20								
		.40								
	5	.10								
		.20								
		.40								
	6	.10								
		.20								
		.40								

Note: There are four between-subjects factors: Sample size, tests length or number of items, proportion of missing observations or missing data by persons (mP), and proportion of missing items (mI). In addition to four levels of DIF magnitude, between-subject factors are also crossed with 6 missing data methods and a complete data case, and two levels of significance at *α* = .01 and .05, and two levels of ability distribution. This factorial design (4 x 3 x 3 x 2 x 4 x7 x 2 x 2) had a total of 8064 conditions.

Between-Subjects Factors

Sample Size

Two total sample sizes were simulated. A sufficiently large total sample size (e.g., *N*=1000) to allow for stable parameter estimates and a minimum total sample size (e.g., *N*=500) for implementing the GRM was also simulated. A DIF analysis encompasses both a reference (R) and a focal group (F), of which the focal group tends to be smaller. That is, in empirical research it is usually the case to have an unbalanced sample when carrying out DIF analyses, in which the focal group is normally smaller than the reference group. Thus, in addition to balanced samples (sample size ratio 1:1), unbalanced samples (sample size ratio 3:2) were generated:

- Balanced sample sizes (1:1 sample size ratio)

  n_R=250: n_F=250 and n_R=500:n$_F$=500),

- Unbalanced sample sizes (3:2 sample size ratio)

  n_R=300:n_F=200 and n_R=600:n_F=400)

Proportion of Missing Observations (mP)

The proportion of observations or cases having missing data (i.e., 10%, 20%, and 40%). in the two total sample size conditions were manipulated as follows,

- *N*=500

  10% missing observations (*m*P = 50 observations missing)

  20% missing observations (*m*P = 100 observations missing)

  40% missing observations (*m*P = 200 observations missing)

- *N*=1000

  10% missing observations (*m*P = 100 observations missing)

  20% missing observations (*m*P = 200 observations missing)

  40% missing observations (*m*P = 400 observations missing)

Number of Missing Items (mI)

Item missingness was generated by deleting ~20% and ~40% of items in each scale, which was 1 and 2 items, respectively.

- Scale J (4 items)

  1 item missing (J=3)

  2 items missing (J=2)

- Scale H (5 items)

  1 item missing (H=4)

  2 items missing (H=3)

- Scale G (6 items)

  1 item missing (G=5)

  2 items missing (G=4)

Within-Subjects Factors

Missing Data Methods

The effect of six missing data methods on DIF were investigated in this simulation study: FIML, multiple imputation (MI), person mean substitution (PMS), single regression substitution (SRS), relative mean substitution (RMS) and Listwise deletion. DIF under complete data (as if no missing data was present) was also evaluated for comparison.

Ability Group Distributions (Impact)

- No impact θ_R~N (0, 1) / θ_F~N(0,1)
- Small impact θ_R ~N(0, 1) / θ_F~N(-.5, 0)

DIF Magnitude

For the null conditions, response data for both the reference and focal groups will be generated using the same item parameters (Table 6). For DIF conditions, a shift up in the item parameters thresholds of last item in each scale for the focal group ($\Delta b_{jk\mathrm{F}}$) will be imposed to simulate small (0.25), moderate (0.50), and large (0.75) magnitude of DIF, assuming an equal shift across *all* b_{jk} parameters (i.e., uniform DIF) of the DIF item as shown in Table 8. That is,

$$b_{\mathrm{jkR}} = b_{\mathrm{jkF}} + 0.25 \text{ for each } k$$

where k = item response categories

Table 8

Item Parameter Modification to Simulate DIF

Item		Generating Parameters			DIF Modification		
	α	b_{jk1}	b_{jk2}	b_{jk3}		Δb_{jk} =.25	
J5	1.8265	-2.4991	-1.6811	0.2184	-2.2491	-1.4311	0.4684
H5	2.8120	-1.9703	-1.1506	0.2442	-1.7203	-0.9006	0.4942
G13	2.4786	-1.8894	-1.2225	-0.1894	-1.6394	-0.9725	0.0606
						Δb_{jk} =.50	
J5	1.8265	-2.4991	-1.6811	0.2184	-1.9991	-1.1811	0.7184
H5	2.8120	-1.9703	-1.1506	0.2442	-1.4703	-0.6506	0.7442
G13	2.4786	-1.8894	-1.2225	-0.1894	-1.3894	-0.7225	0.3106
						Δb_{jk} =.75	
J5	1.8265	-2.4991	-1.6811	0.2184	-1.7491	-0.9311	0.9684
H5	2.8120	-1.9703	-1.1506	0.2442	-1.2203	-0.4006	0.9942
G13	2.4786	-1.8894	-1.2225	-0.1894	-1.1394	-0.4725	0.5606

Note: DIF is simulated only in the last item of each scale for the focal group by adding .25, .50 and .75 to the designated DIF item, thus simulating negligent, moderate, and large DIF respectively. DIF is simulated to be uniform; thus, only location parameters are modified.

Computer Software

SAS 9.4 was used to execute the simulation, generating 1000 samples or replications (nR) for each condition in the study. The use of 1000 samples provided a maximum standard error (*SE*) for any proportion estimated no greater than .015 and a confidence interval width ±.03 around observed rates of Type I error (Robey & Barcikowski, 1992).

The simulated data for this study was fitted to the GRM, which as its name implies, models graded responses (e.g., polytomous, Likert-type scored items). The MULTILOG 7.03 (Thissen, 2003), a computer program "for the analysis and scoring of items with MULTIple alternatives" (p. 345), was used to run the DIF analyses. This computer program provided baseline item parameter estimates for data generation and tests of the null hypothesis about DIF using the G^2 statistics, which is the chi-square index of the goodness-of-fit Likelihood Ratio (LR) test. Rather than allowing MULTILOG run using the default number of cycles, the number of cycles used to run the DIF analyses was set to 300 since they allowed reasonable parameter estimates (i.e., converge). The SAS IRTGEN macro (Whittaker et al., 2003) was used to generate data response for the GRM for three scale lengths with four Likert-type categories.

Missing Data Mechanism

Finally, the missing data mechanism was constant. That is, only data missing completely at random (MCAR) was simulated in the present study. Figure 8 illustrates the process followed to implement MCAR. As shown in the Figure 8, a random number between 0 and 1 was allocated to each observation or person. Missingness was simulated for a person given this random number (e.g., the person was selected for missingness if the corresponding random number was less or equal to the percentage of missing observations targeted (i.e., MISSINGP). Once a person was selected to have missing data, the same process was implemented for each item. That is, a vector with random number was generated for each item; if the random number allocated to each item was less or equal than the percentage of missing targeted (MISSINGI), the item response was set to missing (".").

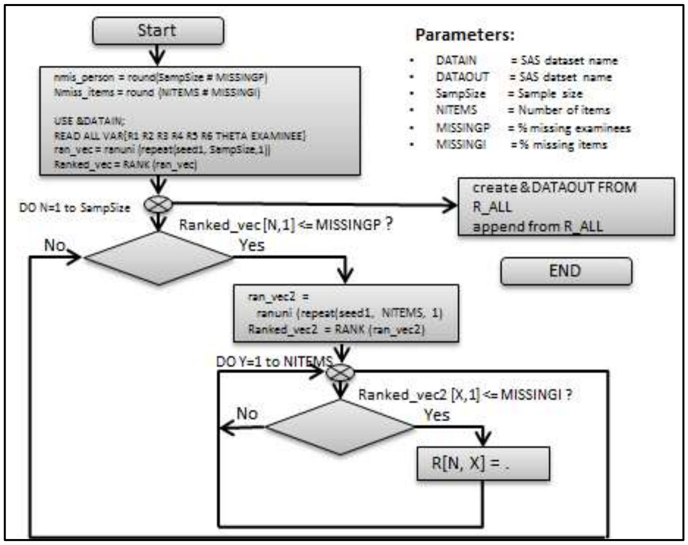

Start
Parameters:
DATAIN = SAS dataset name
DATAOUT = SAS datset name
SampSize = Sample size
NITEMS = Number of items
MISSINGP = % missing examinees
MISSINGI = % missing items
nmis_person = round(SampSize # MISSINGP)
Nmiss_items = round (NITEMS # MISSINGI)
USE &DATAIN;
READ ALL VAR{R1 R2 R3 R4 R5 R6 THETA EXAMINEE}
ran_vec = ranuni (repeat(seed1, SampSize,1))
Ranked_vec = RANK (ran_vec)
DO N=1 to SampSize
create &DATAOUT FROM R_ALL
append from R_ALL
Ranked_vec [N,1] <= MISSINGP ?
No
Yes
END
ran_vec2 = ranuni (repeat(seed1, NITEMS, 1)
Ranked_vec2 = RANK (ran_vec2)
DO Y=1 to NITEMS
Ranked_vec2 [X,1] <= MISSINGI ?
No
Yes
R[N, X] = .

Figure 8. MCAR missing data generation.

The proportion of missing observations was crossed with the proportion of missing items as shown in Table 7.

Analytical Plan

The effect of MDM on the effectiveness of the IRT- LR test for detecting DIF will be evaluated for each combination of factors' conditions (Table 7) across the six MDM and complete data in terms of Type I error rates and statistical power. For the detection of uniform DIF, in which DIF is present only in the location parameter (*b*), item parameter invariance holds if

$$a_{\text{jR}} = a_{\text{jF}} \text{ and } b_{\text{jkR}} = b_{\text{jkF}}$$

That is, there is no change in the discrimination parameter for both the reference group and focal group and the location parameter *b* is invariant for both groups. Thus, the null hypothesis of no DIF for the GRM is stated as,

$$H_0 \colon \alpha_{\mathrm{j}R} = \alpha_{\mathrm{j}R}\ ;\ and\ \beta_{\mathrm{jkR}} = \beta_{\mathrm{jkF}} \text{ for all } k \text{ of item } j\ ,$$

and the alternative is

$$H_1 \colon \text{not all parameters for item } j \text{ are group invariant}$$

The implementation of the IRT-LR test for detecting DIF in polytomous items using a free-baseline approach implies the testing of the null hypothesis (H_0) by comparing the fit of successive nested, paired models, one for each studied item (i.e., augmented model against baseline model), to determine "whether the additional parameters in the augmented model are significantly different" (Thissen et al., 1993).

Type I Error (Rejection of the Null Hypothesis)

For testing the null hypothesis (i.e., no DIF items or DIF effect size = 0), the same parameters were used to generate item response data for both the reference and focal groups (see Table 6). The robustness of the IRT-LR test for controlling Type I error was tested at the nominal α =.01 and α = .05 significance levels. However, because the baseline approach to the IRT-LR for DIF implies a series of comparisons of model fit of the augmented models to the compact model, a Bonferroni correction to the significance level was implemented to avoid the inflation of the nominal alpha. The decision whether H_0 is rejected for the studied item was made on the basis of comparing the G^2 statistic to both χ^2 and Bonferroni critical values. Table 9 shows the tabled χ^2 and corrected Bonferroni critical values for α =.01 and α = .05.

Table 9

Rejection Criteria: Adjusted P, and χ2 and Bonferroni Critical Values for the Evaluation of the Null Hypothesis of No DIF (4 df)

	χ2 Critical Values		Bonferroni χ2 Critical Value	
Subscale	0.01	0.05	0.01	0.05
G	13.28	9.49	16.92	13.28
H	13.28	9.49	16.42	12.76
J	13.28	9.49	15.78	12.09

Note. For each of the IRT-LR tests, the Bonferroni corrections to the *p*-values depend on the subscales' number of items. The Bonferroni χ2 critical values are estimated on the Bonferroni corrected *p*-values. χ^2 critical values do not change across subscales because they are computed based on the same number of degrees of freedom (4).

In DIF analysis, Type I error is the incorrect identification of an item as displaying DIF when it is not a DIF item (i.e., rejection of the null hypothesis). The free-baseline approach to the IRT-LR test for DIF detection compares G^2 differences for a series of nested models with respect to a χ^2 critical value with degrees of freedom (*df*) equal to the difference in the number of estimated parameters at the specified critical values (e.g., for this study, $\chi^2_{(4)}$=13.28 and $\chi^2_{(4)}$=9.49 at α =01 and α =.05 respectively) used as measures of statistical significance. In a simulation study, this multiple significance testing occurs in a single experiment or replication. Thus, following the χ^2 and Bonferroni adjustment for the significance testing of each item, a Type I error was computed as a familywise error rate (FWER) per experiment (sample or replication).

Familywise Type I Error Rate

For this study, the null conditions were established by using the same item parameter estimates for generating the response data for both the focal and reference group. Thus, for the null conditions, DIF = 0, meaning that all hypothesis tested within a set of items were considered true. Each item in a scale, was evaluated for DIF by comparing the observed value of the G^2 statistic to the χ^2 and Bonferroni adjustment critical values (i.e. cutoff criteria) at the specified significance levels. If the change in the observed value of the G^2 exceeded these critical values (see Table 9) thus indicating an unlikely value, the null hypothesis of no DIF was rejected for a given item. Type I error rates were computed as any-hypothesis test or FWER, per experiment (each null condition in the study), as recommended by Ryan (1959; p. 54), over the number of samples, at two levels of significance ($\alpha = .01$ and $\alpha = .05$). That is, the probability of at least one Type I error (Toothaker, 1991; Ryan, 1959) within an experiment, is the FWER or,

$$\alpha_{\mathrm{FW}} = \frac{k_I}{k}$$

where

$k_{\mathrm{I}} =$ number of samples in which at least one test of the null hypothesis was rejected

k = number of samples

In the case of the MI method, 5 imputed datasets were generated. For explanation purposes, an example of the process for estimating Type I in the context of multiple testing and averaging across imputations for 10 samples or replications is shown in Table 10.

Table 10

Hypothesis Testing in the Context Multiple Testing for Multiple Imputation. Averaging Type I Error Rates across Imputations

	Imp_1				Imp_2			→	Imp _5			
*n*R	H_{01}	H_{02}	H_{03}	α_{FW}	H_{01}	H_{02}	H_{03}	α_{FW}	H_{01}	H_{02}	H_{03}	α_{FW}
1	0	0	0	0	1	0	1	1	0	0	0	0
2	1	0	0	1	0	0	0	0	1	1	0	1
3	0	0	0	0	1	0	0	1	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0
6	1	0	0	1	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0
10	1	1	0	1	0	1	0	1	0	0	0	0
				.03				.03				.01
												.02

Note. Within each imputation, three items were tested for DIF (items were flagged 1 if the DIF test was significant). Type I error was estimated per experiment (sample or replication). Within each imputation, Type I error per experiment (α_{FW}) was estimated estimated by dividing number of samples in which at least one test of the null hypothesis was rejected by the number of samples or replications (*n*R). In this example, the Type I error rate for the condition simulated was .02.

Overall Distributions of Familywise Error Rates

After the FWER was estimated across the study conditions (see Table 7), the overall distribution of the rejection rates was examined using boxplots, which allowed observing graphically side by side the distribution shapes of the rejection rates across all factors by MDM and significance level. This examination helped comparing at a glance, for example, if the distributions for a particular MDM tended to skew toward higher levels of rejection.

Impact of Study Factors on Familywise Error Rates

To analyze the total variability on the FWER accounted by the interaction of factors, the effect size η^2 was computed for all first-order interactions by method, to determine which interactions contributed significantly to the variability of the rejection rates. When none of the first-order interactions had a significant effect size, further analyses were examined for those main effects having a moderate to large effect size ($\eta^2 \geq .0588$) if any was found. For those interactions or main effects showing an effect size on FWER, bar graphs were constructed for the mean point estimates of statistically significant interactions or main effects.

Statistical Power

Cohen (1992) defined the power of a statistical test of the null hypothesis as "the probability that the H_0 will be rejected when it is false" (p. 98). Within the context of multiple comparisons, as explained in the previous section, statistical power of the free-baseline IRT-LR test was estimated as per test power; then, for a meaningful power analysis, power analyses were conducted only for those conditions having adequate Type I error control by Bradley's criterion (1978). For Bradley's implementation at α =.01 and α =.05 power was estimated for the proportion of conditions rejecting the null hypothesis if they were in the following ranges,

If rejection rate at .01 is > .005 and < .015 then Type I error control is adequate

If rejection rate at .05 is > .025 and < .075 then Type I error control is adequate

Overall Distributions of Statistical Power

Once the conditions over which the IRT-LR test is robust (i.e., adequate Type I error control) are determined by Bradley's criterion (1978), only these conditions were used to examine the overall power of the IRT-LR test for detecting DIF across all conditions by MDM using box plots.

Comparison of Statistical Power across MDM

For power comparisons across MDM, only those conditions over which all MDM had adequate type I error control by Bradley’s criterion were considered. Power comparisons were displayed using boxplots.

Impact of Study Factors on Statistical Power

An analysis of effect size (η^2) on all main effects and first-order interactions were conducted to determine which factors contributed significantly to the variability of power estimates within each MDM. For those factors and interactions showing a significant effect size on statistical power, that is, having a moderate to large effect size ($\eta^2 \geq .0588$), bar graphs were constructed for the mean point estimates of each statistically significant factor or interaction.

CHAPTER FOUR

RESULTS

This study compared the effect of six missing data methods (MDM) on the performance of the IRT-LR test for detecting DIF in polytomously scored items, in terms of Type I error and statistical power. Research question 1 is addressed by presenting first the overall Type I error rates of each MDM across all simulation conditions for both χ^2 and Bonferroni adjustment critical values at $\alpha = .01$ and $\alpha = .05$. To analyze simulation conditions with adequate Type I error control, Bradley's (1978) criteria for robustness was implemented and the extent of the relationship of the research factors and familywise error rates under each MDM were estimated on those conditions having adequate Type I error control. Type I error results are presented in terms of effect size estimates (η^2) for first-order interactions or main effect. Then, research question 1 is summarized. Next, research question 2 is addressed by presenting first the overall statistical power of each method across all simulation conditions for both χ^2 and Bonferroni adjustment critical values at $\alpha = .01$ and $\alpha = .05$. Then, distributions of power estimates by method only for the conditions with adequate Type I error control are examined using boxplots. Next, methods are compared by comparing only those methods that had power over the same conditions and the extent of the relationship of each research factor and power estimates under each MDM are presented. Lastly, research question 2 is also summarized.

Research Question 1. Effect of Missing Data one Type I Error Rates

As explained in the Chapter 3, once the significance testing comparisons of the nested hypotheses for DIF detection were conducted, Type I error rates were computed as any hypothesis test or FWER; that is, the probability of at least one Type I error in a group of significance tests, as recommended by Toothaker (1991) and Ryan (1959), over the number of samples, at two levels of significance (α = .01 and α = .05). The overall distributions of familywise error rate (FWER) estimates at α = 01 for both critical χ^2 and Bonferroni adjustment for all missing data methods across all simulation conditions are presented in Figure 8. As displayed in Figure 8, the overall distributions of the FWER estimates showed an inflated error rate across all missing data methods when $\chi^2_{(4)}$=13.28 was applied to testing for DIF in each item within a subscale at α = 0.1 (*M* = .04 to *M* = .07). Multiple imputation (MI) and single regression substitution (SRS) missing data methods (MDM) showed the larger mean familywise error rates (*M* = .07 and *M* = .06 respectively), with SRS showing the largest upward dispersion (max = .185), followed by MI (max = .149) and RMS (max = .015).

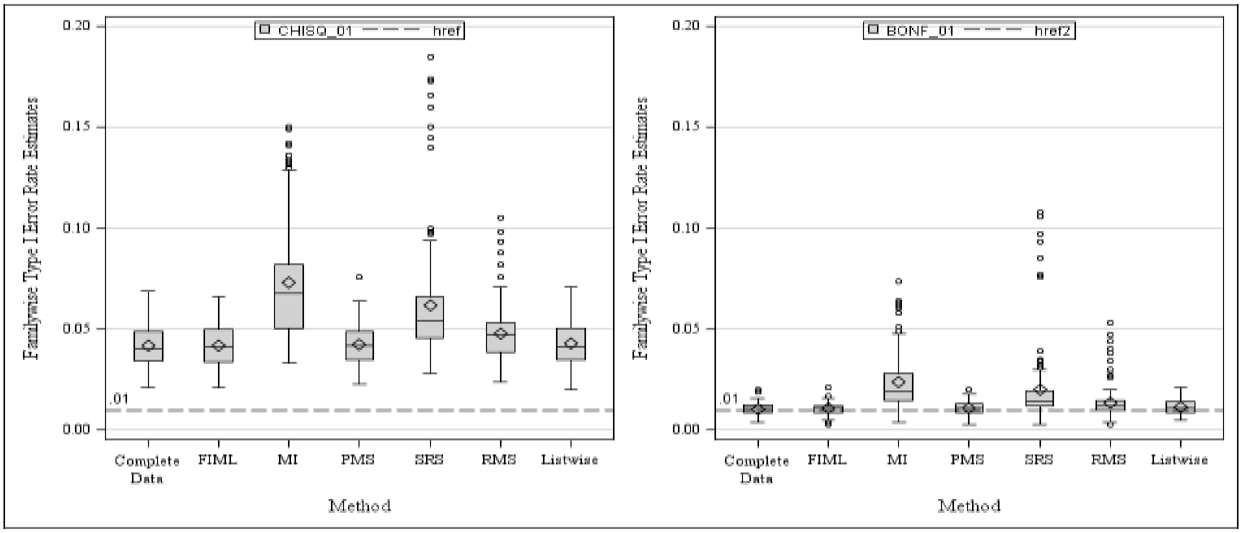


Figure 9. Overall familywise error rate distribution by missing data method and Critical χ2 and Bonferroni adjustment respectively. Estimates are based on 1000 samples of each simulation condition and tests conducted at the nominal α = .01 significance level

In contrast, the use of the Bonferroni adjustment for the testing of DIF for each individual item (see Table 9) resulted in mean FWER at the nominal alpha level (α = .01) for the complete data, FIML, PSM, RMS, and Listwise deletion missing data methods. Both MI and SRS had a slightly higher mean FWER (*M* = .02) but larger upward dispersions (max =.074 and max = .106 respectively). While the RMS method had a slightly higher mean FWER (*M* = .02) than the nominal alpha, it also showed greater upward dispersion (max = .053) than that for the methods with mean FWER at the nominal level. The extreme values (i.e., max.) or outliers for the MI, SRS, and RMS methods (for both χ^2 and Bonferroni adjustment) suggested the impact of factor designs. Bradley' (1978) criteria for robustness was implemented to investigate the effect of MDM and their treatments across simulation factors on those simulations conditions having adequate Type I error control. As explained, these initial inspections of Type I error showed an inflated FWER across all methods when hypothesis tests of significance were conducted using the χ^2 critical value at alpha .01. Under Bradley's criteria for robustness, none of these conditions had an adequate Type I error control and further analyses were not conducted for these conditions.

To What Extent was the Effect of Missing Data Consistent when α = .01?

To respond research question 1 regarding the effect of missing data on Type I error rates of the Likelihood Ratio test for DIF detection, effect size η^2estimates were computed across all MDM for all factors to identify the effect of each factor on the total variability observed in the FWER when α = .01 under each MDM (The complete table of η^2 values (Table D1) is presented in Appendix D). The following sections report effect sizes by method if $\eta^2 \geq .05$.

Complete Data (α = .01)

Effect size estimates were computed for all first-order interactions for the complete data method. This method had an overall mean FWER at the nominal level (*M* = .01). Although a slight variability is observed (min = 0.003; max = 0.02), under this MDM, none of the research factors, either as interactions or main effects, had an effect on FWER.

FIML (*α* = .01)

The overall distribution of FWER for the FIML displayed in Figure 8 showed that the FIML method had a mean FWER at the nominal level (*M* = 01). The slight variability observed in the distribution (min = 0.003; max = 0.021) suggested the effect of a factor or factors in the study. An analysis of effect size for first-order interactions showed a moderate effect for the interaction of the proportion of missing observations and ability distribution (η^2 = .05) associated with the variability observed in FWER for FIML method. Figure 9 displays the mean FWER for the interaction effects between ability distribution and the proportion of missing observations.

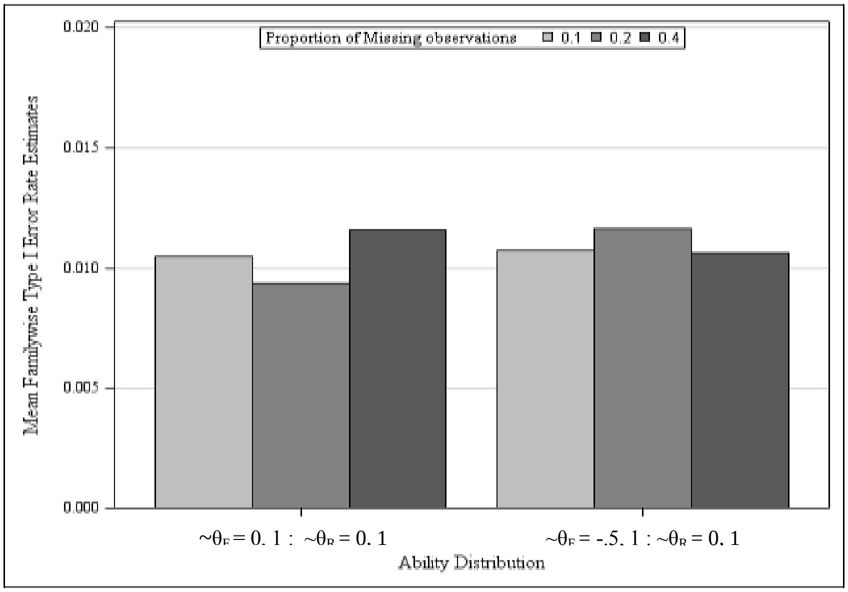


Figure 10. Mean familywise error rates for FIML (α = 01) by ability distribution and proportion of missing observations (η^2 =.05)

As was observed in Figure 10, to the interaction of ability distribution and proportion of missing observations under FIML did not vary greatly. When both focal and reference groups had the same ability distribution (ability distribution = 1 or $\sim\theta_F$= 0, 1 and $\sim\theta_R$=0, 1), the mean FWER were slightly higher when the proportion of missing data was .40. When the focal and reference group differed in ability distribution (ability distribution = 2 or $\sim\theta_F$= -.5, 1 and $\sim\theta_R$=0, 1), the mean FWER was slightly higher when the proportion of missing observations was .02.

Multiple Imputation (α = .01)

The overall distributions of FWER for the MI method displayed in Figure 8 showed that the MI method had an overall FWER slightly above the nominal level (*M* = 02). The variability observed in the overall distribution (min = 0.003; max = 0.021) suggested the effect of a factor or factors in the study. An analysis of the effect size for first-order interactions showed a large effect for the interaction between the proportion of missing observations and the proportion of missing items (η^2 = .12) associated with the variability observed in FWER. Figure 10 displays the relationship of mean FWER and the interaction effects between the proportion of missing observations and the proportion of missing items under MI.

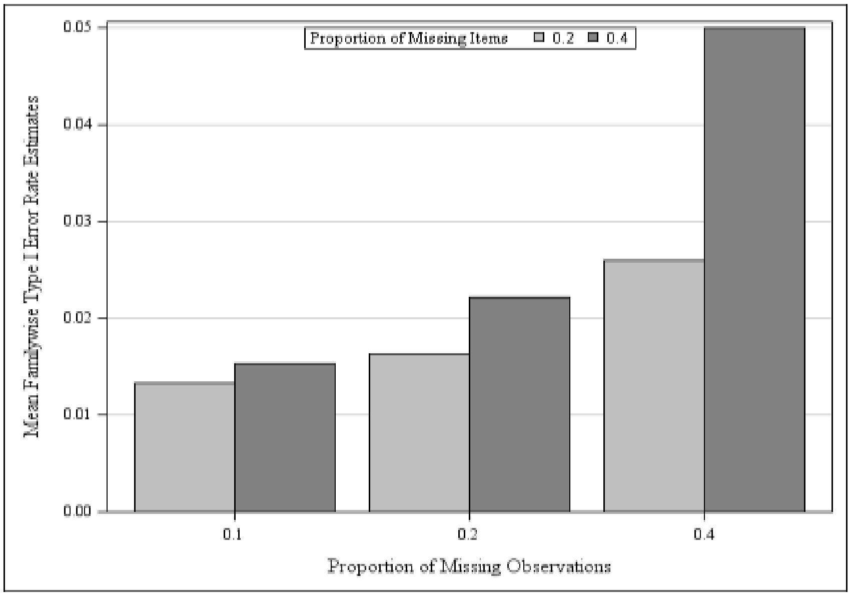


Figure 11. Mean familywise error rates for MI (α = .01) by proportion of missing observations and proportion of missing items (η^2 =.12)

As observed in Figure 10, the proportion of missing observations and the proportion of missing items had a large effect on FWER for the MI method. Mean FWER increased as the proportion of missing observations increased and as the number of missing items increased. When the proportion of missing observations was .10, there was a slight increase in mean FWER from the proportion of missing items .20 to the proportion of missing items .40 (*M* = 0.013 to *M* = 0.015 respectively). In contrast, markedly higher mean FWER were observed from the proportion of missing items = .20 to the proportion of missing items = .40 when the proportion of missing observations was .40 (*M* = 0.025 to *M* = 0.05 respectively).

Person Mean Substitution (α = .01)

Effect size estimates were computed for all first-order interactions for the person mean substitution method. This method had an overall mean family rate at the nominal level (*M* = .01). Although a slight variability in FWER is observed (min = 0.003; max = 0.02), under this MDM, none of the interactions or main effects had a significant effect size.

Single Regression Substitution (α = .01)

The overall distributions of FWER for the SRS displayed in Figure 8 showed that this method had a mean FWER slightly above the nominal level (*M* = 02) but the variability observed in the overall distribution showed very extreme upward values (max = 0.108), suggesting the effect of a factor or factors in the simulation. An analysis of the effect size for first-order interactions under SRS showed large effects for three interactions: 1) the proportion of missing observations and the proportion of missing items (η^2 = .13), 2) the interaction of number of items and proportion of missing observations (η^2 = .11), and 3) the interaction of the number of items and the proportion of missing items (η^2 = .10). Figure 11 to Figure 13 display the mean FWER for these large effect sizes under SRS.

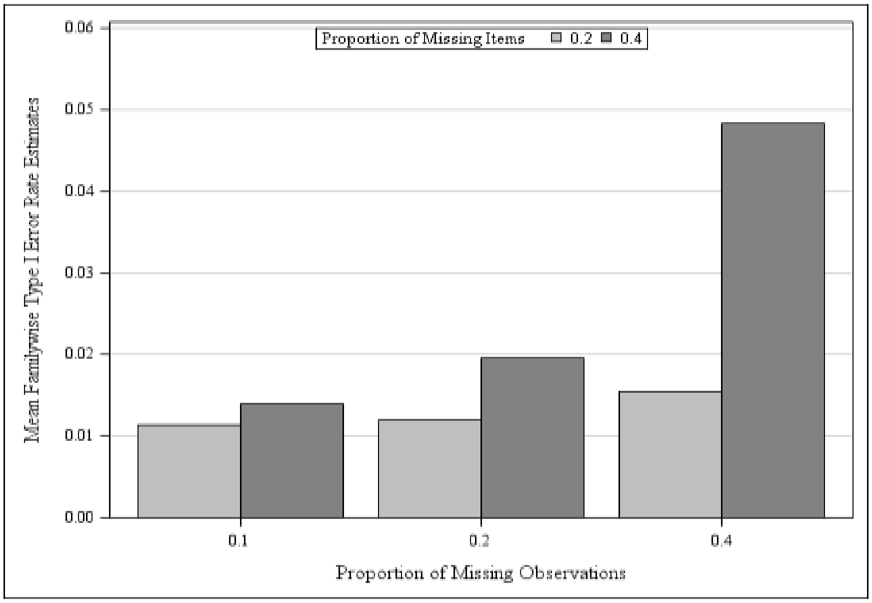


Figure 12. Mean familywise error rates for SRS (α = .01) by proportion of missing observations and proportion of missing items (η^2 =.13)

As observed in Figure 11, the proportion of missing observations and the proportion of missing items had a large effect on FWER for the SRS method. Mean FWER increased as the proportion of missing observations increased and as the number of missing items increased. When the proportion of missing observations was .10, there was a slight increase in mean FWER from the proportion of missing items = .20 to the proportion of missing items = .40 (*M* = 0.011 to *M* = 0.014 respectively). In contrast, markedly higher mean FWER were observed from the proportion of missing items = .20 to the proportion of missing items = .40 when the proportion of missing observations was .40 (*M* = 0.015 to *M* = 0.05 respectively).

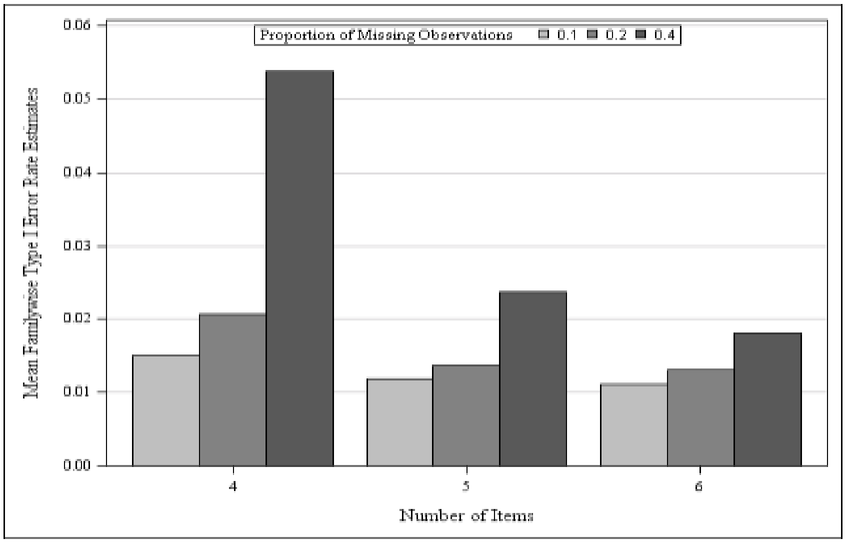


Figure 13. Mean familywise error rates for SRS (α = .01) by number of items and the proportion of missing observations (η^2 =.11)

As observed in Figure 12, number of items and the proportion of missing observations had a large effect on FWER under the SRS method. Mean FWER increased as the proportion of missing observations increased but smaller mean differences were observed as the number of missing items increased. Overall, across all number of items there was a notably greater increase in the mean FWER from missing observations = .20 to missing observations = .40 than there was from missing observations = .10 to missing observations = .20. But there was a marked increase in mean FWER from the proportion of missing observations = .20 to the proportion of missing observations = .40 (*M* = 0.021 to *M* = 0.054 respectively) when the number of items was four.

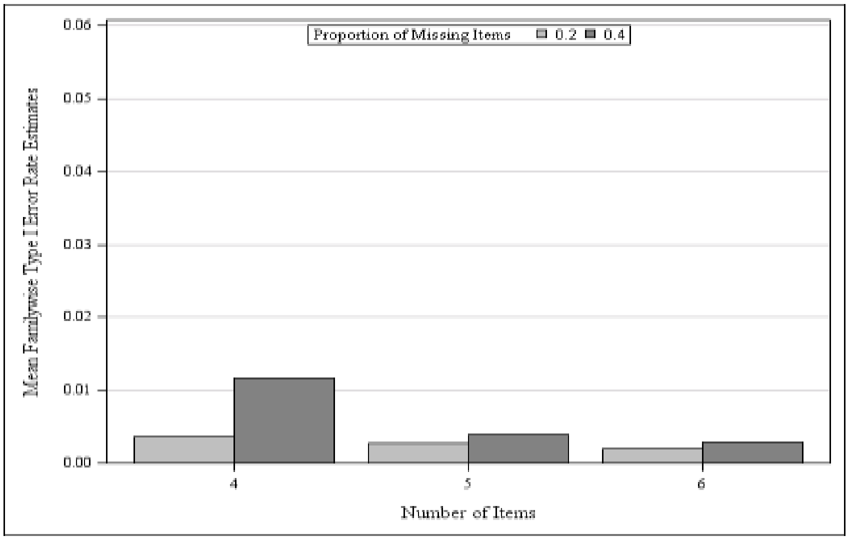


Figure 14. Mean familywise error rates for SRS (α = .01) by number of items and the proportion of missing items (η^2 =.10)

As observed in Figure 13, number of items and the proportion of missing items had a large effect on FWER under the SRS method. Mean FWER increased as the proportion of missing items increased but decreased as the number of items increased. Mean FWER were very similar when the number of items was five and six, with very small mean difference increases observed as the number of missing items increased. There is a notable increase in the mean FWER going from missing items .20 to missing items .40 when the number of items was four (*M* = 0.004 to *M* = 0.012 respectively).

Relative Mean Substitution (α = .01)

The overall distributions of FWER for the RMS displayed in Figure 8 showed that this method had a mean slightly above the nominal level (*M* = 02) but the variability observed in the overall distribution showed very extreme upward values (min = 0.003; max = 0.108), suggesting the effect of a factor or factors in the simulation study. An analysis of the effect size for first-order interactions showed large effects for three interactions: 1) the interaction of the number of items and the proportion of missing items (η^2 = .15), 2) the interactions of number of items and proportion of missing observations (η^2 = .07), and 3) the interaction of the proportion of number of items and the proportion of missing items (η^2 = .07). Figure 14 to Figure 16 display the mean FWER for these effect sizes.

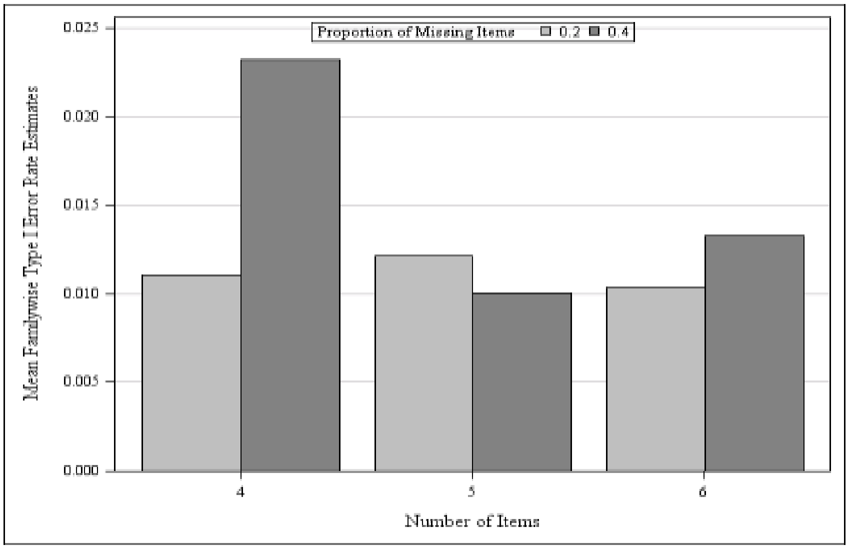


Figure 15. Mean familywise error rates for RMS (α = .01) by number of items and proportion of missing items for (η^2 =.15)

As observed in Figure 14, number of items and the proportion of missing items had a large effect on FWER under the RMS method. Mean familywise error rates were consistent when the number of items was five and six. Markedly, there was a notable increase in the mean

FWER from the proportion of missing items = .20 to the proportion of missing items = 40 (*M* = 0.011 to *M* = 0.023 respectively) when the number of items was four.

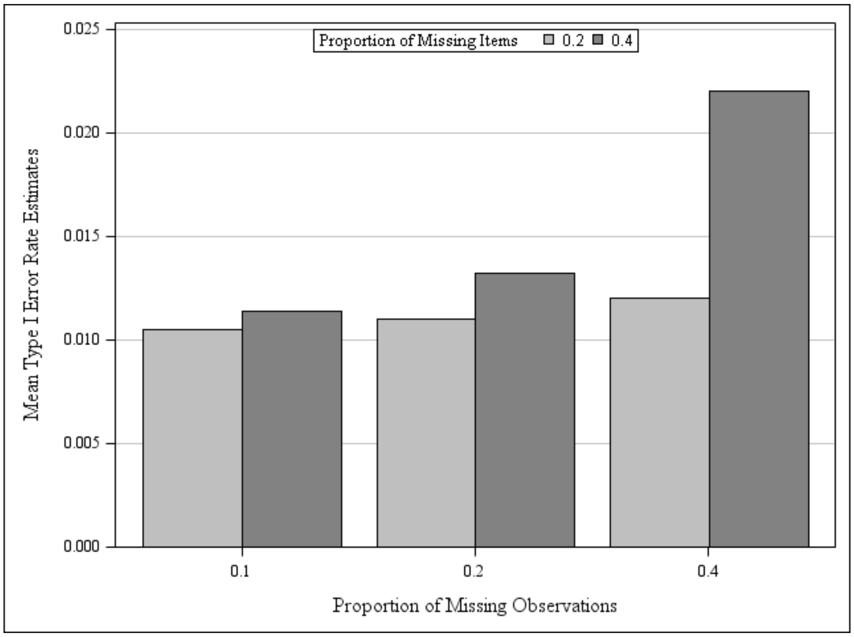


*Figure 16***.** Mean familywise error rates for RMS (α = .01) by proportion of missing observations and proportion of missing items (η^2 =.07)

As observed in Figure 15, the proportion of missing observations and the proportion of missing items had a moderate effect on FWER under the RMS method. Similarly like in the same factors' interaction for the SRS method, in the RMS method the mean FWER increased as the proportion of missing observations increased and as the number of missing items increased. Slight increases in mean FWER for the proportion of missing items were observed when the proportion of missing observations was .10 (*M* = 0.0105 to *M* = 0.0114) and when the proportion of missing observations was .20 (*M* = 0.0105 to *M* = 0.0114). In contrast, markedly higher mean FWER were observed from the proportion of missing items = .20 to the proportion of missing items = .40 when the proportion of missing observations was .40 (*M* = 0.012 to *M* = 0.022 respectively).

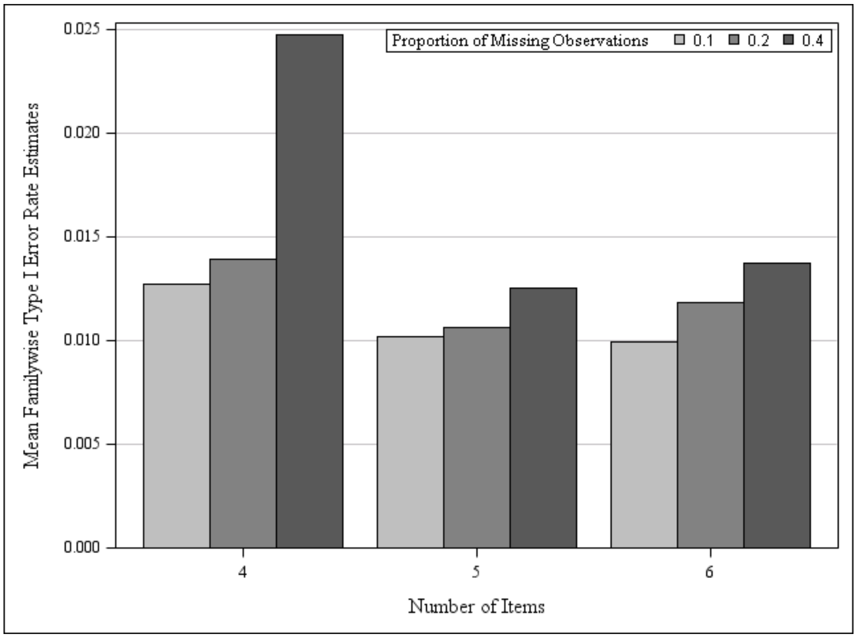


Figure 17. Mean familywise error rates for RMS (α = .01) by number of items and proportion of missing observations (η^2 =.07)

As observed in Figure 16, the interactions of the number of items and the proportion of missing observations under the RMS method had a similar pattern to the interaction for the number of items and the proportion of missing items under the SRS method (Figure 12); that is, mean FWER were consistent when the number of items was five and six and similarly, there was a notable increase in the mean FWER from the proportion of missing items = .20 to the proportion of missing items = 40 (*M* = 0.014 to *M* = 0.025 respectively) when the number of items was four.

Listwise Deletion (α = .01)

Effect size estimates were computed for all first-order interactions and main effects for the Listwise method. Listwise deletion had an overall mean FWER at the nominal level (*M* =

.01). Within this MDM no extreme values were observed (min = 0.005; max = 0.021) and none of the first-order interactions had an effect on FWER. Main effects for the number of items had a moderate effect size (η^2=.07). Figure 17 displays the mean FWER for the number of items.

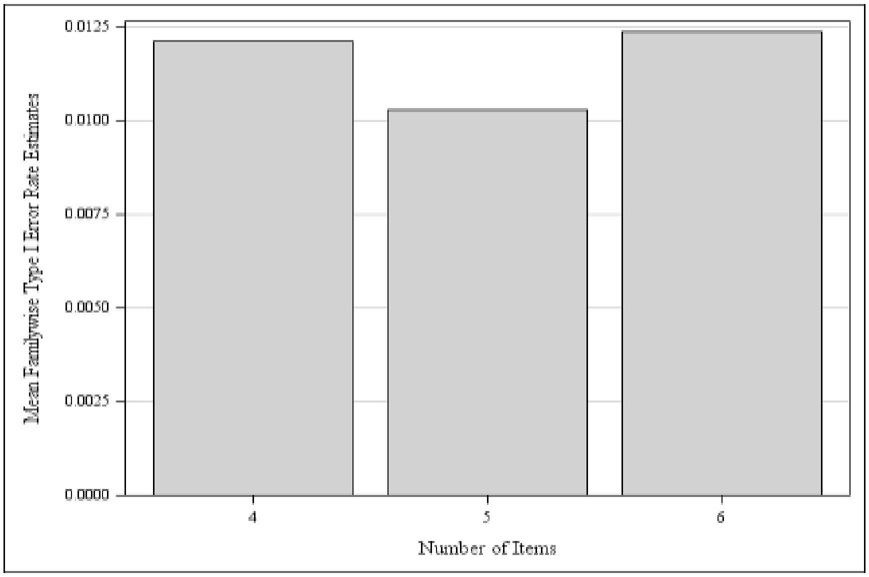


Figure 18. Mean familywise error rates for Listwise deletion (α = .01) by number of items main effect (η^2 =.07)

As observed in Figure 17, the mean FWER were very consistent across the number of items 4 and 6 (*M* = .0121 and *M* =. 0124 respectively) with the mean being slightly lower when the number of items was 5 (*M* = .0103).

Research Question 1 Summary (α = .01)

Inflated FWER resulted across all methods when the significance testing for DIF was conducted using χ^2 critical values at α = .01. In addition to inflated error rates, MI, SRS, and RMS showed also greater dispersion toward higher error rates. On the other hand, when the critical values for significance testing were corrected, using a Bonferroni adjustment, the FWER across all methods resulted in mean distributions around α = .01. However, patterns of upward dispersion of error were also observed for MI, SRS, and RMS under Bonferroni adjustment at α = .01.

An analysis of the effect of each simulation factor on FWER by MDM showed the following results when α = .01:

1. Sample size did not have any effect on FWER for any MDM.
2. None of the simulation factors had any effect on FWER under complete data method.
3. The interaction of the proportion of missing observations and ability distribution showed a moderated effect on FWER under FIML (η^2 = .05)
4. The interaction of number of items and the proportion of missing observations had a moderate effect (η^2 = .07) and large effect (η^2 = .11) on FWER under RMS and SRS respectively.
5. The interaction of the number of items and the proportion of missing items had a large effect large effect on FWER under SRS and RMS (η^2 = .10 and η^2 = .15 respectively).
6. The interaction of the proportion of missing observations and the proportion of missing items had moderate effect (η^2 = 07) on FWER under RMS, and a large effect large effect (η^2 = 12 and η^2 = 13) on FWER under MI and SRS respectively.

To What Extent was the Effect of Missing Data Consistent when α = .05?

The overall distributions of FWER estimates at α = 05 for both critical χ^2 and Bonferroni adjustment for all missing data methods across all simulation conditions are presented in Figure 18.

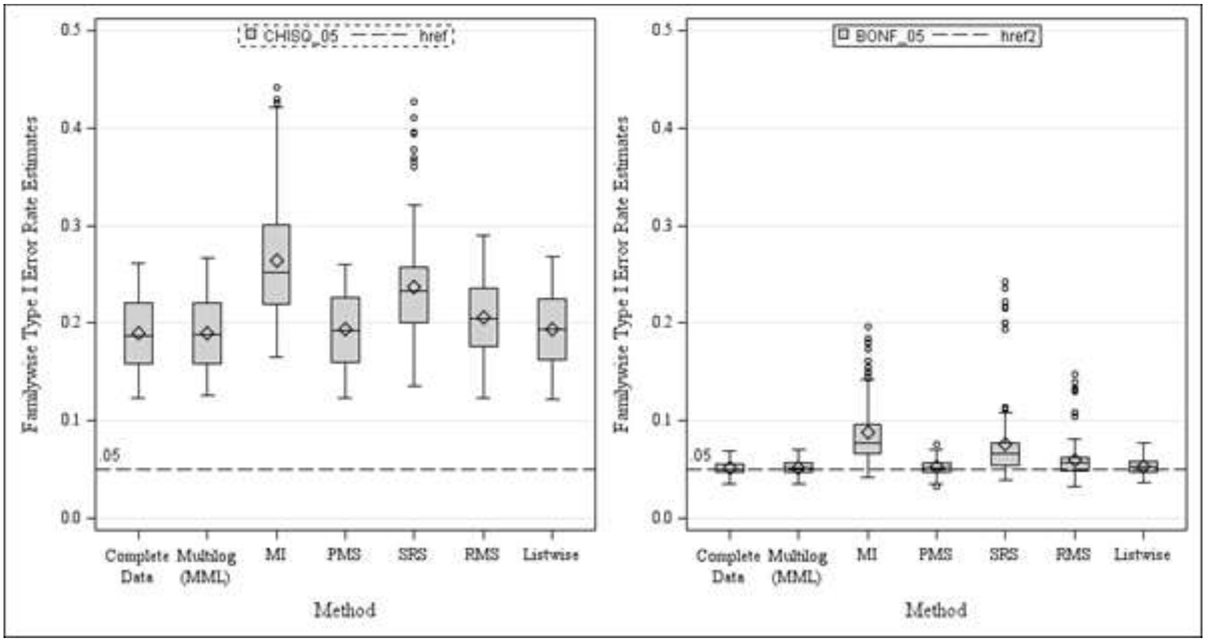


Figure 19. Overall mean error distribution rates (α = .05) by missing data method and Critical χ2 and Bonferroni adjustment. Horizontal reference lines are at the respective nominal α level.

As can be observed in Figure 18, the overall distributions of the FWER estimates across all missing data methods when the $\chi^2_{(4)}$=9.49 critical value was applied to testing for DIF each item within a subscale at α = 0.5 (*M* = .19 to *M* = .26) were entirely consistent with the overall distributions of the FWER obtained when the $\chi^2_{(4)}$=13.28 critical value was applied to testing for DIF each item within a subscale at α =.01. That is, the application of the χ^2 critical values for the test of DIF resulted in inflated mean FWER in both levels of alpha. Similarly consistent with the results at α = .01, the Multiple imputation (MI) and single regression substitution (SRS) methods showed the larger mean FWER ($M = .26$ and $M = .24$ respectively) when $\alpha = .05$, with MI also showing the largest upward dispersion (.442), followed by SRS (.427). In contrast, but consistent with the results for the Bonferroni adjustment at $\alpha = .01$, the use of the Bonferroni adjustment for the testing of DIF for each individual item (see Table 9) resulted in mean FWER for the complete data, FIML, PSM, RMS, and Listwise deletion methods at or slightly above $\alpha = .05$. Consistent too with the results for $\alpha = .01$, both MI and SRS had the higher mean FWER ($M = .09$ and $M = .08$ respectively), with the SRS method having the largest upward dispersion followed by MI (.24 and .20 respectively). While the RMS method had a slightly higher mean FWER than the nominal alpha ($M = .06$), it also showed greater upward dispersion (.15) than that for the methods with mean FWER at the nominal level. These extreme upward values or outliers for the MI, SRS, and RMS methods (for both χ^2 and Bonferroni adjustment) suggested the effect of a factor or factors in the study. However, because these initial inspections showed an inflated family error rate across all methods when hypothesis tests were conducted using the χ^2 critical value at $\alpha = .05$ and consequently, none of the conditions met the Bradley's criteria for robustness, further analyses were not conducted for these conditions. Thus, effect sizes (η^2) only for the Bonferroni adjustment were computed to determine the impact of first-order interactions and main effects, for each method and across all factors. Effect size estimates were computed for all MDM to identify the factors contributing to the variability of the FWER. Table D1 displays the η^2 for all main effect and first-order interactions. The following sections report only effect sizes by method for $\eta^2 \geq .05$.

Complete Data ($\alpha = .05$)

Effect size estimates were computed for all first-order interactions for the complete data method. This method had an overall mean FWER at the nominal level ($M = .05$) and a minimum and maximum value of 0.04 and 0.06 respectively. Under this MDM, and consistent with the results for $\alpha = .01$, none of the interactions or main effects had an effect size $\eta^2 \geq .05$.

FIML ($\alpha = .05$)

The overall distribution of FWER for the FIML displayed in Figure 18 showed that the FIML method had a mean FWER at the nominal level ($M = 05$) and a minimum and maximum values of 0.04 and 0.07 respectively. When $\alpha = .01$, the interaction of ability distribution and the proportion of missing items had an $\eta^2 = .06$ under FIML; however, the analysis of effect size (η^2) for first-order interactions and main effects at $\alpha = .05$ did not suggest the effect of a factor in the study contributing to the FWER.

Multiple Imputation ($\alpha = .05$)

Consistent with the results for MI at $\alpha = .01$, the overall distribution of FWER for the MI displayed in Figure 18 showed that when $\alpha = .05$, the MI method had an overall FWER above the nominal level ($M = 09$). The variability observed in the overall distribution (min = 0.042; max = 0.197) suggested the effect of a factor or factors in the study on the familywise error rates. Similarly to the results for MI when $\alpha = .01$, the analysis of the effect size for first-order interactions showed a large effect for the interaction between the proportion of missing observations and the proportion of missing items ($\eta^2 = .11$) associated with the total variability observed in the FWER. Figure 19 displays the mean FWER for the interaction effects of ability distribution and the proportion of missing observations.

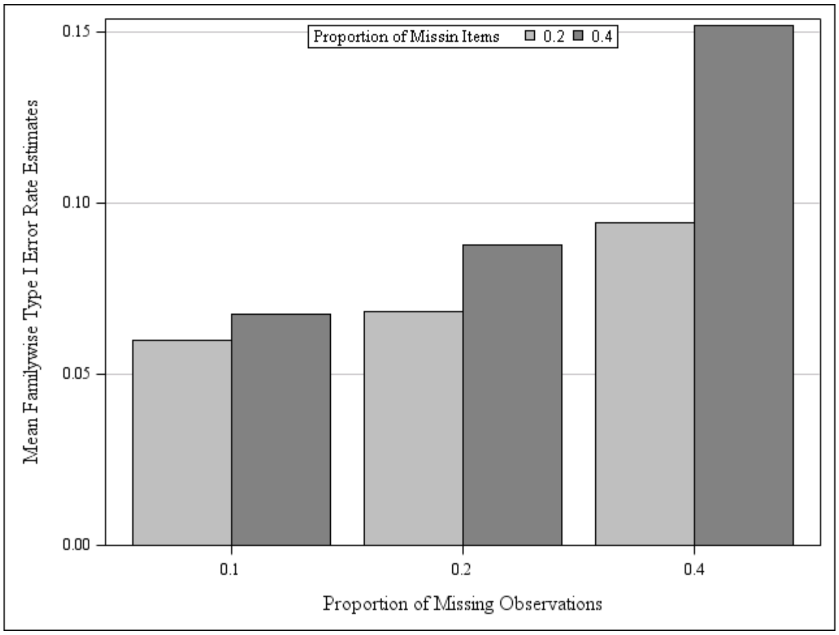


Figure 20. Mean familywise error rates for MI (α = .05) by proportion of missing observations and proportion of missing items (η^2 =.11).

As observed in in the interaction of the proportion of missing observations and the proportion of missing items when α = .01 (Figure 10), Figure 19 also showed that interaction of the proportion of missing observations and the proportion of missing items had a large effect on FWER. Mean FWER increased as the proportion of missing observations increased and as the number of missing items increased. When the proportion of missing observations was .10, there was a slight increase in mean error rates from the proportion of missing items = .20 to the proportion of missing items = .40 (*M* = 0.06 to *M* = 0.07 respectively). In contrast, markedly higher mean error rates were observed from the proportion of missing items = .20 to the proportion of missing items = .40 when the proportion of missing observations was .40 (*M* = 0.09 to *M* = 0.15 respectively).

Person Mean Substitution (α = .05)

Effect size estimates were computed for all first-order interactions for the person mean substitution method. This method had an overall mean FWER at the nominal level (*M* = .05). Although a slight variability was observed (min = 0.033; max = 0.076), under this MDM, none of the first-order interactions or main effects had an effect on FWER, which was consistent with the results for this method at α = .01.

Single Regression Substitution (α = .05)

The overall distributions of FWER for the SRS displayed in Figure 18 showed that this method had an overall FWER above the nominal level (*M* = .08) and the variability observed in the overall distribution showed very extreme upward values (max = 0.243), suggesting factor effects. An analysis of the effect size for first-order interactions showed similarity to the results for α = .01 (see Figure 11 to Figure 13), with large effects for the interaction of the proportion of missing observations and the proportion of missing items ($\eta^2 = .13$), the interaction of number of items and proportion of missing observations ($\eta^2 = .10$), and the interaction of the number of items and the proportion of missing items ($\eta^2 = .09$). Figure 20 to Figure 22 display the mean FWER for the large effects of these interactions.

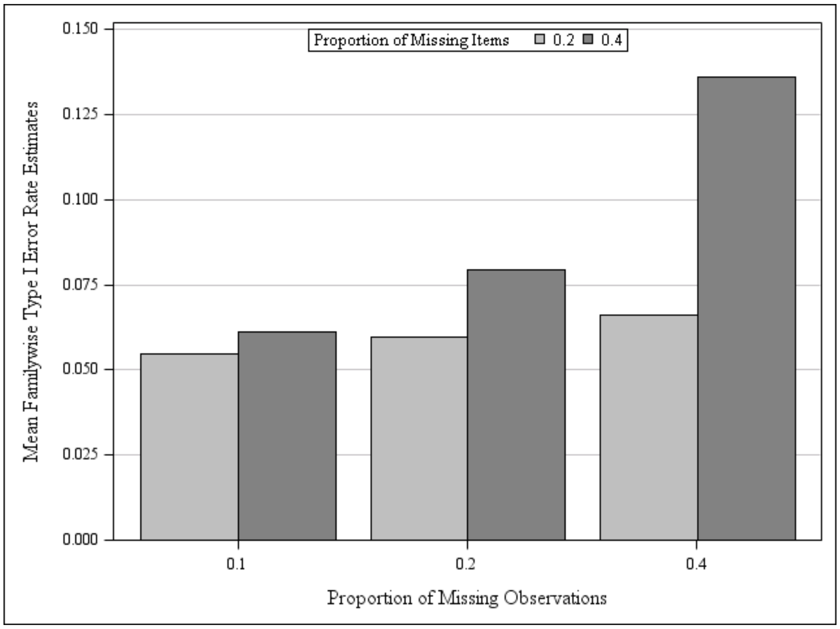


Figure 21. Mean familywise error rates for SRS (α = .05) by proportion of missing observations and proportion of missing items (η^2 =.13).

As observed in Figure 20, the proportion of missing observations and the proportion of missing items had a large effect on FWER for the SRS method. Mean FWER increased as the proportion of missing observations increased and as the number of missing items increased. When the proportion of missing observations was .10, the mean FWER was close to the nominal alpha for both proportions of missing items, increasing slightly from the proportion of missing items = .20 to the proportion of missing items = .40 (*M* = 0.054 to *M* = 0.061 respectively). In contrast, markedly higher mean FWER were observed from the proportion of missing items = .20 to the proportion of missing items = .40 when the proportion of missing observations was .40 (*M* = 0.066 to *M* = 0.135 respectively).

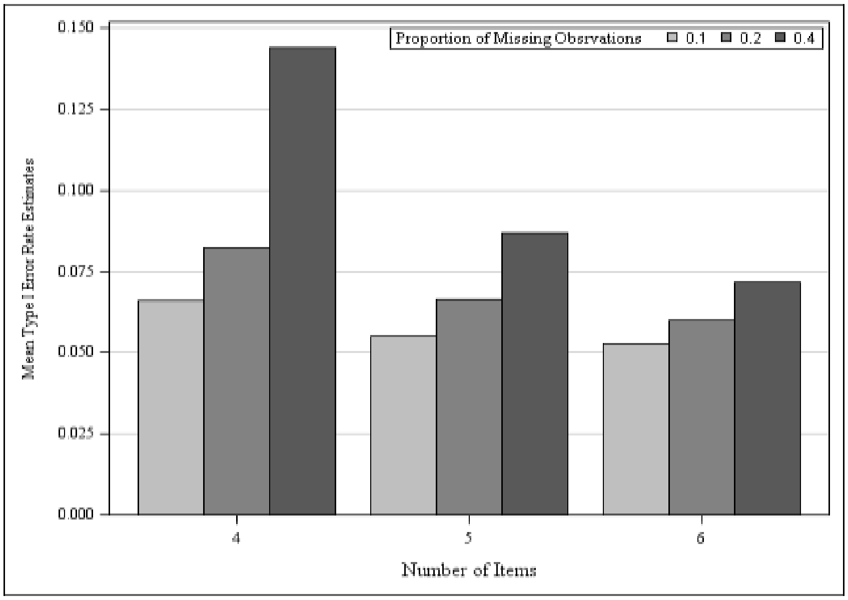


Figure 22. Mean familywise error rates for SRS (α = .05) by number of items and the proportion of missing observations (η^2 =.09).

As observed in Figure 21, number of items and the proportion of missing observations had a large effect on FWER for the SRS method. Mean FWER increased as the proportion of missing observations increased but smaller mean differences were observed as the number of items increased. Overall, across all number of items there was a notable increase in the mean FWER from missing observations = .20 to missing observations = .40 than there was from missing observations = .10 to missing observations = .20. But there was a marked increase in mean FWER from the proportion of missing observations = .20 to the proportion of missing observations = .40 (*M* = 0.021 to *M* = 0.054 respectively) when the number of items was four.

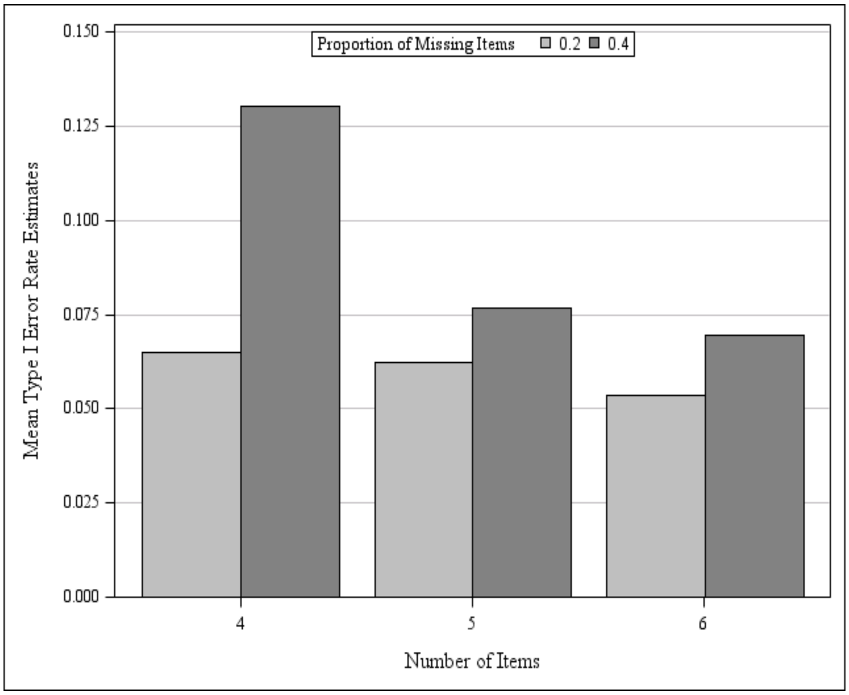


Figure 23. Mean familywise error rates for SRS (α = .05) by number of items and the proportion of missing items (η^2 =.10).

As observed in Figure 22, number of items and the proportion of missing items had a large effect on FWER for the SRS method. Mean FWER increased as the proportion of missing items increased but decreased as the number of items increased. Mean FWER were very similar when the number of items was five and six, with very small mean difference increases observed as the number of missing items increased. There is a notable increase in the mean FWER going from missing items .20 to missing items .40 when the number of items was four (*M* = 0.06 to *M* = 0.13 respectively).

Relative Mean Substitution (α = .05)

The overall distributions of FWER for the RMS displayed in Figure 18 showed that this method had a FWER slightly above the nominal level (*M* = 06) but the variability observed in the overall distribution showed very extreme upward values (max = 0.15), suggesting factor effects. An analysis of the effect size for first-order interactions for RMS showed similar to those for *α* = .01; that is, there were large effects for the interaction of the number of items and the proportion of missing items (η^2 = .16), as well for the interactions of number of items and proportion of missing observations, and the interaction of the proportion of missing observations and the proportion of missing items (η^2 = .11 and η^2 = .08 respectively). Figure 23 to Figure 25 display the mean FWER for these significant interactions.

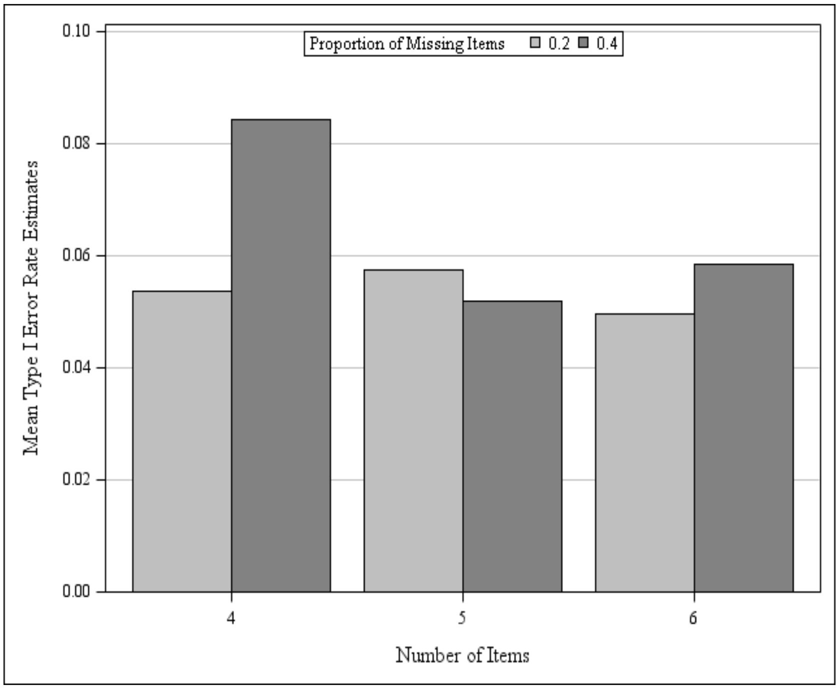


Figure 24. Mean familywise error rates for RMS (α = .05) by number of items and the proportion of missing items (η^2 =.16)

As observed in Figure 23, number of items and the proportion of missing items had a large effect on FWER for the RMS method. Mean FWER were consistent when the number of items was five and six. Markedly, there was a notable increase in the mean FWER from the proportion of missing items = .20 to the proportion of missing items = 40 (*M* = 0.05 to *M* = 0.08 respectively) when the number of items was four.

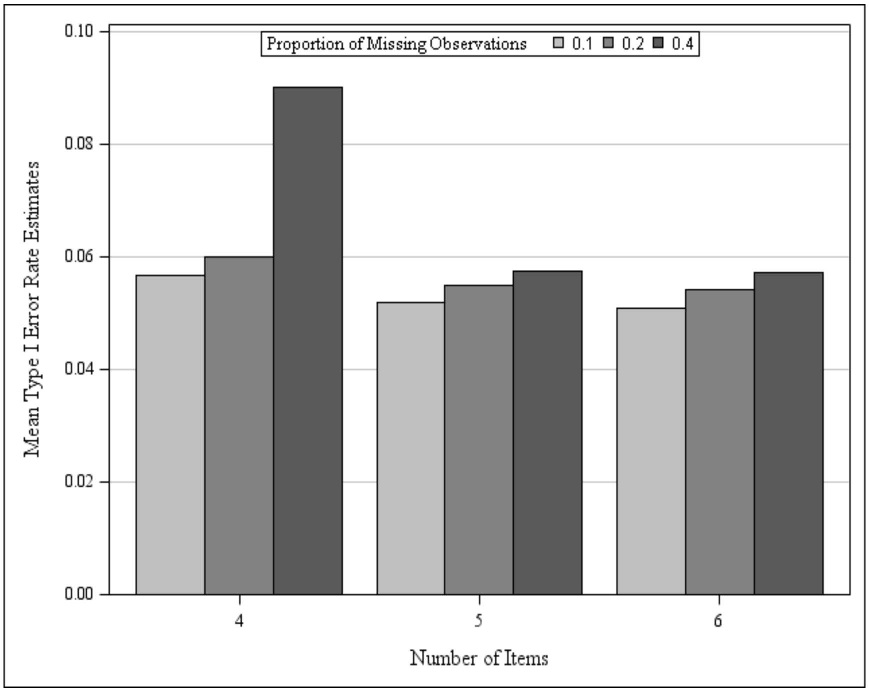


Figure 25. Mean familywise error rates for RMS (α = .05) by number of items and the proportion of missing observations (η^2 =.11).

As observed in Figure 24, the interaction of the number of items and the proportion of missing observations for α = .05 had a similar pattern like the interaction of number of items and the proportion of missing items for α = .01 (see Figure 16). Mean FWER for the interaction of the number of items and the proportion of missing observations were consistent when the number of items was five and six. Markedly, there was a notable increase in the mean FWER from the proportion of missing items = .20 to the proportion of missing items = 40 (*M* = 0.06 to *M* = 0.09 respectively) when the number of items was four.

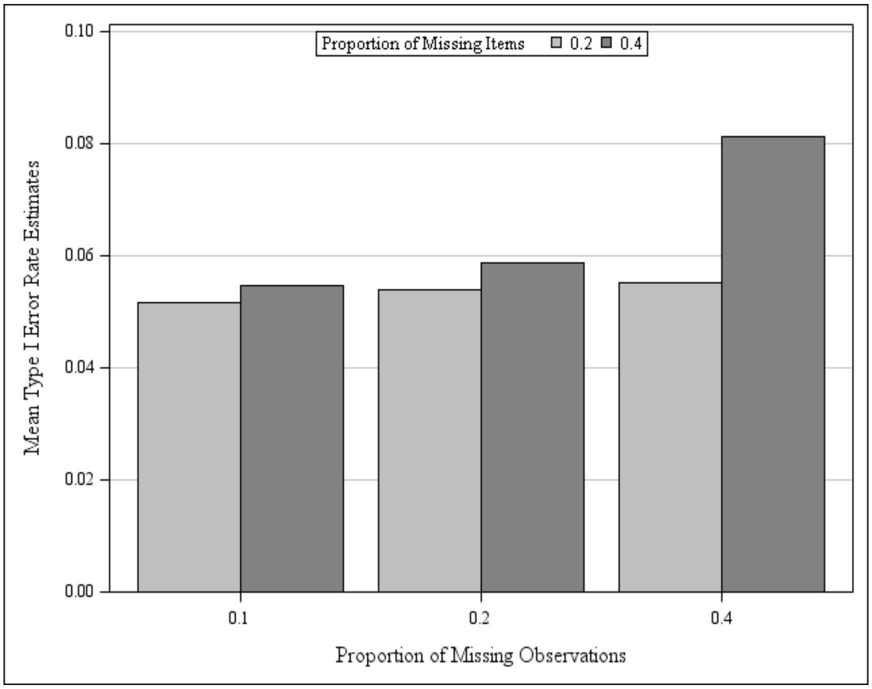


Figure 26. Mean familywise error rates for RMS (α = .05) by number of items and the proportion of missing observations (η^2 =.08)

As observed in Figure 25, the proportion of missing observations and the proportion of missing items had an effect on FWER for the RMS method. Similarly like in the same interaction for the SRS method for α = .01, in the RMS method the mean FWER increased as the proportion of missing observations increased and as the number of missing items increased when α = .05. Slight increases in mean FWER for the proportion of missing items were observed when the proportion of missing observations was .10 (*M* = .051 to *M* = .054) and when the proportion of missing observations was .20 (*M* = 0.054 to *M* = 0.058). In contrast, markedly higher mean FWER were observed from the proportion of missing items = .20 to the proportion of missing items = .40 when the proportion of missing observations was .40 (*M* = 0.05 to *M* = 0.08 respectively).

Listwise Deletion (α = .05)

Effect size estimates were computed for all first-order interactions and main effects for the Listwise deletion method. Listwise deletion had an overall mean FWER at the nominal level (*M* = .05). Within this MDM no extreme values were observed (min = 0.005; max = 0.021). Whereas this method showed only main effect for the number of items factor when α = .01 (see Figure 17), at α = .05, the interaction of sample size and the proportion of missing observations had an effect (η^2 = .07). Figure 26 displays the mean FWER for this interaction.

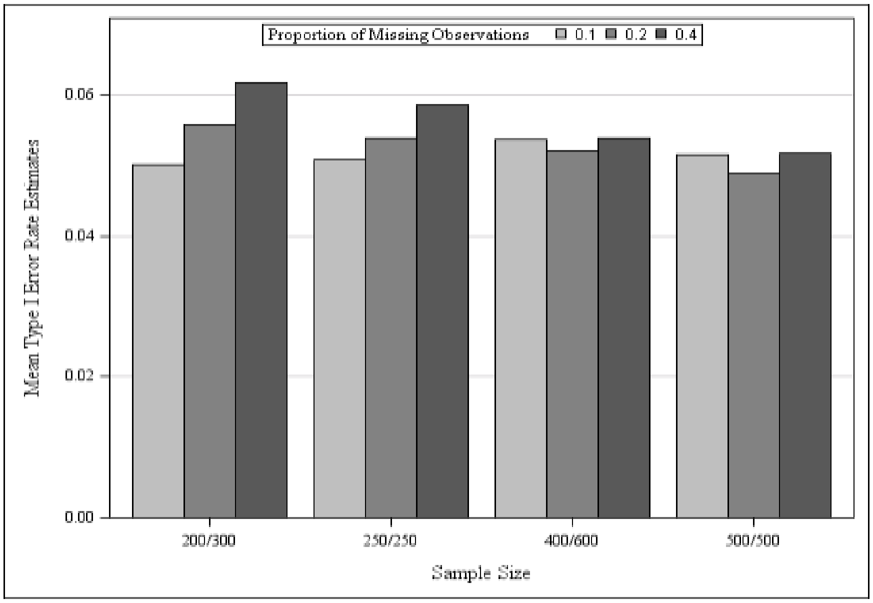


Figure 27. Mean familywise error rates for Listwise deletion (α = .05) by sample size and the proportion of missing observations (η^2 =.08)

As observed in Figure 26, the mean FWER were very consistent around the nominal alpha .05 for the largest sample size (*N* = 1000) both for the balanced (n_F = 500 and n_R = 500) and unbalanced samples (n_F = 400 and n_R=600) across all levels of the proportion of missing observations but slightly smaller means were observed in the balanced samples. A consistent increase in mean FWER was observed across all levels of the proportion of missing observations in the smaller sample size (N = 500) for both the balanced (n_F=250 and n_R = 250) and unbalanced samples (n_F = 200 and n_R = 300). The largest proportion of missing observations (.4) for these smaller samples had mean FWER slightly above of the nominal alpha (M = .061 and M = 0.58 respectively.

Type I Error Control (Bradley's Criteria)

Bradley (1978) provides the qualifying and quantifying criteria for what constitutes robustness when examining Type I error rates; that is, criteria under which it is valid to make inferences about the probability of making a Type I error. Factors such as the nominal alpha, and the direction and location of the critical rejection regions come into place when determining criteria for robustness (i.e., range of robust *p*s in terms of α). Under Bradley's criteria, this study investigated the effect of missing data and missing data methods on the robustness of the IRT-LR tests for an upper tail χ^2 and Bonferroni adjustment at α =.01 and α = .05. Because of the inflated Type I error rates obtained when χ^2 was used to tests the null hypothesis, Bradley's criteria for robustness was not met for both .01 and .05 significance levels. Only the proportions of conditions meeting Bradley's criteria for robustness for the Bonferroni adjustment are reported, both for .01 and .05 significance levels. Table 10 shows the proportion of conditions meeting Bradley's criteria at Bonferroni adjustment α=.01 and α = .05 respectively across methods and by research factors.

Table 11

Proportion of Conditions with Adequate Type I Error Control (Bradley's Criterion) by Research Design Factors and Bonferroni Adjustment (α = .01 *and* α = .05)

Factor	Complete Data		FIML		Multiple Imputation		Person Mean Substitution		Single Regression Substitution		Relative Mean Substitution		Listwise	
Sample Size	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
250/250	.81	1.0	.78	1.0	**.22**	**.44**	.86	1.0	**.47**	**.69**	.69	.94	.78	1.0
500/500	.86	1.0	.89	1.0	**.22**	**.39**	.78	1.0	**.53**	**.64**	.61	.94	.81	1.0
200/300	.92	1.0	.83	1.0	**.25**	**.50**	.78	.97	**.47**	**.69**	.75	.92	.61	.97
400/600	.83	1.0	.86	1.0	**.25**	**.47**	.86	1.0	**.47**	**.75**	.75	.94	.89	1.0
Scale														
J (4 items)	.85	1.0	.81	1.0	**.15**	**.33**	.83	1.0	**.35**	**.50**	.58	.83	.73	1.0
H (5 Items)	.85	1.0	.85	1.0	**.29**	**.52**	.85	1.0	**.52**	**.75**	.81	1.0	.88	.98
G (6 items)	.85	1.0	.85	1.0	**.27**	**.50**	.77	.98	**.58**	**.83**	.71	.98	.71	1.0
Missing Items														
.20	.83	1.0	.83	1.0	**.29**	**.58**	.82	1.0	**.64**	**.90**	.79	1.0	.72	.99
.40	.88	1.0	.85	1.0	**.18**	**.32**	.82	.99	**.33**	**.49**	.61	.88	.82	1.0
Missing Observations														
.10	.83	1.0	.83	1.0	**.48**	**.90**	.83	1.0	**.67**	**.92**	.83	1.0	.85	1.0
.20	.83	1.0	.83	1.0	**.23**	**.44**	.88	1.0	**.56**	**.77**	.73	1.0	.81	1.0
.40	.90	1.0	.85	1.0	**.00**	**.02**	.73	.98	**.23**	**.40**	.54	.81	.65	.98
Ability Distribution														
0,1:0,1	.83	1.0	.82	1.0	**.24**	**.44**	.83	.99	**.53**	**.69**	.71	.93	.79	.99
0,1:-.5, 1	.88	1.0	.86	1.0	**.24**	**.46**	.81	1.0	**.44**	**.69**	.69	.94	.75	1.0

Note. Estimates were based on 1,000 samples of each condition. Bolded figures represented the lower estimates by MI and SRS methods

Using a Bonferroni adjustment for nominal α = .01, the Bradley's criterion indicates that Type I error control is considered adequate if the estimated Type I error rate falls within $\alpha_{nominal}$ ± $0.5_{nominal}$ (i.e., 005 < rejection rate < .015). When using Bonferroni adjustment at a nominal α = .05, Bradley's liberal criterion indicates that Type I error control is adequate if the estimated

Type I error rate falls within $\alpha_{nominal} \pm 0.5_{nominal}$ (i.e., 025 < rejction rate < .075). As observed in Table 10, the multiple imputation (MI) method had the smaller proportions of conditions having adequate Type I error control across all factors for both α = .01 and α = .05 compared to all other methods, followed by the single regression substitution (SRS) method. Within MI, the proportions of conditions meeting Bradley's criteria for robustness were very consistent across all factors and levels of significance. That is, within MI, smaller proportions meeting Bradley's criteria for robustness were observed for the balanced samples and proportions of Bradley's rates were markedly lower when the number of items was four, compared to the Bradley's rates for scales with five and six items. Within MI and the number of missing observations, none of the conditions met Bradley's criteria when the proportion of missing observations was .40 at α = .01 and only .02 met Bradley's criteria for robustness when α = .05. Complete data and FIML, on the other hand, had the larger proportions of conditions with adequate Type I error control over all simulation conditions and levels of significance, followed by the person mean substitution (PMS) method. Figure 27 and Figure 28 show the proportions having adequate Type I error control by method.

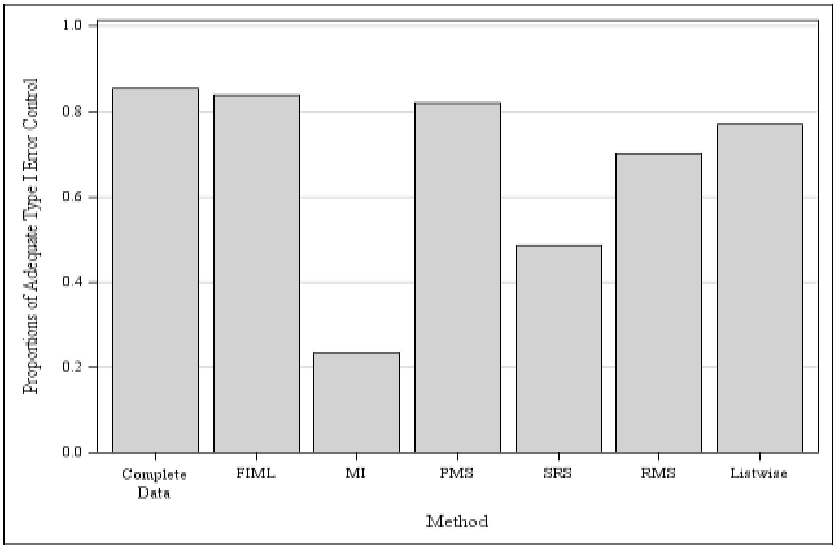


Figure 28. Proportion of conditions with adequate Type I error control by method (α = .01).

As shown in Figure 27, complete data, FIML, and PMS had the larger proportion of conditions having adequate Type I error control at α = .01 by Bradley's (1978) criteria (.85, .84, and .82 respectively). That is, from a total of 144 conditions, complete data, FIML, and PMS had 123, 121, and 118 conditions having adequate Type I error control respectively. When α = .01, Listwise deletion had 111 conditions meeting Bradley's criteria for robustness or .77. Among all methods, MI had the lowest proportion of conditions meeting Bradleys' criteria (.24 or 34 conditions out of 144) followed by SRS (.49 or 70 conditions out of 144). Figure 28 shows the proportion of conditions meeting Bradley's criteria for robustness at α = .05.

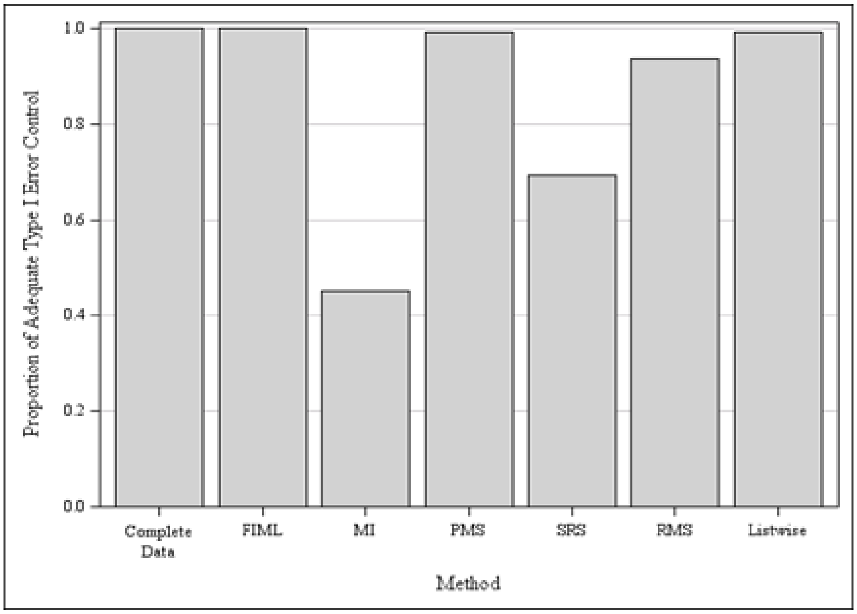


Figure 29. Proportion of conditions with adequate Type I error control by method (α = .05).

As shown in Figure 28, when the tests of the null hypothesis were conducted using a Bonferroni adjustment for α = .05, complete data and FIML had all 144 conditions meeting Bradley's criteria for robustness. Both PSM and Listwise had .99 of conditions (i.e., 143 conditions) having adequate Type I error control. From all MDM, MI had the smaller proportion of conditions (.45) or only 65 conditions having adequate Type I error control, followed by the SRS (.69) with 100 out of 144 having adequate Type I error control.

Research Question 1 Summary (α = .05)

An inflated FWER resulted across all methods when the significance testing for DIF was conducted using χ^2 critical values at $\alpha = .05$. In addition to inflated error rates, MI, SRS, and RMS showed also greater dispersion toward higher error rates. On the other hand, when the critical values for significance testing were corrected, using a Bonferroni adjustment, the FWER across all methods resulted in mean distributions around $\alpha = .05$. However, patterns of upward dispersion of error were also observed for MI, SRS, and RMS.

An analysis of the effect of each simulation factor on FWER by MDM showed the following results when $\alpha = .05$:

1. None of the simulation factors had any effect on FWER under complete data and FIML method
2. The interaction of ability distribution and number of items had a moderate effect on FWER only under PMS ($\eta^2 = .05$).
3. The interaction of number of items and proportion of missing observations and the interaction of number of items and proportion of missing observations had a large effect on FWER under SRS and RMS ($\eta^2 = .09$ and $\eta^2 = .11$ respectively).
4. The interaction of number of items and the proportion of missing items had also a large effect on FWER under SRS and RMS ($\eta^2 = .10$ and $\eta^2 = .16$ respectively).

5. The interaction of the proportion of missing observations and the proportion of missing items had an effect (η^2 = 08) on FWER under RMS, and a large effect large effect (η^2 = 11 and η^2 = 13) on FWER under MI and SRS respectively.
6. The implementation of Bradley's (1978) criteria for robustness showed similar patterns across factors and levels of significance.
7. Complete data, FIML, and PMS had the largest proportions of conditions meeting Bradley's criteria for robustness, followed by Listwise deletion.
8. Notably, MI was the MDM having the lowest proportions of conditions meeting Bradley's criteria for robustness, followed by the SRS method for both $\alpha = .01$ and $\alpha = .05$.
9. Across factors, none of the Type I error rates under MI and SRS met the Bradley's criteria for robustness when $\alpha = .01$; thus, these methods were not included in power comparison across methods.

Research Question 2. Effect of Missing Data on Statistical Power

Within the context of multiple comparisons, as explained previously, statistical power of the free-baseline IRT-LR test was estimated as per test power (Toothaker, 1991). Meaningful power analyses (Ankenmann et al., 1999) were conducted after adequate Type I error control was established. Thus, to address research question 2 when $\alpha = .01$ power comparisons across methods were made only for those conditions in which all methods provided adequate Type I error control.

Power - Bonferroni Adjustment (α = .01)

Power was estimated as a per test power for all those items that were manipulated to simulate DIF within each subscale. DIF was simulated so that the last item of each scale was slightly, moderately, and highly more difficult for the focal group compared to the reference group by shifting the items' location parameters by .25, .50, and .75 respectively. Then, as explained in the previous section, power comparisons were made only for conditions in which all methods provided adequate Type I error control. When α = .01, none of the conditions for MI and SRS met this criterion and these MDM were not considered for power analyses.

Differences in mean power estimates by method were relatively small, ranging from .62 for Listwise deletion method to .68 for complete data. Figure 29 shows the overall distribution of Power estimates by method across all simulation factors.

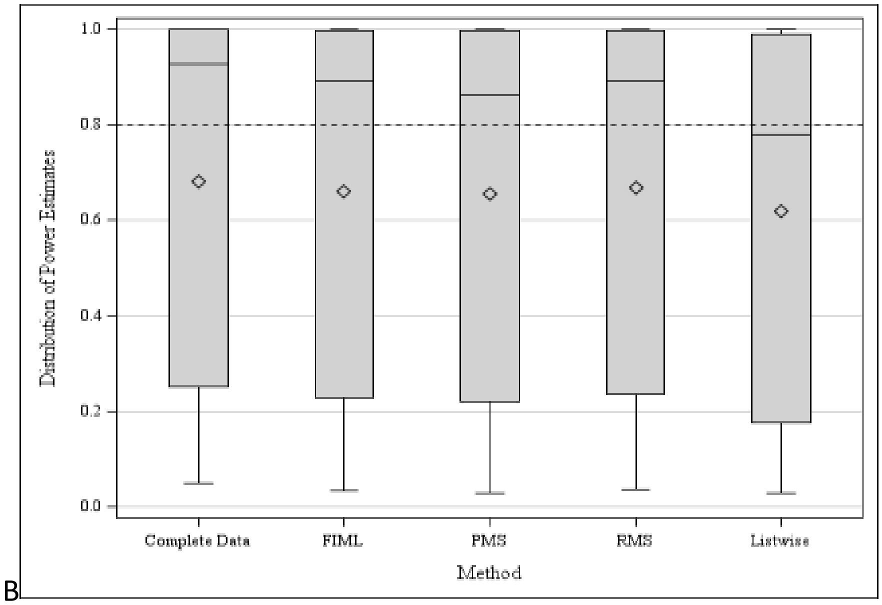


*Figure 30***.** Overall distribution of power estimates by method (*α* = .01) across all simulation factors. Dashed reference line for power is set at .80. Only methods having the same conditions with adequate Type I error control were considered for power analysis.

As observed in Figure 29, the overall distribution of the power estimates by method across all simulation factors were similar in how they spread (i.e., negatively skewed); however, some differences were observed. Within each distribution, the median of all MDM except Listwise deletion (.78) were above the .80 dashed power reference line. Notably observed was the larger length of the lower half of the distribution compared to the length of the upper half for all methods; that is, the lower half of the data for all distributions was more dispersed than the data in the upper half, pulling the mean from the median. Mean power estimates for all methods were above .60, with complete data and RMS having the largest means (.66 and .65 respectively). The smallest mean power estimate was for Listwise (.61). Additionally, the upper half of all methods overlapped but for Listwise there was more dispersion. The next step was to investigate the impact of factors to determine whether their interactions (i.e., first-order interactions) or main effects explained the differences in the distributions of mean power estimates observed across methods. However, none of the first-order interactions or main effect by method resulted in significant effect size η^2. Thus, the distributions of power estimates were reviewed by method and magnitude of DIF. Table 11 shows the overall mean power by method across DIF.

Table 12

Mean Power Estimates by Method and DIF for Bonferroni Adjustment .01

	DIF Effect Size		
Method	.25	.50	.75
Complete data	.20	.85	.99
FIML	.17	.82	.99
Person Mean Substitution	.16	.81	.99
Relative Mean Substitution	.18	.83	.99
Listwise Deletion	.13	.75	.97

Note. Estimates were based on 1,000 samples of each condition.

As can be observed in Table 11, statistical power to detect the smallest DIF effect size (.25) was very low across all MDM. On the other hand, when DIF was .75, all methods detected this largest effect size at an extremely high rate. The effect size η^2 for first-order interactions and main effects was estimated to investigate the impact of the simulation factors on the variability of power estimates only for detecting DIF = .50. Results indicated that the interaction of number of items and sample size had an effect on power estimates for the complete data, FIML, PSM and RMS methods. Figure 30 to Figure 33 show this interaction for the mentioned MDM.

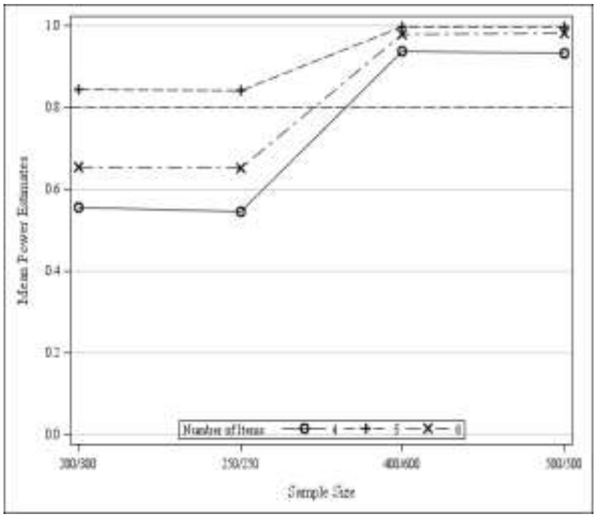


Figure 31. Mean power estimates for complete data (α = .01) by sample size and number of items (η^2 =.08)

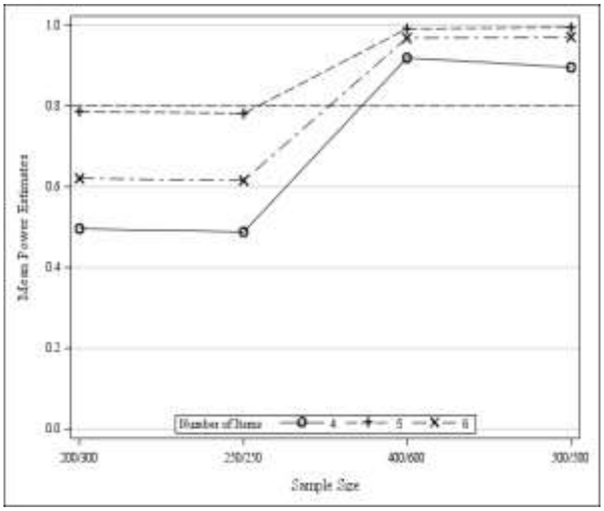


Figure 32. Mean power estimates for FIML (α = .01) by sample size and number of items η^2 =.05)

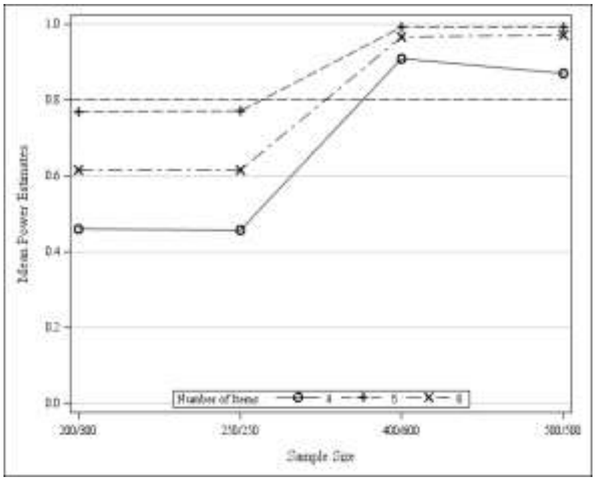


Figure 33. Mean power estimates for PMS (α = .01) by sample size and number of items (η^2 =.05)

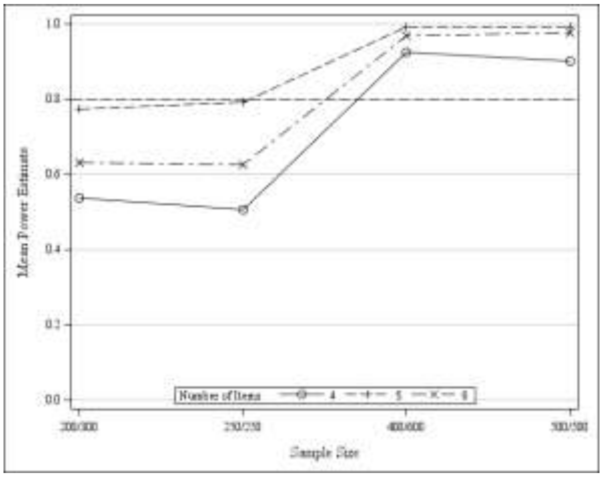


Figure 34. Mean power estimates for RMS (α = .01) by sample size and number of items (η^2 =.05)

As shown in Figure 30 to Figure 33, an effect on power estimates was observed for the interaction of sample size and number of items across complete data, FIML, PSM, and RMS methods. Power increased as total sample size increased. Also, it was observed under these MDM, power estimates varied slightly by number of items within a total sample size level (e.g., complete data power estimates for the small total sample size (*N* = 500) for both unbalanced (200/300) and balanced (250/250) sample sizes were very similar when the number of items was four (*M* = .56 and *M* = .55 respectively). While larger power estimates were observed as the number of items increased, power estimates were slightly higher when the number of items was five than when the number of items was six. Within the small total sample size, some differences were noted. FIML, PMS, RMS, and Listwise mean power estimates were slightly lower than those power estimates for complete data (i.e., all mean power estimates for these methods were below the .80 power reference line for the total small sample size). As with complete data (Figure 30), and FIML (Figure 31), sample size and the number of items had similar effect on the mean power estimates for the PMS method (η^2 = .05). Within the RMS method, both sample size and the number of items had a moderate significant effect (η^2 = .05) on mean power estimates. The same trends are observed in the RMS, such as consistent mean power estimates within a total sample size (small total sample and large total sample) within the same number of items. A sharp increase in power estimates were observed for the total large sample size. Thus, across all methods, statistical power was consistently higher for the larger sample size, both for balanced and unbalanced samples.

Mean power estimates were examined for the Listwise deletion method. Results showed that none of the first-order interactions had an effect on power estimates. However, main effects were observed for number of items (η^2 = .15), sample size (η^2 = .67), and the proportion of missing observations (η^2 = .12). Figure 34 to Figure 36 display these main effects.

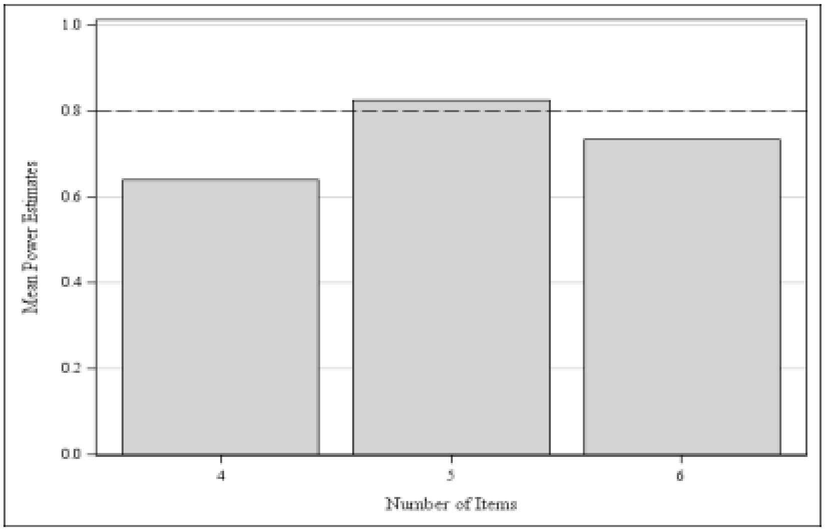


Figure 35. Mean power estimates by number of items for Listwise deletion method (η^2 =.15)

The number of items had a large effect (η^2 =.15) on the mean power estimates for the Listwise deletion method as observed in Figure 34. Results showed that the smaller mean power (.63) was estimated when the number of items was four while the largest estimate of mean power was obtained when the number of items was 5 (.83).

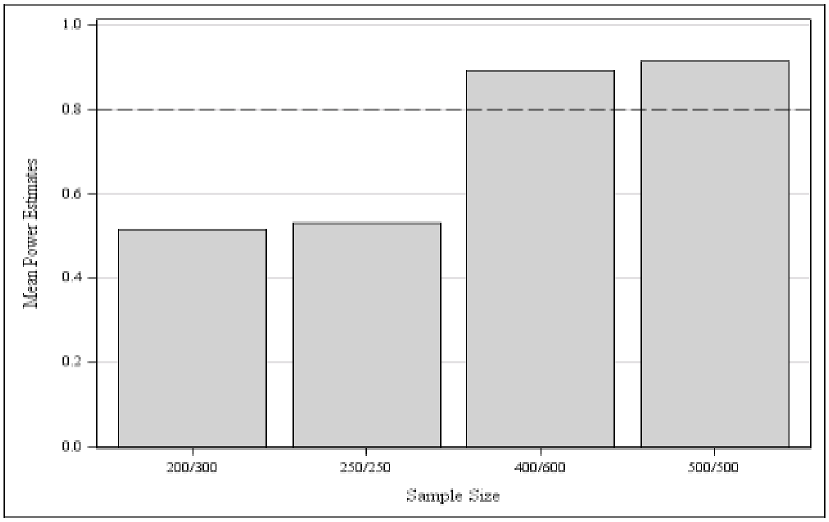


*Figure 36***.** Mean power estimates by sample size for Listwise deletion method (η^2 =.67)

Sample size had a large effect on mean power estimates (η^2 =.67) under Listwise deletion method. As observed in Figure 35, estimates were very consistent within total sample size, with lower power estimates for the smaller total sample size. A sharp increase in mean power estimates was observed for the larger total sample size (from .52 and .53 for the unbalanced and balanced small total sample, to .89 and 91 for both levels of the total large sample size).

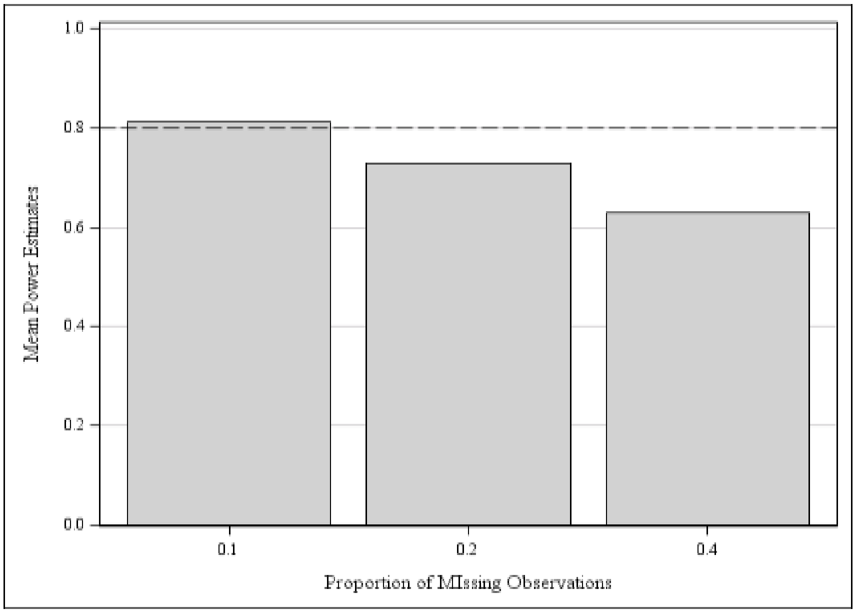


Figure 37. Mean power estimates by the proportion of missing observations for Listwise deletion method (η^2 =.12).

The proportion of missing observations had also a large effect size on the power mean estimates (Figure 36) under Listwise deletion method. Power estimates decreased as the proportion of missing observations increased. Overall, Listwise deletion had the smaller mean power estimates than any of the other methods.

Power - Bonferroni Adjustment (α = .05)

As explained in the previous section, statistical power was estimated as a per test power and power comparisons were made only for conditions in which all methods provided adequate

Type I error control when α = .05; MI and SRS were poor Type I error rate methods (see Table 10). and were dropped from further analyses of power.

Differences in mean power estimates by method were relatively small, ranging from .56 for Listwise deletion method to .65 for complete data. Figure 37 shows the overall distribution of power estimates by method across all simulation factors.

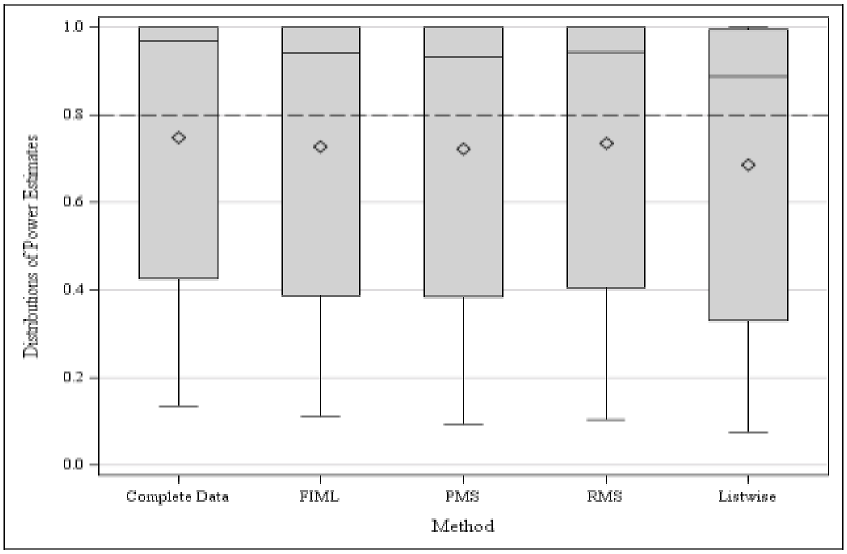


Figure 38. Overall distribution of power estimates by method (α = .05) across all simulation factors. Dashed reference line for power was set at .80

As observed in Figure 37, the overall distribution of the power estimates by method across all simulation factors were similar to those observed when α = .01, in how the distributions of power estimates spread (i.e., negatively skewed). That is, when α = .05, all methods had similar, consistent distributions of power. However compared to α = .01, when alpha was set to .05 all methods had medians above .80, including Listwise deletion, and slightly higher means, ranging from .69 (for Listwise deletion) to .75 (for complete data). Similarly, it was observed a larger length of the lower half of the distributions compared to the length of the upper half; that is, the lower half of the data for all distributions across methods was more dispersed than the data in the upper half, pulling the mean from the median. Additionally, the upper half of all methods overlapped. The mean power estimates were reviewed by method and magnitude of DIF. Table 12 shows the overall mean power by method across DIF.

Table 13

Mean Power Estimates by Method and DIF for Bonferroni Adjustment .05

	DIF Effect Size		
Method	.25	.50	.75
Complete data	.33	.91	.99
FIML	.30	.89	1.00
Person Mean Substitution	.29	.88	.99
Relative Mean Substitution	.31	.90	1.00
Listwise Deletion	.24	.83	.99

As can be observed in Table 12, estimates of power to detect the smallest DIF effect size (.24 to .34) were very low. On the other hand, when DIF was .75, all methods detected this largest effect size perfectly. The effect size η^2 for first-order interactions and main effects was estimated to investigate the impact of the simulation factors on the variability of power estimates for detecting DIF = .50. Effect size analyses for all MDM indicated that the interaction of items and sample size was significant ($\eta^2 = .05$ to η^2.14). Figure 38 to Figure 44 displays the mean power estimates for the interaction of items and sample size for $\alpha = .05$ across methods.

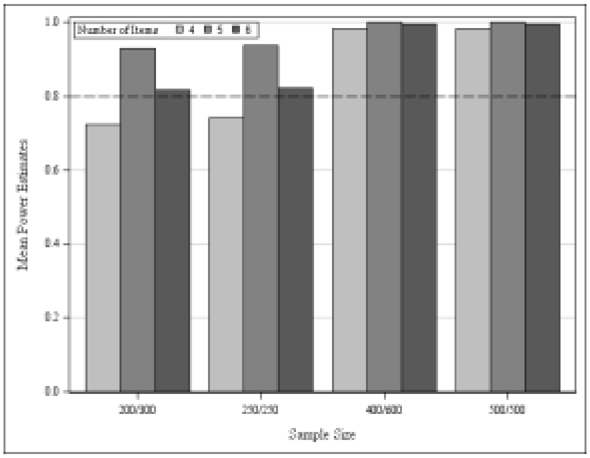


*Figure 39***.** Mean power estimates for complete data by sample size and number of items (η^2 =.14)

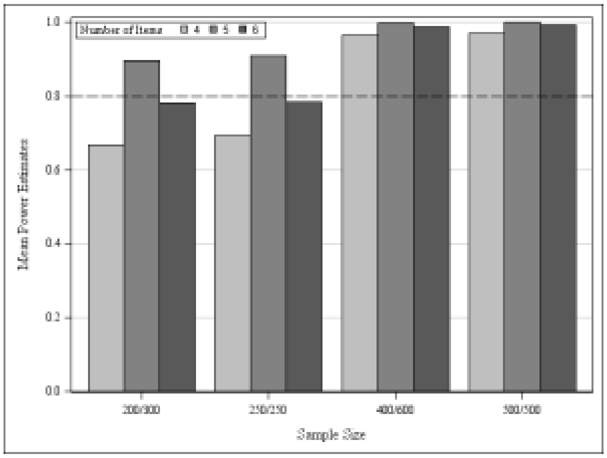


Figure 40. Mean power estimates for FIML by sample size and number of items (η^2 =.11)

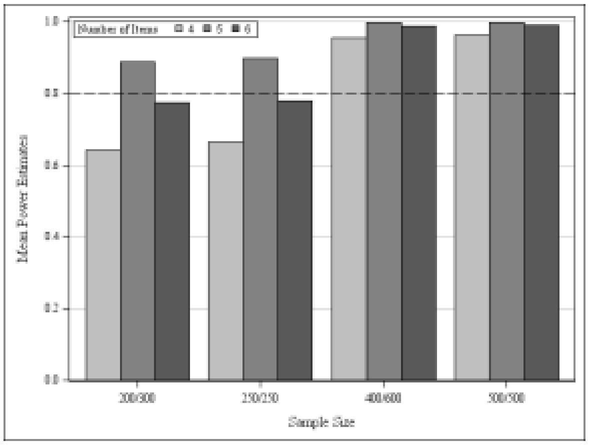


Figure 41. Mean power estimates for PSM by sample size and number of items (η^2 =.11)

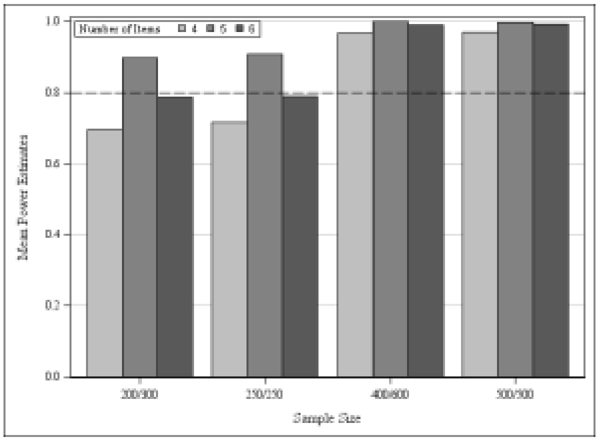


Figure 42. Mean power estimates for RMS by sample size and number of items (η^2 =.10)

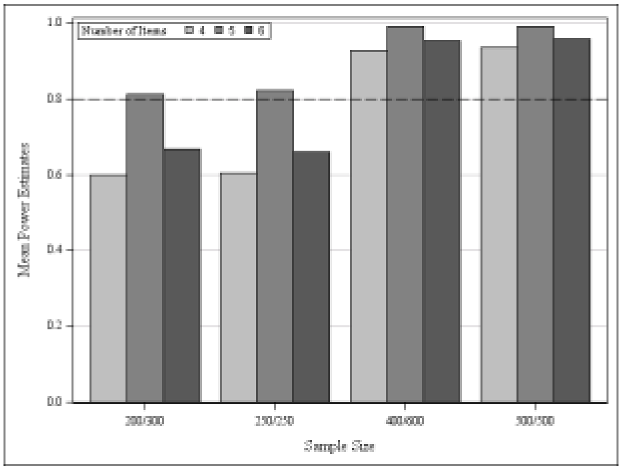


Figure 43. Mean power estimates for Listwise by sample size and number of items (η^2 =.05)

As shown in Figure 39 to Figure 43, an effect between sample size and number of items on power estimates was observed across all MDM (range between η^2 = .05 to η^2 = .14). That is, power increased as total sample size increased and increased as the number of items increased. Also, it was observed that power estimates were constant within a sample size range but larger when the number of items was five and six. A sharp increase in mean power was observed from the total small sample size to the total large sample size. Although the pattern of the relationship between sample size and number of items was very consistent, some differences are noted. Across all MDM mean power estimates for the small total sample size (*N*=500), mean power estimates for the unbalanced samples were slightly smaller. Across all MDM, the mean power estimates were below .80 for the small total sample size when the number of items was four; under Listwise deletion, mean power estimates were also below .80 when the number of items was six. Overall, Listwise mean power estimates were slightly lower than those estimated for all the other MDM. Like with complete data (Figure 39), and FIML (Figure 40), both sample size and the number of items had a significant effect on the mean power estimates for the PMS method (η^2 = .11). Within the RMS method, both sample size and the number of items had a moderate significant effect (η^2 = .10) on mean power estimates. The same trends are observed in the RMS, such as stable mean power estimates within sample size range (small total sample and large total sample), and a sharp increase in power estimates from the small total sample size to the large sample size. Thus, across all methods, statistical power was consistently higher for the larger sample size, both for balanced and unbalanced samples.

Research Question 2 Summary

Power estimates were very consistent across levels of significance and by MDM.

1. The power of the IRT-LR method for DIF detection when DIF = .25 was low regardless of missing data treatment.

2. The power of the IRT-LR method for DIF detection when DIF = .75 was entirely detectable. When = .01, detection was not perfect, but nearly so.

3. MI and SRS were poor Type I error rate methods and were removed from further analyses of power.

3. Factors that had an effect in the power of the IRT-LR test were further examined when DIF = .50.

4. Across significance levels a nd across all MDM, the power of the IRT-LR was entirely consistent.

5. The interaction of sample size and number of items had an effect on power estimates across all methods.

CHAPTER FIVE

DISCUSSION

“Trust, but verify” – Ronald Reagan’s signature phrase

Researchers have addressed missing data as a common problem in empirical research; as Widaman (2006) stated, “the presence of missing data is the rule, not the exception” (p. 42). Thus, whenever a research study is conducted, one fact is known: data will probably be missing. In educational research, specifically in test validation procedures such as differential item functioning or DIF, missing data are not the exception either. Methodologically speaking, missing data are considered a problem because most statistical procedures require complete data—thus the need for addressing missing data in applied research. Research on missing data and treatments in the context of differential item functioning (DIF) has been limited and has focused primarily on achievement assessment using binary items using large scales. However, as it has been stated in this study, noncognitive assessment and the use of polytomous data are relevant for measuring students’ progress.

The overall purpose of this study was to compare the effect of missing data, in the form of item nonresponse, on the Type I error and statistical power of the IRT-LR test for detecting DIF in polytomous items. The graded response model (GRM; Samejima, 1969, 2010), within the IRT framework, provided an adequate statistical model for the type of data in this study. In addition to complete data analyses, six methods for treating missing data in Likert-type items in scales measuring students’ attitudes were examined across factors such as levels of significance, examinees' ability distribution or impact, sample size, number of items, proportion of missing observations, proportion of missing items, and DIF magnitude. The two outcomes of the study were addressed with the following questions, 1) to what extent is the effect of missing data on the Type I error control of the IRT-LR tests for DIF detection in polytomously scored items consistent across study's factors, under complete data and missing data methods (MDM)?, and 2) to what extent is the effect of missing data on the statistical power of the IRT-LR test for DIF detection in polytomously scored items consistent across study factors, under complete data analyses and MDM?.

Before turning to the discussion of findings, it is important to point out some limitations noted in the previous research reviewed. In DIF analyses, the purpose of the studies is the identification of DIF items by conducting significance testing and estimating rejection error rates. However, inferences on the estimated rejection rates should not be made on the estimated rates themselves only (e.g., considering a given error rate as adequate if the proportion itself does not exceed, for example, .05). Thus, criteria for determining whether a test is robust (i.e., its properties and behavior are not affected when assumptions are not met), should be established when reporting Type I error results. In addition, in DIF studies, it is very common to have a research situation in which multiple testing is present and consequently, significance levels should be corrected to control Type I error rates. Furthermore, when conducting a study, it is important that researchers design and conduct their studies so that their inferences about significance hold. This is also true in simulation studies. When making inferences of significance out of the probability of being wrong, we tend to think of the likeliness of our findings, as Weisberg (2014) said, as being "the" probability rather than being an estimate of the probability (p. 349). Thus the recommendation of conducting a simulation study using field data (e.g., datasets or parameters from datasets) becomes relevant. A brief description of such limitations follows.

Criteria for Robustness

The criteria for addressing the adequacy of Type I error rates (i.e., when the test is robust to the violation of assumptions) must address the quantifying and qualifying conditions under which Type I error rates are controlled. The robustness of the test, in turn, can aid in making valid inferences and comparisons across experimental conditions. That is, once the robustness criteria is established, researchers can summarize the conditions for which the test is robust (see Table 10) and be more confident, for example, about making statements on the level of missingness at which missing data becomes problematic or under which MDM, Type I error is controlled. In addition to reporting the results for Type I error in terms of Bradley's criteria for robustness (another helpful criteria is the false discovery rate), this study also presented the results for Type I error considering the multiple testing for significance context.

The Multiple Significance Testing Problem

As was also discussed in the methods section, the implementation of the free-baseline approach to the IRT-LR method for DIF detection in polytomous items implies a series of comparison of nested models. Thus, the nature of this multiple testing required the adjustment of the levels of significance. While some researchers have dissented with the application of appropriate error rates in the presence of multiple testing (e.g., Saville, 1990), Mantel and Haenszel (1959) explained,

> For the usual problem of multiple significance testing, this would be equivalent to allocating a large part of the desired risk of erroneous acceptance of an association as real to a small group of comparison where fruitful results were anticipated, and parceling out the reminder of the available risk to the large bulk of comparisons. (p. 724)

Furthermore, Ryan (1959), reinforcing the idea of applying an appropriate level of significance when dealing with multiple testing, stated that, “there are important questions of logic involved in the use of these methods” (p. 26). That is, in the presence of multiple testing, valid inferences require the assignment of an appropriate level of significance to the specific research at hand, considering all the conditions involved. Otherwise, there is a risk of inflated Type I error rates. Take for example, Kromrey and Dickinson’s (1995) statement,

> If 100 experiments are conducted and each experiment includes 10 hypothesis tests, then 1,000 hypothesis tests have been conducted in total. If 50 of these hypothesis tests have resulted in Type I errors, and if these 50 Type I errors occurred in 40 of the experiments (i.e., some experiments have more than 1 Type I error), then the per hypothesis error rate is .05 (50 Type I errors / 1,000 hypothesis tests), the per experiment rate is .50(50 Type I errors / 100 hypothesis tests), and the experimentwise or familywise error rate is .40(40 experiments with at least 1 Type I error / 100 experiments). Of the three error rates, the family wise error rate has generally been accepted as the Type I error. (p. 54)

Thus, this study estimated Type I error rates as familywise error rates to further avoid an inflated rejection rate.

Statistical Power in the Context of Multiple Significance Testing

Because it is not meaningful to analyze power for conditions that do not have adequate Type I error control, only conditions with adequate Type I error control by Bradley's (1978) criteria for robustness were considered for the analysis of power. In addition, the estimation of statistical power in the context of multiple significance testing requires the selection of a type of power accordingly, depending on the tests conducted. Toothaker (1991) described three types of power in the context of multiple significance testing: 1) any test power (i.e., probability of rejecting at least one false hypothesis in a set of comparisons), 2) all tests power (i.e., probability of rejecting all false null hypothesis in a set of comparisons, and 3) per test power (i.e., probability of rejecting each false null hypothesis in a set of comparisons). This simulation study manipulated only one item in each scale so that it exhibited DIF. Thus, for this study, power was estimated, as explained in Chapter 3, as per test power.

Precision of Simulation Results

One last point to consider before the discussion of results is that of the precision of the findings in the previous research summarized in Chapter 2. The number of samples generated in some of the revised studies were insufficient for providing stable estimates (see, for example, Cohen, Kane, & Kim (2001) and Robey & Barcikowski (1992) for formulae for estimating the number of replications needed for a desired outcome).

In sum, the limitations discussed above (e.g., lack of a robustness criteria, no adjustment to the significance levels in the presence of multiple testing, and the insufficient number of replications for precise results) made it difficult to compare the results of some studies summarized in Chapter 2 with the results of the present study. In addition to the limitations mentioned, some of the summarized studies addressed binary items in the context of achievement testing using a large number of items whereas this study addressed the effect of missing data and missing data treatments on type I error and statistical power for DIF detection in polytomous items, in the context of attitudinal assessment, and using small scales. For example, Finch's (2011) study of the effect of missing data on the detection of uniform DIF used only 100 replications and conducted a study for DIF in binary items. Robitzsch and Rupp (2009) conducted their study using also binary items. In addition, both studies consisted of 40, and 20 and 40 binary items respectively. In both studies, only one item was the designated DIF item and also the targeted item for manipulating missingness. In the Robitzsch and Rupp's (2009) study, the detection of DIF by the methods used (M-H and LR) was by estimating BIAS and RMSE as outcomes. Thus, results from Robitzsch and Rupp's study were not used for discussion and comparison with the present study. The research design of this study was more aligned to Garrett's (2009) study in that this study and Garrett's evaluated the effect of missing data on the performance of procedures for detecting uniform DIF in polytomous items under similar DIF magnitudes. Additionally, the same type of missingness was implemented (MCAR), and ability distributions or impact had the same values for the reference and focal groups. Differences to note between this study and Garrett's were the number of items or test length, which for the Garrett's study was set to 20 while this study considered small scales (4-, 5-, and 6-item scales) which is common in attitude assessment. It is important to note that Garrett generated the response data from the parameters of a previous study which consisted of only 14 items and in order to increase the number of items to 20, 6 items were repeated. This practice of repeating items so more items are included in a scale is not psychometrically sound because items with the same parameters are literally the same items and do not add anything to the scale. Finally, in terms of total sample size, Finch, used only balanced groups for their DIF analyses, while Garrett in addition to a balanced sample size condition, included also unbalanced groups.

Discussion of Findings: Type I Error

To analyze the effect of missing data and MDM on the Type I error of the IRT-LR test for DIF detection, the overall distributions of the rejections rates were examined for the χ^2 and Bonferroni adjusted nominal level. The distributions of Type I error rates by MDM across all factors showed that when the nominal level of significance was not protected (i.e., using χ^2 critical values for the significance tests), the error rates were inflated. Thus, only the rejection rates obtained using the Bonferroni adjustment to the nominal levels were further analyzed. The results of this study for Type I error were consistent across nominal alpha levels and effect size, and entirely consistent across MDM. The DIF analyses conducted using complete data (i.e., as if no data are missing), the FIML and the person mean substitution method (PMS) were very consistent. That is, the overall Type I error rates of the IRT-LR test were at the nominal level of significance (for both α = .01 and α = .05) under complete data, FIML, and PMS, and little variability was observed in the overall Type I error distributions across all factors. An analysis of effect size suggested that none of the factors had an effect size $\eta^2 \geq .05$ for complete data and ability distribution had a moderate effect on Type I error rates when interacting with the proportion of missing observations under FIML, and when interacting with the number of items in PSM. Sample size and the proportion of missing observations had also a moderate effect on the Type I error rates for Listwise in this study. Under different research conditions, this same interaction was also found in Finch's study.

Although Finch's (2011) and Garrett's (2009) studies were not comparable to the present study in terms of the type of data analyzed and methods for DIF detection, their Type I error results and effect sizes were also consistent within each study research design. Take for instance their estimated Type I error rates for complete data. For all the methods Finch and Garrett examined (M-H, LR, SIBTEST, and Mantel test and OLR respectively), Type I error rates were also near the nominal level when α = .05 and the reference and focal groups had the same ability distribution. For Garrett's, Type I error rates were higher than .05 when groups differed in ability. The same consistency in effect size results was observed in these studies, in that none of the factors had an effect on Type I error rates when the DIF analysis was performed using complete data. Garrett also examined the PMS under the name of WMS (within person-mean-substitution). At the significance level examined in Garrett's, the WMS had rejection rates slightly above the nominal level for both DIF detection methods (Mantel and OLR), and in addition, rejection rates were slightly larger with higher proportions of missing data. In Finch's study, it is worth to mention that under Listwise deletion, the interaction of sample size and proportion of missing observations had also an effect on Type I error rates.

One missing data method implemented in the present study and also in Finch's and Garrett's was multiple imputation (MI). As explained earlier, in this study Type I error analyses for MI were conducted using the rejection rates for the Bonferroni adjustment for significance. Under significance by the Bonferroni adjustment, the MI method had entirely consistent results for both levels of significance. Across α = .01 and α = .05, 1) MI had the largest mean error rate across methods. 2) The distributions of Type I error rates for MI across all factors were upwardly dispersed, more so when $\alpha = .05$ (max = 0.07 and max = 0.20 for $\alpha = .01$ and $\alpha = .05$ respectively). 3) The interaction of the proportion of missing observations and the proportion of missing items had a large effect on the Type I error rates under MI, with mean error rates for the proportion of missing observations increasing with the increase of missing items. 4) MI had the smallest proportions meeting Bradley's (1978) criteria for robustness on each factor. In Finch's (2011), MI treated missing data but the type of data used (dichotomous items), the large number of items (40 items), and the DIF methods used (M-H, SIBTEST, and LR), made it difficult to compare the effect of MI on the Type I error rates of his study with the Type I error rates obtained in the present study. However, it is worthtly to mention that in Finch's study for identifying the DIF item, the DIF methods examined had Type I error rates near $\alpha = .05$ under MI, when missing data were MCAR, and groups had the same ability distribution regardless of the proportion of missing data (5% and 15%). But error rates tended to be above .05 for the smallest sample size (250 / 250). Garrett's DIF methods for identifying the DIF item had Type I error rates below the nominal level when MI was used to treat missing data. But there were differences by DIF methods and the proportion of missing data. Surprisingly, rejections rates decreased as the proportion of missing data increased, with rejections rates slightly higher for the Mantel DIF method when groups differed in ability distribution.

The present study also examined MDM that were developed for imputing missing data in Likert-type data specifically. Single regression substitution (SRS) bases its effectiveness in substituting missing values in Likert-type items on the fact that Likert-type items are correlated to some degree. Thus, for the SRS, the observed value most correlated with the missing value is used to predict the missing item response. However, this method was not effective in the context of this study. 1) The mean Type I error rates of the SRS were above the nominal levels and the distribution of error rates were substantially dispersed out toward higher rejection rates. 2) SRS had smaller proportions of conditions having adequate Type I error control than the other MDM (except for MI, which had the smallest proportions). 3) Under the SRS, several interactions of factors had an effect of Type I error rates, from moderate to large (η^2 = .09 to η^2 = .13). Markedly, Type I error rates were larger at higher proportions of missing observations and missing items as well as when the number of items was 4.

The relative mean substitution (RMS) was another MDM developed for missing data in Likert-type data. This MDM bases its effectiveness in estimating an item missing value using three sources of information, including the grand mean of all valid item scores. Considering that in DIF analyses at least two groups are compared, the imputation of missing values was implemented separately by each group in the study before conducting the DIF detection. The Type I error results for RSM were similar to the results of SRS across nominal levels. Mean Type I error rates were at the nominal level (but slightly higher when α = .05) and the distribution of rejection rates were dispersed toward higher rejection rates. The analysis of effect size was entirely consistent in both SRS and RMS although the RMS had higher proportions of conditions meeting the Bradley's criteria for robustness than the SRS.

The results of this study are difficult to compare with the reviewed literature on DIF detection in the presence of missing data because while all of them evaluated the efficacy of the methods studied in terms of Type I error rates, these rates were not adjusted to meet a criteria for robustness; thus, the reported Type I error rates in the literature reviewed cannot be meaningfully interpreted. But within each research context, in the present study and in the studies reviewed, results for Type I error control were very similar, across the DIF methods and MDM implemented when missing data was MCAR. In each study, consistency of Type I error results was observed regardless the factors under which the effect of missing data on Type I error was investigated. In the case of the IRT-LR DIF detection method, the inflated Type I error rates and the large effect of the interaction of the proportion of missing observations and the proportion missing items suggested that MI and SRS were the less effective MDM for treating missing data in the Likert type items in the short scales examined in this study. But MI seemed to be an effective MDM for treating missing data in dichotomous items in large sets of items as evidenced in the Type I error rates of the M, M-H and logistic regression DIF detection methods for dichotomous items, which were at the nominal level when missing data was MCAR, and the number of items was large.

As for Listwise deletion and within each study research context, this MDM had similar performance with both dichotomous and polytomous items. While the Type I error rates for the Listwise deletion method were at the nominal levels of significance, the number of items had an effect on this method which suggested that even few items impact the performance of this MDM in the context of DIF. As expected, sample size and the proportion of missing observations had an impact on the Type I error rates for Listwise deletion with more noticeable differences in rejections rates for the small total sample size (*N* = 500).

Discussion of Findings: Statistical Power

In this study, power comparisons across MDM were made only over conditions that had adequate Type I error control and that were present in all MDM. The power of the IRT-LR method for detecting DIF was very consistent across levels of significance, factors, and MDM. Across all MDM, the power of the IRT-LR test for DIF detection was very consistent across DIF magnitude (e.g., low power was observed for detecting DIF = .25 and large power estimates were observed for detecting DIF = .75). Within this study, only the interaction of the number of items and sample size had a large effect on power estimates, across all MDM. Entirely consistent with theory, in this study and in Finch's and Garrett's, sample size had an effect on the power estimates across all MDM. That is, power decreased as sample size decreased, regardless of the proportion of missing data. Within Finch's (2011) and Garrett's (2009) research conditions, power estimates across the methods investigated were consistent across DIF magnitude; that is, low power for small DIF and power increases were observed with increases of DIF magnitude).

Compared to M, M-H, LR, and SIBTEST, whose effectiveness for detecting DIF has been amply demonstrated (Finch, 2011), the IRT-LR on the other hand has been less implemented due to the complexity of testing nested hypothesis. The software needed to run IRT-LR was also more specialized than common statistical packages. . While there have been recent studies for fitting IRT models using SAS (e.g., Chen, Li, & Kromrey, 2013) and research on the IRT-LR DIF detection method should be encouraged, the implementation of the IRT-LR has not really taken off.

Last Thoughts: Recommendations and Future Research

Tukey (192), speaking of the future of data analysis, said that we have watched mathematical statistics evolve. As data have become more complex, statistical methods have become more complex too. However, the same challenges that Tukey observed, are still crucial today. The challenges of incomplete data are still not met and as Tukey recommended, one should be willing to work with data as it exists. The recommendation of conducting simulation studies with field data is not new, though. Kromrey and Hines (1991), for example, recommended that research on missing data should reflect data characteristics as observed in the field. Emenogu's (2006) study clearly showed the difference between using computer generated data and using field data. For instance, Emenogu found 12 items exhibiting DIF in a set of 41 items. Moreover, the type of DIF items identified also varied, with some items favoring the focal group and some items favoring the reference group. Additionally, DIF items varied in the magnitude of DIF exhibited. Missing data also varied across variables. Thus, it is clear that conducting a DIF study with such large number of items (20 or 40) and manipulating just one item, does not reflect what one would be observing in reality. That is, we should leave our comfort zone and "tackle old problems in more realistic frameworks" (Tukey, 1962; p. 4). . In sum, when treating missing data, the selection of a missing data method should be done according to the research problem at hand. Thus, researchers doing research on DIF must screen their data for missingness before conducting any analysis. However, in the context of short scales, the proportion of missing values in the variables under study is a relevant factor on the performance of a DIF method and should provide a good indicator as to the amount of data that would be lost if Listwise deletion is implemented. Also, the number of items is also a relevant factor as even one more item improves the performance of the DIF method. Although there is no such thing as a best method per se, in the context of DIF and short scales, using all the information available implementing FIML seems to be the best option for handling missing data in this scenario. Although MI is advocated in the literature, caution is recommended when used to treat missing data in Likert-type scales. The Rounding needed will infuse upward or downward bias to the variability of the plausible values imputed.

The following topics are recommended for further research. First, real data are likely to have other missing data mechanisms present in addition to MCAR. Thus, in addition to MCAR missingness, research on MAR and MNAR mechanisms are recommended. The problem of rounding in missing data research should be studied further. For example, one of the main advantages of MI is that this method captures for the uncertainty of the missing data. That is, the imputed value is not “the value” but an estimate of the value and this uncertainty is captured in the variability that is estimated across the multiple imputations. However, when the plausible imputed values are forced to round so they fit how Likert-type items are scored, this rounding up or down biases the estimates across imputations. In addition, in the context of DIF, further research is needed to investigate whether contextual variables improve the accuracy and precision of imputation in short scales

REFERENCES

Ake, C. F. (2005). Rounding after multiple imputation with non-binary categorical covariates. *Proceedings of the Thirtieth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute, Inc.

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Long Grove, IL: Waveland Press, Inc.

Allison, P. D. (2001). *Missing data.* Sage University Papers Series on Quantitative Applications in the Social Sciences. Thousand Oaks, CA: Sage.

Allison, P. D. (2005). Imputation of categorical variables with PROC MI. *Proceedings of the Thirtieth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute, Inc.

Alwin, D. F. (1992). Information transmission in the survey interview: Number of response categories and the reliability of attitude measurement. *Sociological Methodology, 22,* 83-118.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrich, D. (2002). Implications and applications of modern test theory in the context of outcomes based education. *Studies in Educational Education, 28*, 103-121.

Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics*, *37*(3), 387-416.

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement*, *36*(4), 277-300.

Baker, F. B. (1977). Advances in item analysis. *Review of Educational Research*, *47*(1), 151-178.

Baker, F. B., & Kim, S-O. (2004). *Item response theory: Parameter estimation techiques*. New York, NY: Marcel Dekker, Inc.

Baldi, S., Perie, M., Skidmore, D., Greenberg, E., & Hahn, C. (2001). *What democracy means to ninth graders: U.S. results from the international IEA Civic Education Study* (NCES-2001-096). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Balsis, S., Gleason, M.E. J., Woods, C. E., & Oltmanns, T. F. (2007). An item response theory analysis of DSM-IV personality disorder criteria across younger and older age groups. *Psychology and Aging, 22*(1), 171-185.

Beale, E. M. L., & Little, R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society. Series B(Methodological), 37*(1), 129-145.

Benson, P. H. (1971). How many scales and how many categories shall we use in consumer research?—A comment. *Journal of Marketing, 35,* 56-61.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*(2), 179-197.

Bock, R. D., Mislevy, R., & Woodson, C. (1982). The next stage in educational assessment. *Educational Researcher, 11*(3), 4-16.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152

Brockmeier, L. L., Kromrey, J. D., & Hogarty, K. Y. (2003). Nonrandomly missing data in multiple regression analysis: An empirical comparison of ten missing data treatments. *Multiple Linear Regression Viewpoints*, *29*(1), 8-20.

Brophy, J., & VanSledright, B. (1997). *Teaching and learning history in elementary schools*. New York, NY: Teacher College Press

Buhi, E. R., Goodson, P., Neilands, T. B. (2008). Out of sight, not out of mind: Strategies for handling missing data. *American Journal of Health Behavior*, *32*(1), 83-92.

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

Carey, L. M. (2001). *Measuring and evaluating school learning* (4th Edition). Boston, MA: Pearson Custom Publishing.

Chen, Y-H., Li, I., & Kromrey, D. K. (2013). *GLIMMIX_Rasch: A SAS macro for fitting the dichotomous Rasch model*. Proceedings of the South East SAS Users conference. Cary, NC: SAS Institute, Inc.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretation of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31-43.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential functioning tests items. *Educational Measurement: Issues and Practice, 17*(1), 31-44.

Clearly, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*(2), 115-124.

Clearly, T. A., & Hilton, T. L. (1966). *An investigation of item bias*. Research Bulletin RB-66-17. Princeton, NJ: Educational Testing Service.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98-101.

Cohen, A. S., Kane, M. T., & Kim, S-H. (2001). The precision of simulation studies. *Applied Psychological Measurement, 25*(2), 136-145).

Cohen, A. S., Kim, S.-H., Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*(4), 335-350.

Cohen, A. S., Kim, S-H., & Wollack, J. A. (1998). A comparison of item response theory and observed score DIF detection measures for the graded response model. *Paper presented at the Annual Meeting of the National Council on Measurement in Education*, San Diego, CA.

Coleman, J. S. (1966). *Equality of educational opportunity*. Washington, D.C.: Government Printing Office.

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

Cronbach, L. J., & Gleser, G.(1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.

Davey, A., Savla, J., & Luo, Z. (2005). Issues in evaluating model fit with missing data. *Structural Equation Modeling*, *12*(4), 578-597.

De Ayala, R. J., & Sava-Bolesta M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement, 23*(1), 3-19.

de la Torre, J., & Hong, Y. (2010). Parameter Estimation with small sample size: A higher-order IRT model approach. *Applied Psychological Measurement*, *34*(4), 267-285.

DeMars, C. (2003). Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement*, *27*(4), 274-288.

Dodd, B, G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, *19*(1), 5-22.

Dodd, B. G., Koch, W., & De Ayala, R. J. (1989). Opeational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Testing*, *13*(2), 129-143.

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*(1), 43-68.

Downey, R. G., & King, C. (1998). Missing data in Likert ratings: A comparison of replacement methods. *The Journal of General Psychology*, *125*(2), 175-191.

Downing, S. M. (2003). Item response theory: Applications of modern test theory in medical education. *Medical Education*, *37,* 739-745.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous Item Response Theory models to multiple-choice tests. *Applied Psychological Measurement, 19*(2), 143-165.

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist, 16,* 640-647.

Embretson, S. E. (2004). The second century of ability testing: Some predictions and speculations. *Measurement, 2*(1), 1-32.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NY: Lawrence Erlbaum Associates, Publishers.

Emenogu, B. C. (2006). *The effect of missing data treatment on Mantel-Haenszel DIF detection.* (Unpublished dissertation). University of Toronto, Ontario, Canada.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381.

Finch, H. (2011). The use of multiple imputation for missing data in uniform DIF analysis: Power and Type I error rates. *Applied Measurement in Education, 24*(4), 281-301.

Finch, H. (2011). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and Psychological Measurement, 71*(4), 663-683.

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice, 24*(3), 21-28.

Garcia, G. E., & Pearson, P. D. (1994). *Assessment and diversity*. In L. Darling-Hammond (Ed.), *Review of Research in Education, 20* (pp.337-391). Washington, DC: American Educational Research Association.

Garrett, P. L. (2009). *A Monte Carlo study investigating missing data, differential item functioning, and effect.* (Doctoral dissertation). Georgia State University, Atlanta, GA. Retrieved from ProQuest LLC. (UMI 3401601).

Gold, M. S., & Bentler, P. M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, Iterative Stochastic regression Imputation, and Expectation-Maximization. *Structural Equation Modeling, 7*(3), 319-355.

Graham, J. W., & Hofer, S. M. (2000). Multiple imputation in multivariate research. In T. D. Little, K. U. Schanaber, & J. Baumert (Eds.). *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 201-218). Mahwah, NJ: Lawrence Erlbaum Associate Publishers

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York: American Council on Education.

Hambleton, R. K. (2004). Theory, methods, and practices in testing for the 21st century. *Psicothema, 16*(4), 696-701.

Hambleton, R., K., & Slater, S. C. (1997). Item response theory models and testing practices: Current international status and future directions. *European Journal of Psychological Assessment, 13*(1), 21-28.

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

Harel, O., & Zhou, X-H. (2007). Multiple imputation: Review of theory, implementation, and software. *Statistics in Medicine*, *26,* 3057-3077.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational Statistics, 13*(3), 243-271.

Harwell, M., Stone, C. A., Hsu, T-C, & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101-125.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology, 67*(6), 818-825.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39*, 1-123.

Kane, M. (1996). The precision of measurements. *Applied Measurement in Education, 9*(4), 355-379.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. (2013). Validating the interpretation and uses of tests scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov Chain Monte Carlo estimation. *Applied Psychological Measurement, 36*(5), 399-419.

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

Kim, S-O, & Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods, 6*(2), 215-240.

Kline, P. (2000). *Handbook of psychological testing* (2nd ed.). New York, NY: Routledge.

Koch, W. R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement*, *7*(1), 15-32.

Kromrey, J. D., & Bacon, T. P. (1992). Item analysis of achievement tests based on small numbers of examinees. *Paper presented at the annual meeting of the American Educational Research Association*, San Francisco, CA.

Kromrey, J. D., & Dickinson, W. B. (1995). The use of an overall F test to control Type I error rates in factorial analyses of variance: Limitations and better strategies. *Journal of Applied Behavioral Science, 31*(1), 51-64.

Kromrey, J. D., & Hines, C. V. (1994). Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments. *Educational and Psychological Measurements, 54*(3), 573-593.

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement, 33*(1), 71-92.

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1995). Examining the assumptions and properties of the graded response model: An example using a mathematics performance assessment. *Applied Measurement in Education, 8*(4), 313-340.

Lin, C.-J. (2008). Comparisons between classical test theory and item response theory in automated assembly of parallel test forms. *The Journal of Technology, Learning, and Assessment, 6*(8). Retrieved from http://www.jtla.org.

Linn, R. L. (1990). Has item response theory increased the validity of achievement test scores? *Applied Measurement in Education, 3*(2), 115-141.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Little, R. J. A. (July, 1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, *6*(3), 287-296.

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

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719-729.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

McDonald, R. P. (1999). *Test theory: A unified treatment.* New York, NY: Routledge.

McKennell, A. C. (1974). Surveying attitude structures: A discussion of principles and procedures. *Quality and Quantity, 7*(2), 203-296.

McPhee, D., Kreutzer, J. C., & Fritz, J. J. (1994). Infusing a diversity perspective into human development courses. *Child Development, 65,* 699-715.

Mellenbergh, G. J. (1982). Contingent table models for assessing bias. *Journal of Educational and Behavioral Statistics, 7*(2), 105-118.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*(2), 127-147.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*(3), 215-237.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–100). Washington, DC: American Council on Education.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association, 44(247), 335-341.*

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297-334.

Mislevy, R. J. (1989). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, and I. I. Bejar (Eds.). *Test theory for a new generation of tests* (pp. 19-77). Hillsdale: NJ, Lawrence Erlbaum Associates, Inc., Publishers.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education, 30,* 109-162.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 International report: IEA's study of reading literacy achievement in primary schools.* Chestnut Hill, MA: Boston College.

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

No Child Left Behind (NCLB) Act of 2001. (2002). Pub. L. No. 107-110, § 115, Stat. 1425.

O'Rouke, T. W. (2003). Methodological techniques for dealing with missing data. *American Journal of Health Studies, 18*(2/3), 165-168.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning.* Thousand Oaks, CA: Sage Publications.

Paek, I., & Guo, H. (2011). Accuracy of DIF estimates and power in unbalanced designs using the Mantel-Haenszel DIF detection Procedure. *Applied Psychological Measurement, 35*(7), 518-535.

Patrician, A. P. (2002). Multiple imputation for missing data. *Research in Nursing Health, 25,* 76-84.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8*(2), 287-312.

Progar, Š., & Sočan, G. (2008). An empirical comparison of item response theory and classical test theory. *Horizons of Psychology, 17*(3), 5-24.

Raaijmakers, Q. A. (1999). Effectiveness of different missing data treatments in surveys with Likert-type data: Introducing the relative mean substitution approach. *Educational and Psychological Measurement, 59*(5), 725-748.

Raymond, M. R. (1987). Missing data in evaluation research. *Evaluation & the Health Professions*, *19*(4), 395-420.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45,* 283-288.

Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and logistic regression analysis. *Educational and Psychological Measurement, 69*(1), 18-34.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47,* 537-560.

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Willey & Sons.

Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When Pearson correlations are not enough. *Educational and Psychological Measurement, 65*(4), 588-599.

Ryan, T. A. (1959). Multiple comparisons in psychological research. *Psychological Bulletin, 56*(1), 26-47.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17, 34*(4, Pt. 2).

Samejima, F. (2010). The general graded response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory* (pp. 77-107), New York, NY: Routledge, Taylor & Francis Group.

SAS Institute Inc. (2012). SAS (version 9.3)[Computer Software]. Cary, NC: SAS Institute Inc.

Saville, D. J. (1990). Multiple comparison procedures: The practical solution. *American Statistician, 44,* 174-180.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*(4), 545-571.

Schulte Nordholt, E. (1998). Imputation: Methods, simulation experiments and practical examples. *International Statistical Review, 66*(2), 157-180.

Schulz, W. (2004). Scaling procedures for Likert-type items on students' concepts, attitudes, and actions. In W. Schulz & H. Sibberns (Eds.), *IEA Civic Education Study technical report* (pp. 93–126). Amsterdam, The Netherlands: Paula Wagemaker Editorial Services.

Sijtsma, K., & Junker, B. W. (2006). Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika, 33*(1), 75-102.

Sinharay, S., Haberman, S. J., & Lee, Y-H. (2011). When does scale anchoring work? A case study. *Journal of Educational Measurement, 48*(1), 61-80.

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology, 81*(2), 332-342.

Stark, S. (2007). *MODFIT: A computer program for model-data fit* [Computer program v2.0]. Urbana–Champaign: University of Illinois.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, New Series, 103,* 677-680.

Stodolsky, S. S. (1988). *The subject matters: Classroom activity in math and social studies*. Chicago, IL: The University of Chicago Press.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*(1), 1-16.

Stone, C. A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education, 16*(1), 1-16.

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

Thissen, D. (2003). *MULTILOG 7.03 User's guide: Multiple categorical item analysis and test scoring using item response rheory*. Mooresville, IN: Scientific Software International.

Thissen, D., Steinberg, L., Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*(1), 118-128.

Thissen, D., Steinberg, L., Pyszcznski, T., & Greenberg, J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. *Applied Psychological Measurement*, *7*(2), 211-226.

Thissen, D., Steinberg, L., & Wainer, H. (1988). *Use of item response theory in the study of group differences in trace lines*. In H. Wainer & H. I. Braun (Eds.). *Test Validity,* (pp. 147-170). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Thissen, D., Steinberg, L., & Wainer, H. (1993). *Detection of differential item functioning using the parameters of item response models.* In P. W. Holland, & H. Wainer (Eds.). *Differential item functioning,* (pp. 130-215). Hillsdale, England: Lawrence Erlbaum Associates, Inc.

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement, 8*, 63-70.

Toothaker, L. E. (1991). *Multiple comparisons for researchers.* Newbury Park, CA: Sage Publications, Inc.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 1-67.

Tyler, R. W. (1951). The functions of measurement in improving instruction. In E. F. Lindquist, (Ed.). *Educational Measurement,* (pp. 47-67). Washington, DC: American Council on Education.

U.S. Department of Education, Institute of Education Sciences. (2002). *IEA Civic Education Study* (CD-ROM). Washington, DC: National Center for Education Statistics.

U.S. Department of Education. (2002). No Child Left Behind: Title I – Improving the academic achievement of the disadvantaged. Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/pg1.html#sec101

Vogler, K.E., Lintner, T., Lipscomb, G. B., Knopf, H., Heafner, T. L., & Rock, T. C. (2007). Getting off the back burner: The impact of testing elementary social studies as part of state-mandated accountability program. *Journal of Social Studies, 31*(2), 20-34.

Vogler, K. E., & Virtue, D. (2007). Just the Facts, Ma'am: Teaching social studies in the era of standards and high-stakes testing. *The Social Studies, 98*(2), 54-58.

Wang, W.-C., & Chen, T.-E. (2005). Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. *Educational and Psychological Measurement, 65*(3), 376-404.

Ward, A. W., Stoker, H. W., & Murray-Ward, M. (Eds.). (1996). *Educational measurement: Origins, theories, and explanations.* Boston: University Press of America, Inc.

Weisberg, H. I. (2014). *Willful ignorance*. Hoboken, NJ: John Willey & Sons, Inc.

Whittaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models [manual and program].

Widaman, K. F. (2006). Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development, 7*(13), 42-64.

Wiberg, M. (2004). *Classical test theory vs. item response theory: An evaluation of the theory test in the Swedish driving-license test*. (No. 29). Umeå University: Department of Educational Measurement.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychological journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Woods, C. M. (2008). Likelihood-ratio DIF testing: Effects of nonnormality. *Applied Psychological Measurement, 32*(7), 511-526.

Wright, B. D. (1968). Sample free test calibration and person measurement. *Proceedings of the 1967 ETS invitational conference on testing problems,* Princeton, NJ: Educational Testing Service.

Yeager, E. A., & Davis, O. L. (2005). *Wise social studies teaching in an age of high-stakes testing: Essays on classroom practices and possibilities*. Greenwich, CT: Information Age Publishing.

Yenduri, S., & Iyengar, S. S. (1994). Performance evaluation of imputation methods for incomplete datasets. *International Journal of Software Engineering and Knowledge Engineering, 17*(1), 127-152.

Zhang, P. (2003). Multiple imputation: Theory and method. *International Statistical Review, 71*(3), 581-592.

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., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education*, *19*(4), 321-344.

APPENDIX A

TESTING ITEM RESPONSE THEORY ASSUMPTIONS

Within IRT or modern test theory, items' parameters and examinees' abilities determine the probability of responding correctly or endorsing a given item. The item-free and person-free parameters, along with the item parameter invariance property of IRT models are important characteristics of this measurement theory, and are the main advantages of IRT over CTT. While IRT measurement models could be selected on sound theoretical grounds, the effectiveness of a given IRT model requires that the data fit the selected IRT model, and that model assumptions of dimensionality and local independence are met. Thus, to ensure the precision required for conducting this dissertation study, both assumptions were tested. Table A1 summarizes the descriptive characteristics of the subscales' items.

Table A1

Test Statistics for Subscales G, H, and J

Subscale	Mean	SD	Reliability	Avg. Interitem
G	20.44	3.49	.81	0.72
H	15.80	3.12	.85	0.79
J	12.70	2.32	.78	0.78

Source: *Civics Education Study*. Copyright © 2002 International Association for the Evaluation Educational Achievement (IEA). Publisher: National Center for Education Statistics,

Model Data Fit

The fit of the data generated from the Civics Education study (IEA, 1999) to the IRT GRM was assessed using the MODFIT (Stark, 2007) software. The CivEd subscales for this dissertation study (G, H, and J) were selected in terms of items having the same response type. That is, the items of the selected subscales have the same four response options or categories type (i.e., strongly disagree, disagree, agree, and strongly agree). The six items of subscale G addressed 14-year old U.S. students' positive attitudes toward opportunities which members of certain groups should have in the United States (i.e., women's political and economic rights). Item G1 from the subscale G (Opportunities) is presented to exemplify the response type (Source: Appendix F: The CIVED instruments, p. 255; Schulz & Sibberns, 2004).

G1 Women should run for public office [a seat in the legislature] and take part in the government just as men do?

1 *Strongly disagree*
2 *Disagree*
3 *Agree*
4 *Strongly disagree*

Subscale H included five items measuring 14-year old U.S. students' positive attitudes toward immigration. Item H1 from the subscale H (Opportunities) is presented to exemplify the response type (Source: Appendix F: The CIVED instruments, p. 257; Schulz & Sibberns, 2004).

H1 Immigrants should have the opportunity [option] to keep [continue speaking] their own language

1 *Strongly disagree*
2 *Disagree*
3 *Agree*
4 *Strongly disagree*

Subscale J included four items measuring 14-year old U.S. students' positive attitudes toward participation in school life. Item J1 from the subscale J (Opportunities) is presented to exemplify the response type (Source: Appendix F: The CIVED instruments, p. 259; Schulz & Sibberns, 2004).,

J1 Electing student representatives to suggest changes in how the school is run [how to solve school problems] makes schools better

1 *Strongly disagree*
2 *Disagree*
3 *Agree*
4 *Strongly disagree*

The graphical model fit of data generated from subscales G, H, and J was examined by observing the theoretical ORF's and empirical ORF's of each item and fit plots for each item response option. Figures A1, A2, and A3 show the constructed items' ORF's plots as well as the fit plots for each scale item's options. In each plot constructed, the correspondence of both ORF's is well within the 95% confidence interval, which suggests sufficient fit of the CivEd data to the GRM. Tests of goodness-of-fit chi-square fit statistics provided by MODFIT (Stark, 2007) are summarized in Table A1.

The magnitude of the ratios of χ^2 to their degrees of freedom for each subscale items are summarized in four intervals (<1, 1<2, 2<3, and 3<4), where ratios were considered very small, small, moderately large, and large respectively. For single items, the mean of χ^2 /df for the subscale G (6 items) was 1.46, and the mean of χ^2 /df for the subscale H (5 items) was 0.92, which was smaller than the mean of χ^2 /df for subscale J (4-items), 1.25. Thus better fit seems to be shown for subscale H, with 4 items of such scale within the very small range for singlets.

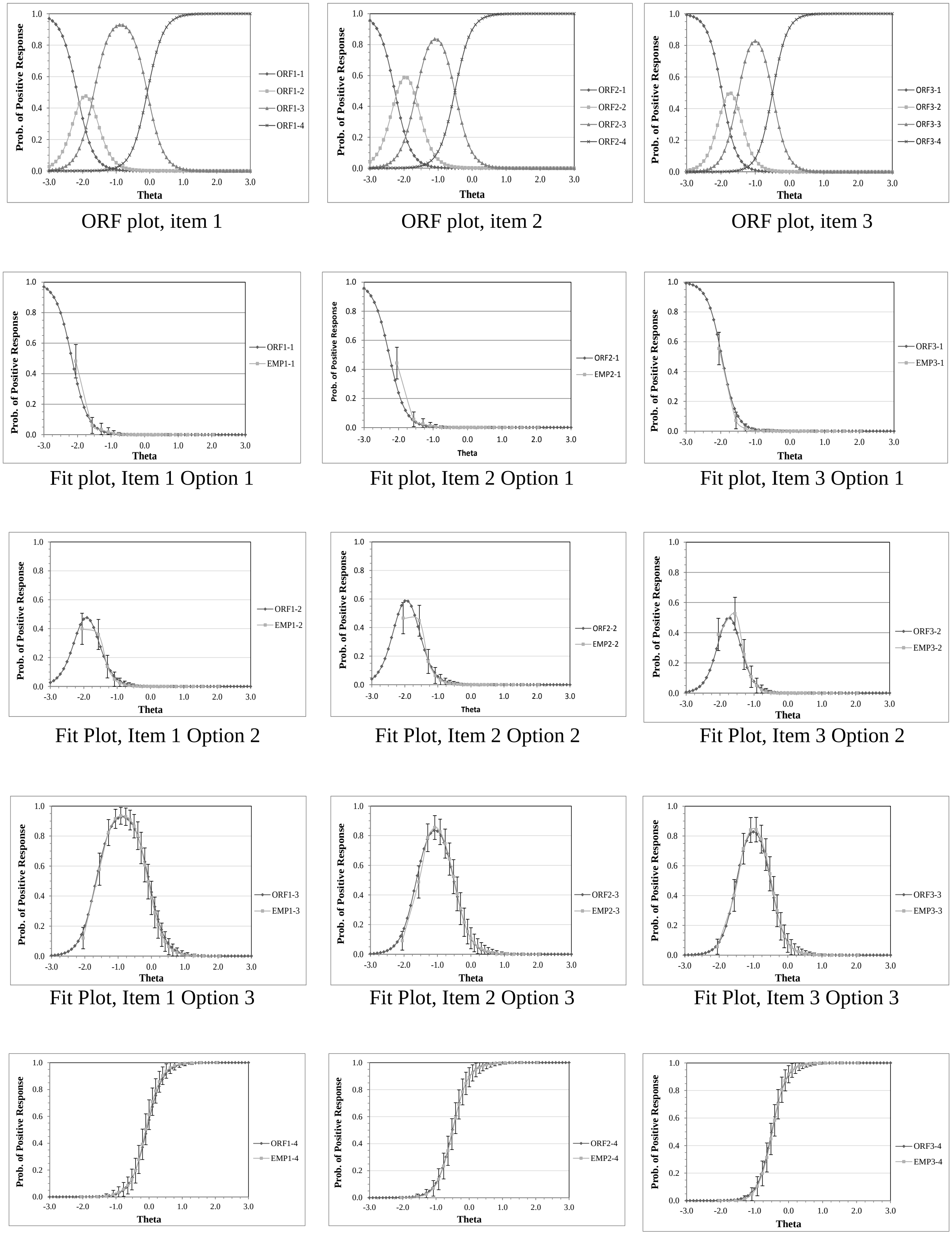
When χ^2/df was computed for doublets and triplets, subscale H has more items within the very small range compared to the number of items in this range for subscales G and F.

Table A2

Summary of Frequencies, Means, and Standard Deviations of χ^2/df Ratios (*GRM*)

	Frequency of χ^2/df Ratios					
Subscales	< 1	1 < 2	2 < 3	3 < 4	*Mean*	*SD*
G (6 items)						
Singlets	**1**	**4**	**1**	0	**1.46**	0.67
Doublets	1	11	3	0	1.67	0.49
Triplets	1	16	3	0	1.53	0.37
H (5 items)						
Singlets	**4**	0	**1**	0	**0.96**	0.86
Doublets	3	5	3	0	1.25	0.54
Triplets	2	8	0	0	1.25	0.31
J (4 items)						
Singlets	**1**	**1**	**1**	**1**	**2.03**	1.39
Doublets	0	4	2	0	1.86	0.38
Triplets	0	4	0	0	1.62	0.16

Note: Because single items are insensitive to unidimensionality and consequently insensitive to local independence violations, MODFIT (Stark, 2007) also computes χ^2 for pairs of items and triples (3-item sets) as recommended in Drasgow et al. (1995) by adjusting each χ^2 to the magnitude that would be expected in a sample of n=3000 then dividing by its degrees of freedom (*df*). Adjusted χ^2/df ratios < 3 indicate good model-data fit



ORF plot, item 1 | ORF plot, item 2 | ORF plot, item 3

Fit plot, Item 1 Option 1 | Fit plot, Item 2 Option 1 | Fit plot, Item 3 Option 1

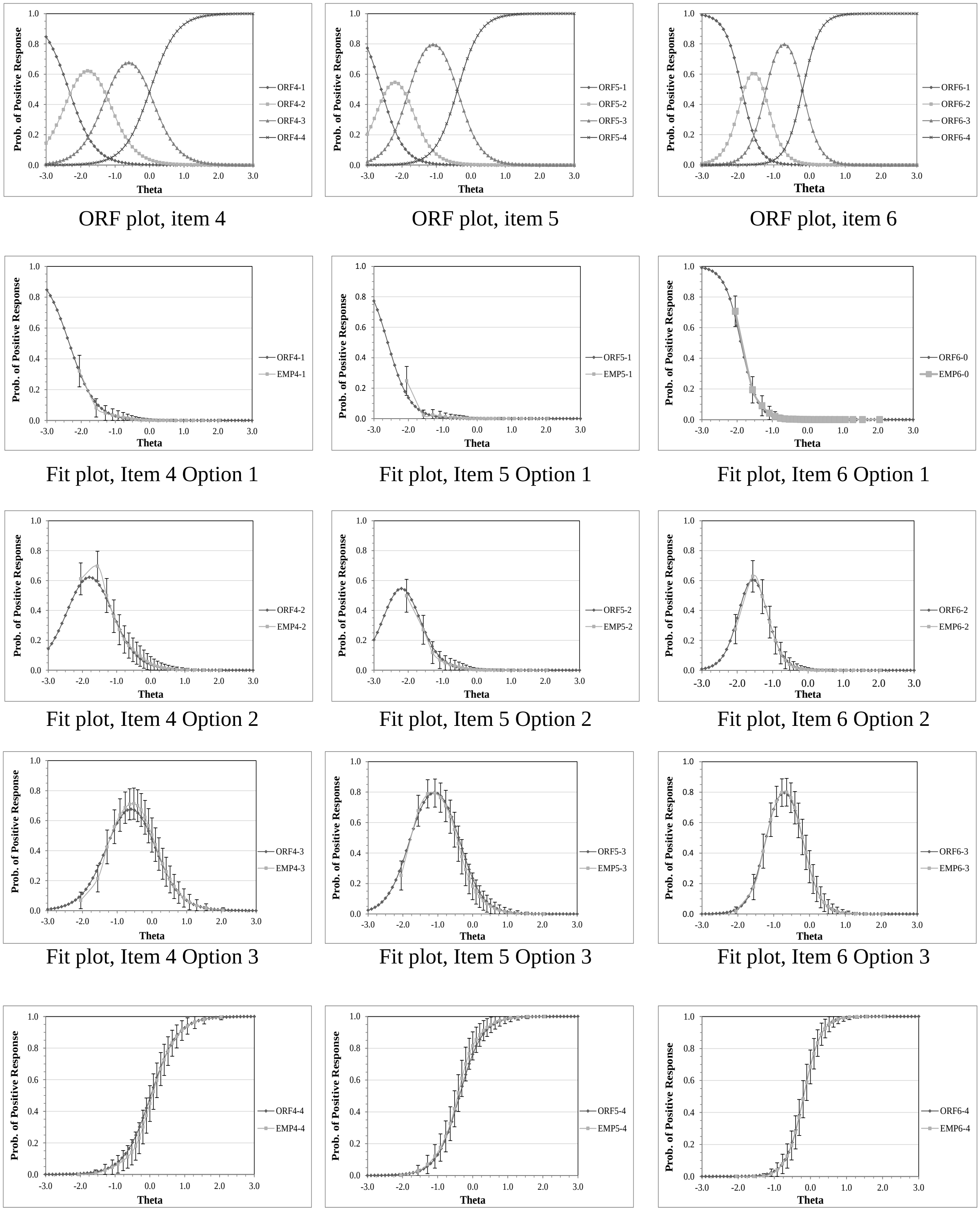
Fit Plot, Item 1 Option 2 | Fit Plot, Item 2 Option 2 | Fit Plot, Item 3 Option 2

Fit Plot, Item 1 Option 3 | Fit Plot, Item 2 Option 3 | Fit Plot, Item 3 Option 3

Fit Plot, Item 1 Option 4 | Fit Plot, Item 2 Option 4 | Fit Plot, Item 3 Option 4

Figure A1. Item ORF’s and fit plots for scale G items 1 to 3

Figure A1 (Cont')



ORF plot, item 4

ORF plot, item 5

ORF plot, item 6

Fit plot, Item 4 Option 1

Fit plot, Item 5 Option 1

Fit plot, Item 6 Option 1

Fit plot, Item 4 Option 2

Fit plot, Item 5 Option 2

Fit plot, Item 6 Option 2

Fit plot, Item 4 Option 3

Fit plot, Item 5 Option 3

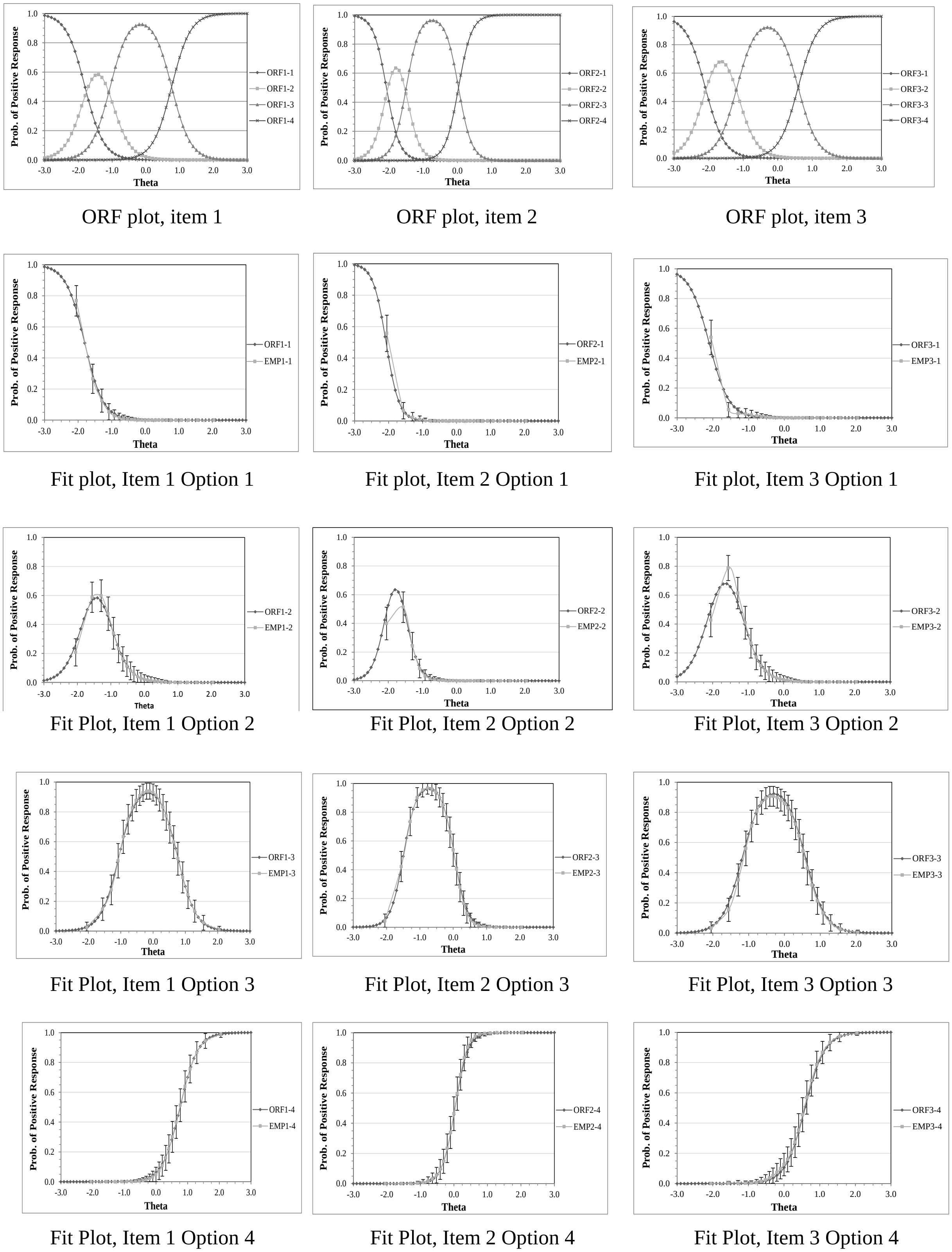
Fit plot, Item 6 Option 3

Fit plot, Item 4 Option 4

Fit plot, Item 5 Option 4

Fit plot, Item 6 Option 4

Figure A1. Item ORF's and fit plots for scale G items 4 to 6



ORF plot, item 1

ORF plot, item 2

ORF plot, item 3

Fit plot, Item 1 Option 1

Fit plot, Item 2 Option 1

Fit plot, Item 3 Option 1

Fit Plot, Item 1 Option 2

Fit Plot, Item 2 Option 2

Fit Plot, Item 3 Option 2

Fit Plot, Item 1 Option 3

Fit Plot, Item 2 Option 3

Fit Plot, Item 3 Option 3

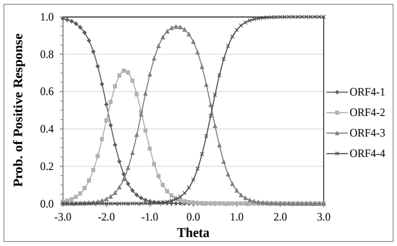
Fit Plot, Item 1 Option 4

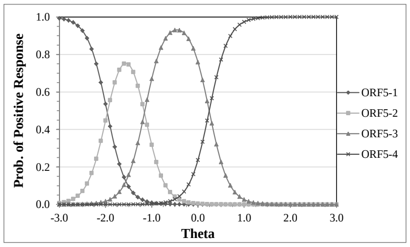
Fit Plot, Item 2 Option 4

Fit Plot, Item 3 Option 4

Figure A2. Item ORF and fit plots for scale H items 1 to 3

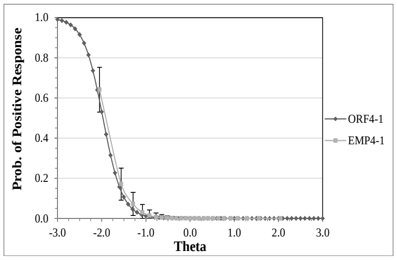
Figure A2 (Cont')



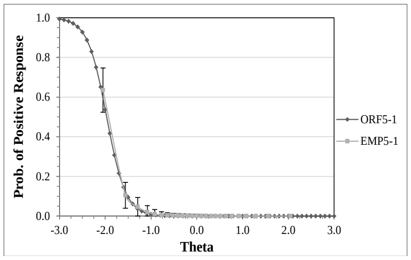


ORF plot, item 4

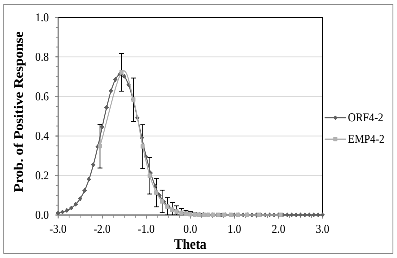
ORF plot, item 5

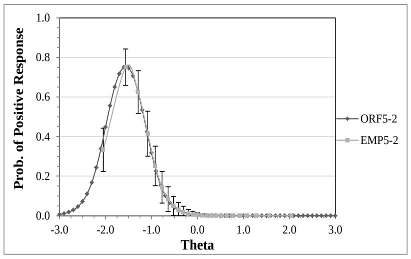


Fit plot, Item 4 Option 1



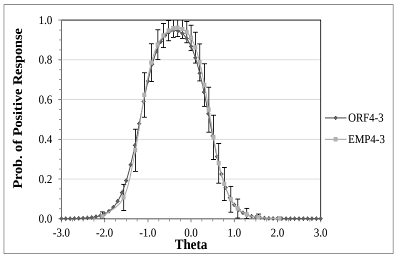
Fit plot, Item 5 Option 1

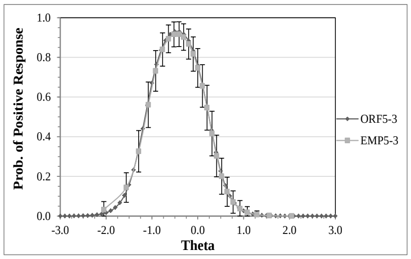




Fit plot, Item 4 Option 2

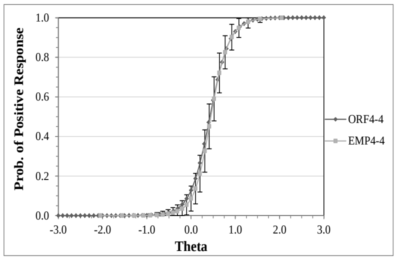
Fit plot, Item 5 Option 2



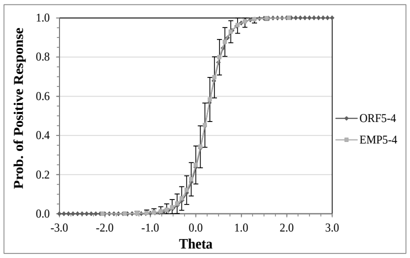


Fit plot, Item 4 Option 3

Fit plot, Item 4 Option 3

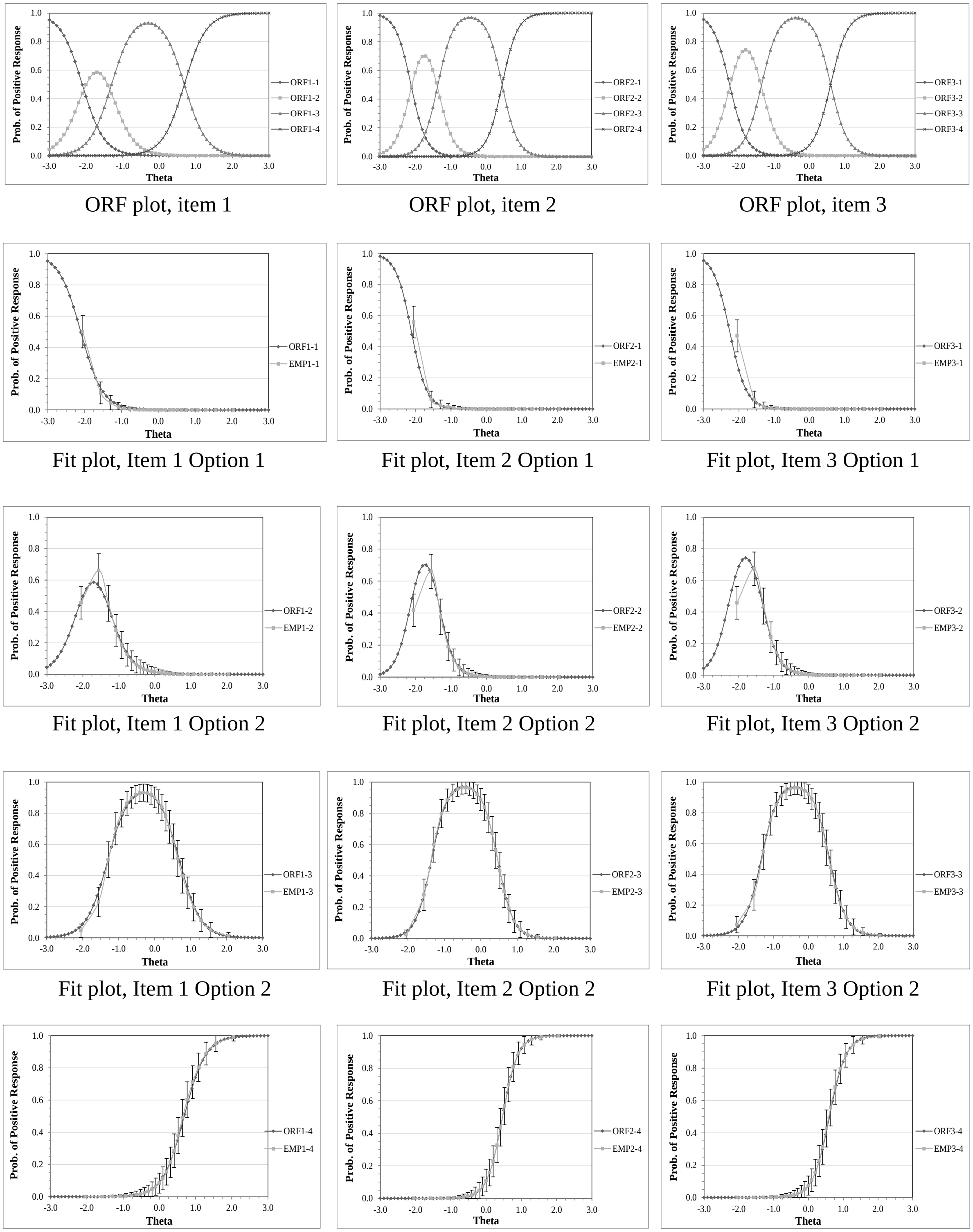


Fit plot, Item 4 Option 4



Fit plot, Item 5 Option 4

Figure A2. Item ORF and fit plots for scale H items 4 and 5



ORF plot, item 1 ORF plot, item 2 ORF plot, item 3

Fit plot, Item 1 Option 1 Fit plot, Item 2 Option 1 Fit plot, Item 3 Option 1

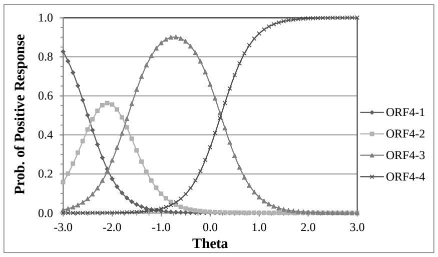
Fit plot, Item 1 Option 2 Fit plot, Item 2 Option 2 Fit plot, Item 3 Option 2

Fit plot, Item 1 Option 2 Fit plot, Item 2 Option 2 Fit plot, Item 3 Option 2

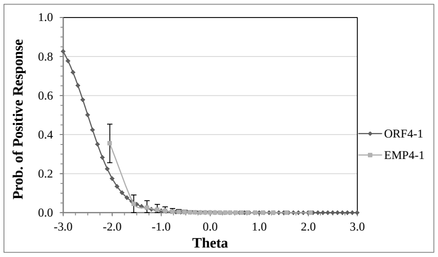
Fit plot, Item 1 Option 4 Fit plot, Item 2 Option 4 Fit plot, Item 2 Option 4

Figure A3. Item ORF and fit plots for scale J items 1 and 3

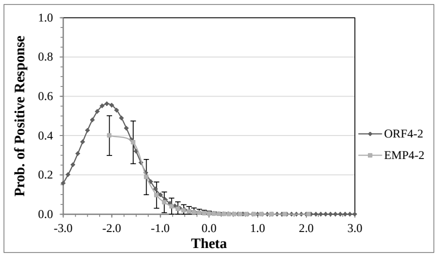
Figure A3 (Cont')



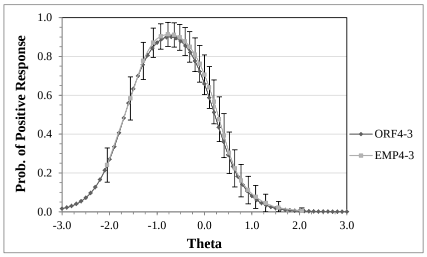
ORF plot, item 4



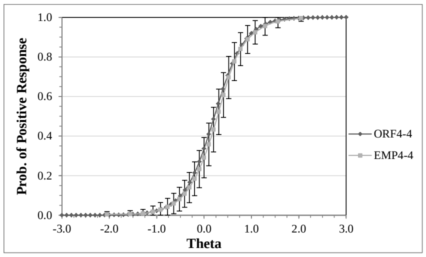
Fit plot, Item 4 Option 1



Fit plot, Item 4 Option 2



Fit plot, Item 5 Option 3



Fit plot, Item 5 Option 4

Figure A3. Item ORF's and Fit Plots for scale J item 4

Dimensionality

The scaling procedures for the Likert-type items in the student data of the CivEd study are described in the IEA Technical Report (2004). Regarding dimensionality, a confirmatory factor analysis (CFA) using structural equation modeling (SEM) was conducted and item fit statistics and scale reliabilities were computed. CFA analyses confirmed the expected item structure dimensions.

Items measuring subscale G attitudes toward desired rights or opportunities for women (*n*=2104) showed a satisfactory unidimentional factor structure (α = .81). The unidimensional factor structure of subscale H items measuring positive attitudes toward immigrants (*n*=2125) was .85 and the factor structure of subscale J measuring the support for women’s rights (*n*=2164) was.78.

APPENDIX B

MULTILOG SYNTAX FILES FOR GRM ITEM PARAMETER CALIBRATION CIVICS EDUCATION STUDY (1999) SUBSCALES G, H, J

Appendix B1. Syntax File for the Civics Education Study (1999) Subscale G

```
MULTILOG syntax generated by ...
>PROBLEM RAMDOM,
        INDIVIDUAL,
        DATA='C:\01-CivEd\DATA\STUDENT_G4.DAT',
        NItems = 6,
        NExaminees = 2104,
        NGROUP=1,
        NCHARS=7;
>TEST ALL, GRADED NCATS=(4(0)6);
>EST NC=50 IT=250;
>SAVE;
>END;
4
1234
114414
223323
332232
441141
(7A1,1X,6A1)
```

Appendix B2. Syntax File for the Civics Education Study (1999) Subscale H

```
MULTILOG syntax generated by ...
>PROBLEM RAMDOM,
        INDIVIDUAL,
        DATA='C:\01-CivEd\DATA\STUDENT_H4.DAT',
        NItems = 5,
        NExaminees = 2125,
        NGROUP=1,
        NCHARS=7;
>TES ALL, GRADED NCATS=(4(0)5);
>SAVE;
>EST NC=50 IT=250;
>END;
4
1234
11111
22222
33333
44444
(7A1,1X,5A1)
```

Appendix B3. Syntax File for the Civics Education Study (1999) Subscale J

```
MULTILOG syntax generated by ...
>PROBLEM RAMDOM,
        INDIVIDUAL,
        DATA='C:\01-CivEd\DATA\STUDENT_J4.DAT',
        NItems = 4,
        NExaminees = 2164,
        NGROUP=1,
        NCHARS=7;
>TES ALL, GRADED NCATS=(4(0)4);
>SAVE;
>EST NC=50 IT=250;
>END;
4
1234
1111
2222
3333
4444
(7A1,1X,4A1)
```

APPENDIX C

ITEM PARAMETER RECOVERY: THE EFFECT OF SAMPLE SIZE, NUMBER OF ITEMS, NUMBER OF REPLICATIONS AND MISSING DATA METHODS ON THE RECOVERY OF THE GRADED RESPONSE MODEL ITEM PARAMETERS

Introduction

Considering that the quality of a research study depends on the quality of the data manipulated for such study, the selection of factors for the study of item parameter recovery should follow not only theoretical but also empirical grounds. While the literature on parameter recovery is not very extensive, sample size is the factor influencing the most the recovery of item parameters across these studies, for different IRT models. However, these studies conducted a limited number of replications, which detracts the accuracy and precision of findings.

Purpose of the Study

This item parameter recovery study was conducted with two purposes: 1) to investigate the effectiveness of the IRTGEN macro for generating item response data for the graded response model, and 2) to investigate the effect of missing data and missing data treatments on the recovery of item parameters in terms of accuracy (i.e., BIAS of the estimates) and precision (i.e., root mean square error or RMSE).

BIAS. The accuracy of each GRM item parameter was evaluated the BIAS (residual error of estimation),

$$BIAS(\Lambda_{jk}) = \frac{\sum_{k=1}^{r}(\hat{\Lambda}_{jk} - \Lambda_{jk})}{r},$$

where

Λ = item parameter of interest (e.g., item discrimination a_j or item threshold b_{jk})

Λ_{jk} = parameter value of item *j* for category *k*

$\hat{\Lambda}_{jk}$ = estimated item parameter value, $\hat{\Lambda}_{j}$, in category *k*

r = number of samples or replications

RMSE. The root mean squared error or RMSE is the square root of the average squared difference between estimated parameter values and the parameters used to generate the data (true parameters (Bolt, 2002; DeMars, 2002). That is, RMSE combines BIAS and the estimated parameter value with sampling error to provide the total error in the estimated parameter value. The RMSE estimates of both the discrimination parameter (*a*) and location parameter (*b*) of the GRM were calculated using the following formula:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{\mathrm{j}=1}^{m}(\hat{\Lambda}_{ij}-\Lambda_{\mathrm{ij}})^2}{n_{\mathrm{i}}m_{\mathrm{j}}}}$$

Where

Λ = a given item parameter, discrimination (a_i) or threshold ($b_{1\mathrm{j}}$, $b_{2\mathrm{j}}$, $b_{3\mathrm{j}}$, $b_{4\mathrm{j}}$),

Λ_{ij} = true parameter value of item *i* for sample *j*, if no data were missing,

$\hat{\Lambda}_{\mathrm{ij}}$ = estimated item parameter value, $\hat{\Lambda}_{\mathrm{i}}$, in sample *j*,

$n_{\mathrm{i}}m_{\mathrm{j}}$ = sample size and number of samples or replications (respectively)

Sample

Data was generated from the item parameters of scales G, H, and J of the Civics Education Study, administered to a standard population of 2811 students from 124 schools in the United States (9th grade students, the grade in which most 14-year olds were at the time of testing). The items used a Likert-type response format using a four-point scale scored 1) *strongly disagree*; 2) *disagree*; 3) *agree;* and 4) *strongly agree*. Observations with items coded 8 (unit nonresponse), 9 (item nonresponse) , and 0, ("don't know" option included in each item), were eliminated from the analysis. Tables C1 - C3 show the frequencies of subscales G, H, and J respective these options per subcale scale. Table C1 shows the option frequencies for the six items of subscale G.

Table C1

Frequency Distributions of Items' Category Options by Subscale J

Item J1				
Options	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**213**	**7.58**	213	7.58
1	166	5.91	379	13.48
2	253	9.00	632	22.48
3	1361	48.42	1993	70.9
4	688	24.48	2681	95.38
8	**37**	**1.32**	2718	96.69
9	**93**	**3.31**	2811	100
Item J2				
Options				
0	**190**	**6.76**	190	6.76
1	91	3.24	281	10.00
2	220	7.83	501	17.82
3	1360	48.38	1861	66.20
4	812	28.89	2673	95.09
8	**37**	**1.32**	2710	96.41
9	**101**	**3.59**	2811	100

			Table C1 cont’	
Options	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Item J3				
Options				
0	**228**	**8.11**	228	8.11
1	87	3.09	315	11.21
2	234	8.32	549	19.53
3	1413	50.27	1962	69.8
4	713	25.36	2675	95.16
8	**37**	**1.32**	2712	96.48
9	**99**	**3.52**	2811	100
Item J5				
Options				
0	**249**	**8.86**	249	8.86
1	88	3.13	337	11.99
2	164	5.83	501	17.82
3	1158	41.2	1659	59.02
4	1012	36.00	2671	95.02
8	**37**	**1.32**	2708	96.34
9	**103**	**3.66**	2811	100

Note. Item options 0 = don’t know; 1 = completely disagree; 2 = disagree; 3 = agree; 4 = completely agree; 8 = not administered; 9 = missing. Percentages deleted across items and items’ options are bolded and correspond to those for option 0 (don’t know), and for options coded 8 (not administered) and 9 (missing).

Table C2

Frequency Distributions of Items’ Category Options by Subscale H

Item H1				
Options	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**190**	**6.76**	190	6.76
1	200	7.11	390	13.87
2	322	11.45	712	25.33
3	1345	47.85	2057	73.18
4	655	23.3	2712	96.48
8	**37**	**1.32**	2749	97.79
9	**62**	**2.21**	2811	100
Item H2				
Options				
0	**123**	**4.38**	123	4.38

1	86	3.06	209	7.44
			Table C2 cont'	
2	152	5.41	361	12.84
3	1181	42.01	1542	54.86
4	1169	41.59	2711	96.44
8	**37**	**1.32**	2748	97.76
9	**63**	**2.24**	2811	100
Item H3				
Options				
0	**281**	**10.00**	281	
1	119	4.23	400	14.23
2	304	10.81	704	25.04
3	1255	44.65	1959	69.69
4	743	26.43	2702	96.12
8	37	1.32	2739	97.44
9	72	2.56	2811	100
Item H4				
Options				
0	**214**	**7.61**	214	7.61
1	116	4.13	330	11.74
2	277	9.85	607	21.59
3	1247	44.36	1854	65.96
4	843	29.99	2697	95.94
8	**37**	**1.32**	2734	97.26
9	**77**	**2.74**	2811	100
Item H5				
Options	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**206**	**7.33**	206	7.33
1	118	4.20	324	11.53
2	263	9.36	587	20.88
3	1119	39.81	1706	60.69
4	978	34.79	2684	95.48
8	**37**	**1.32**	2721	96.80
9	**90**	**3.20**	2811	100

Note. Item options 0 = don't know; 1 = completely disagree; 2 = disagree; 3 = agree; 4 = completely agree; 8 = not administered; 9 = missing. Percentages deleted across items and items' options are bolded and correspond to those for option 0 (don't know), and for options coded 8 (not administered) and 9 (missing).

Table C3

Frequency Distributions of Item's Category Options by Subscale G

Item G1				
Options	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**134**	**4.77**	134	4.77
1	118	4.2	252	8.96
2	157	5.59	409	14.55
3	1066	37.92	1475	52.47
4	1266	45.04	2741	97.51
8	**37**	**1.32**	2778	98.83
9	**33**	**1.17**	2811	100
Item G4				
Options	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**131**	**4.66**	131	4.66
1	91	3.24	222	7.90
2	172	6.12	394	14.02
3	721	25.65	1115	39.67
4	1614	57.42	2729	97.08
8	**37**	**1.32**	2766	98.40
9	**45**	**1.60**	2811	100
Item G6				
Options	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**187**	**6.65**	187	6.65
1	1587	56.46	1774	63.11
2	652	23.19	2426	86.30
3	169	6.01	2595	92.32
4	129	4.59	2724	96.91
8	**37**	**1.32**	2761	98.22
9	**50**	**1.78**	2811	100
Item G9				
Options	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**263**	**9.36**	263	9.36
1	1113	39.59	1376	48.95
2	749	26.65	2125	75.60
3	430	15.3	2555	90.89
4	170	6.05	2725	96.94
8	**37**	**1.32**	2762	98.26
9	**49**	**1.74**	2811	100

			Table C3 cont’	
Item G11				
Options	Frequenc y	Percent	Cumulative Frequency	Cumulative Percent
0	**162**	**5.76**	162	5.76
1	86	3.06	248	8.82
2	200	7.11	448	15.94
3	812	28.89	1260	44.82
4	1451	51.62	2711	96.44
8	**37**	**1.32**	2748	97.76
9	**63**	**2.24**	2811	100
Item G13				
Options	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	**231**	**8.22**	231	8.22
1	1304	46.39	1535	54.61
2	715	25.44	2250	80.04
3	290	10.32	2540	90.36
4	174	6.19	2714	96.55
8	**37**	**1.32**	2751	97.87
9	**60**	**2.13**	2811	100

Note. Item options 0 = don’t know; 1 = completely disagree; 2 = disagree; 3 = agree; 4 = completely agree; 8 = not administered; 9 = missing. Percentages deleted across items and items’ options are bolded and correspond to those for option 0 (don’t know), and for options coded 8 (not administered) and 9 (missing).

As can be observed in Tables C1-C3, the subscales presented different levels of missingness across items and across items’ options. The largest amounts of data deleted for the final samples came from deleting the “don’t know” option across all subscales. The percentage of observations deleted due to the “don’t know” was in the range of 4.66 – 9.36 for subscale G. From each subscale 37 observations (1.32) were deleted due the test not being administered and the range of missing data was between of observations due to participants not being administered the test was in the range of 1.17 to 3.20. Final sample size after deletion of options coded 0, 8, and 9 were N = 2164, N = 2125, and N = 2104 (subscales J, H, and G respectively).

Bias Results

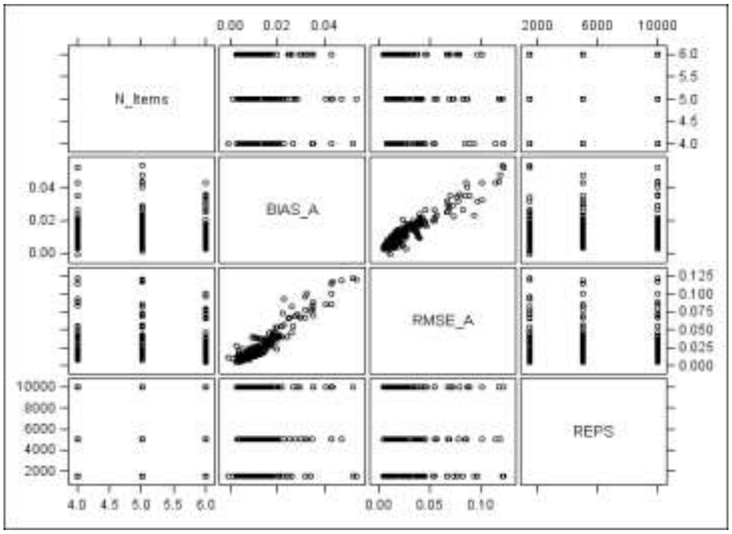


Figure C1. Overall distributions for discrimination parameter *a*: scale length, Bias, RMSE by number of replications.

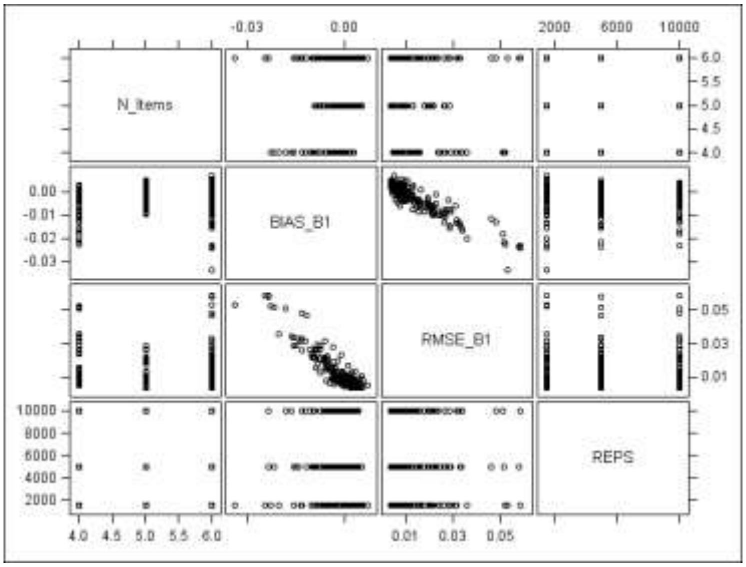


Figure C2. Overall distributions for location parameter b_1: scale length, Bias, RMSE by number of replications

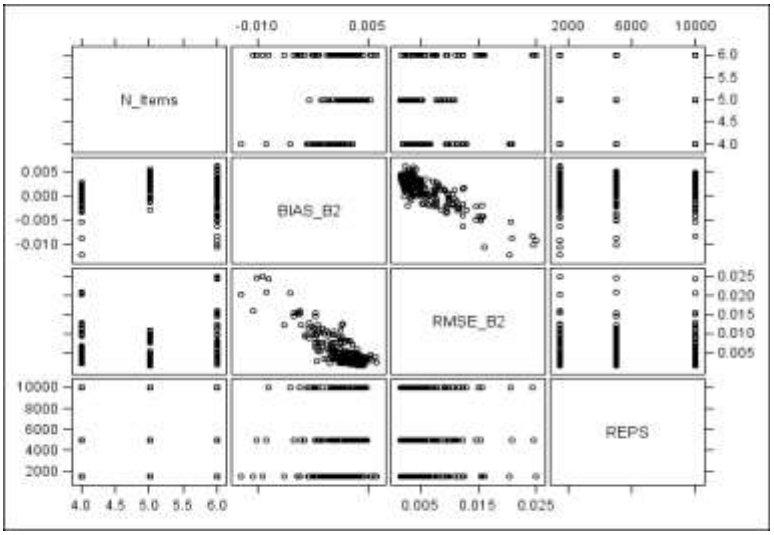


Figure C3. Overall distributions for location parameter b_2: scale length, Bias, RMSE by number of replications

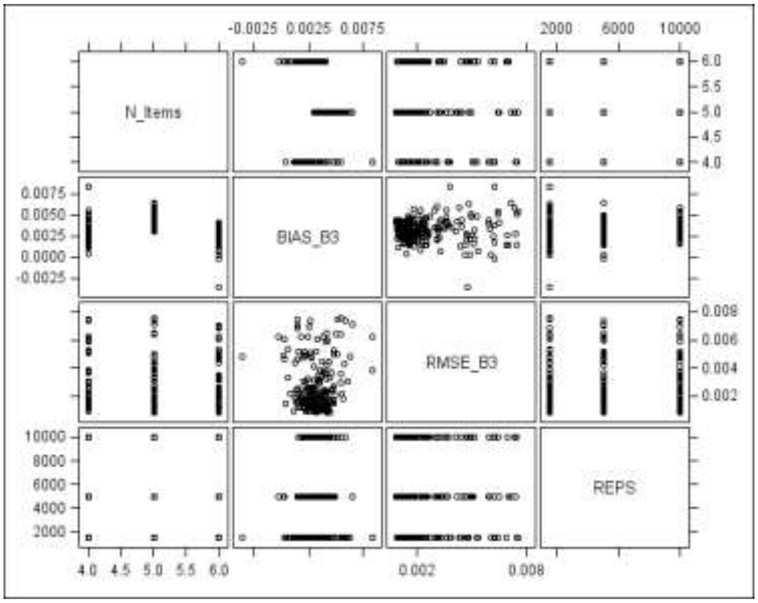


Figure C4. Overall distributions for location parameter b_3: scale length, Bias, RMSE by number of replications

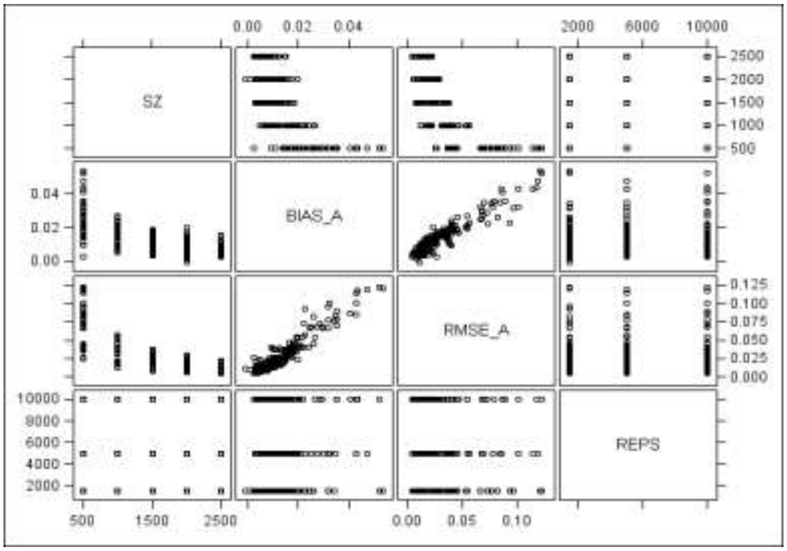


Figure C5. Overall distributions for discrimination parameter a_1: scale length, Bias, RMSE by number of replications

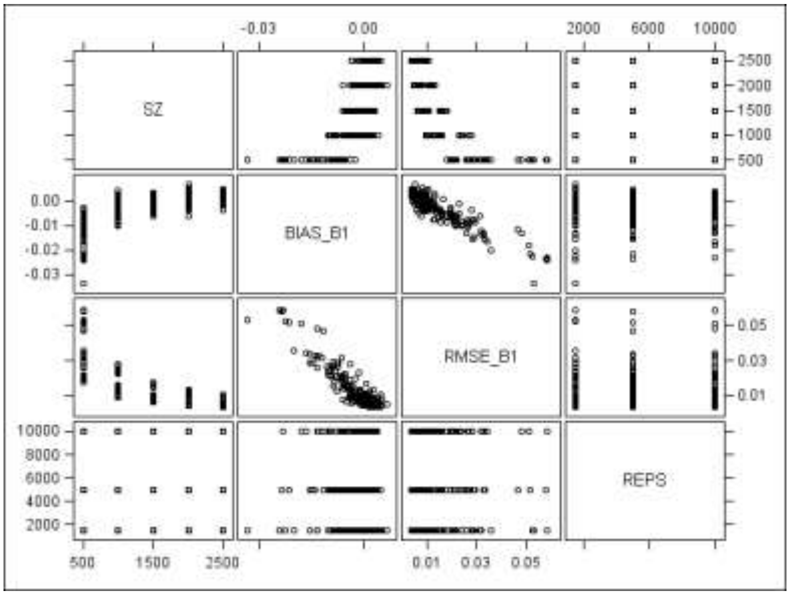


Figure C6. Scale H distributions for location parameter b_1: scale length, Bias, RMSE by number of replications

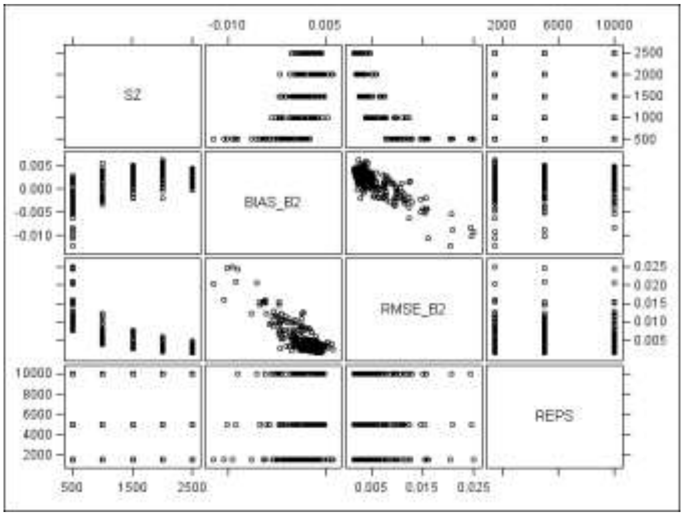


Figure C7. Scale H distributions for location parameter b_2: scale length, Bias, RMSE by number of replications

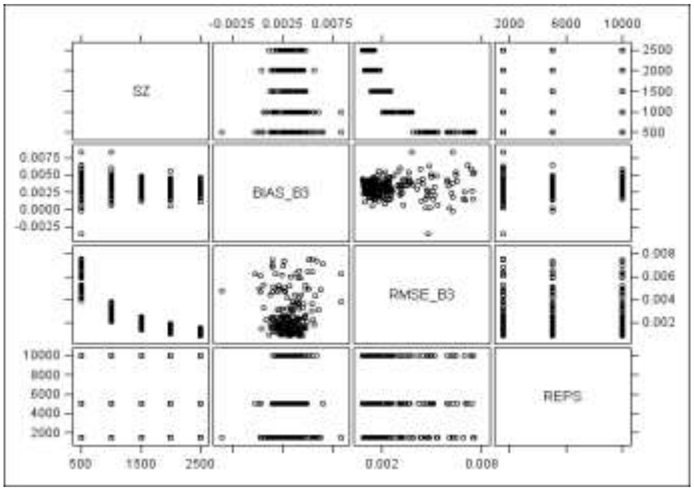


Figure C8. Scale H distributions for location parameter b_3: scale length, Bias, RMSE by number of replications

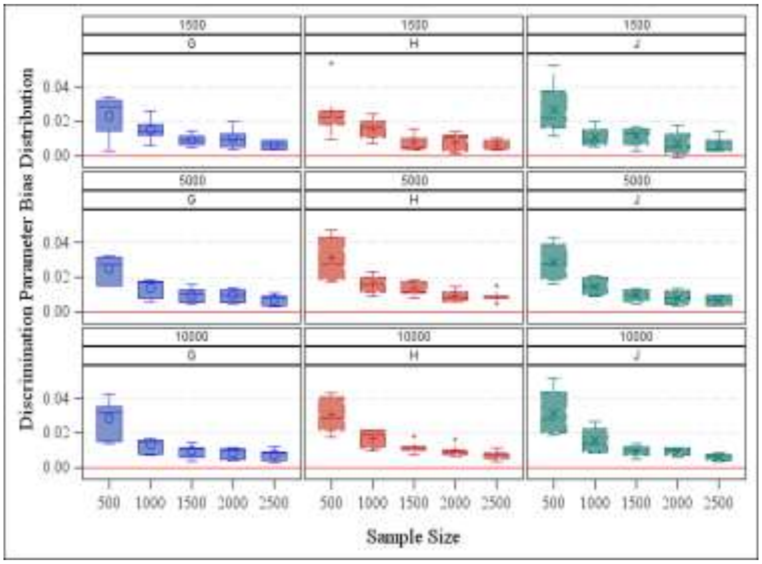


Figure C9. Box plots BIAS distributions discrimination parameter (*a*) by sample size, number of replications and scale

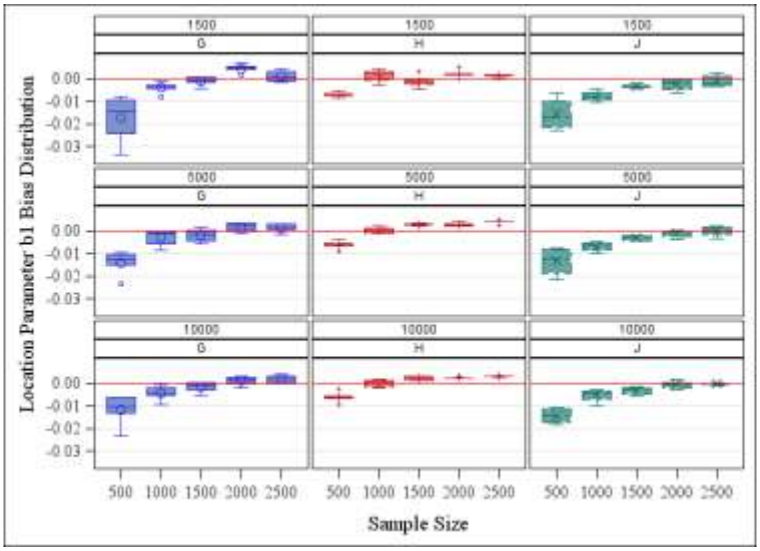


Figure C10. Box plots BIAS distributions for location parameter (b_1) sample size, number of replications, and scale.

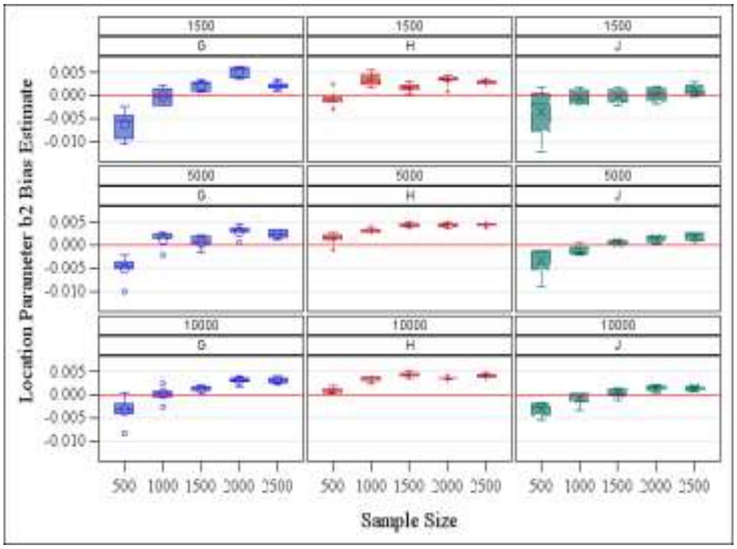


Figure C11. Box plots BIAS distributions for location parameter (b_2) sample size, number of replications, and scale

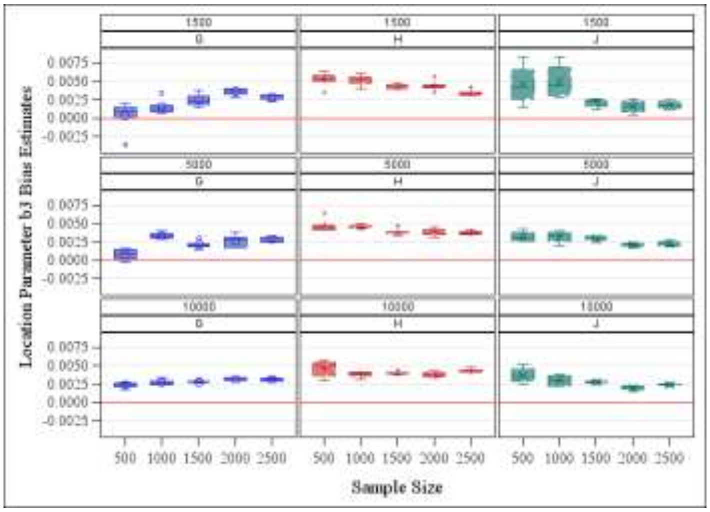


Figure C12. Box plots BIAS distributions for Location parameter (b_3) by sample size, number or replications, and scale.

.

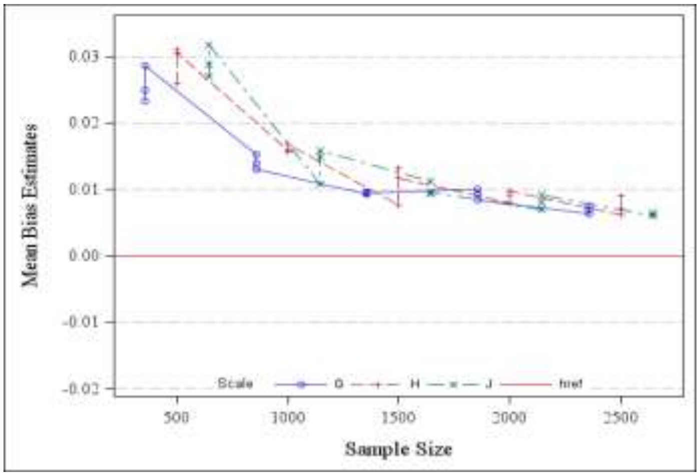


Figure C13. Mean bias estimates discrimination parameter (*a*) mean bias estimates by sample size, number of replications, and scale.

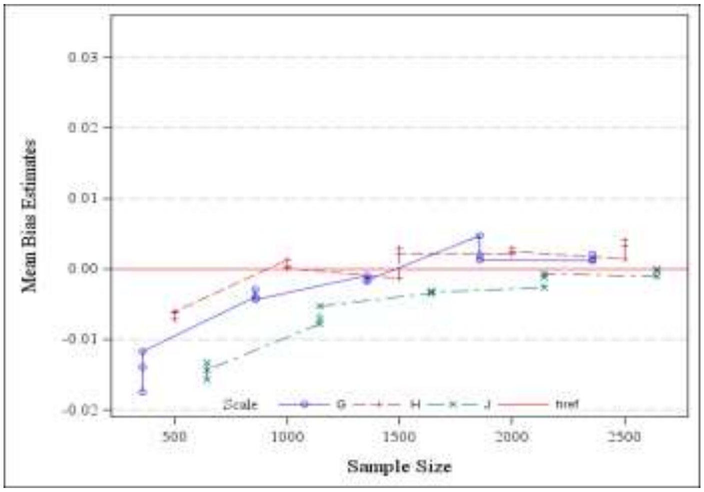


Figure C14. Mean bias estimates location parameter b_1 mean bias estimates by sample size, number of replications, and scale.

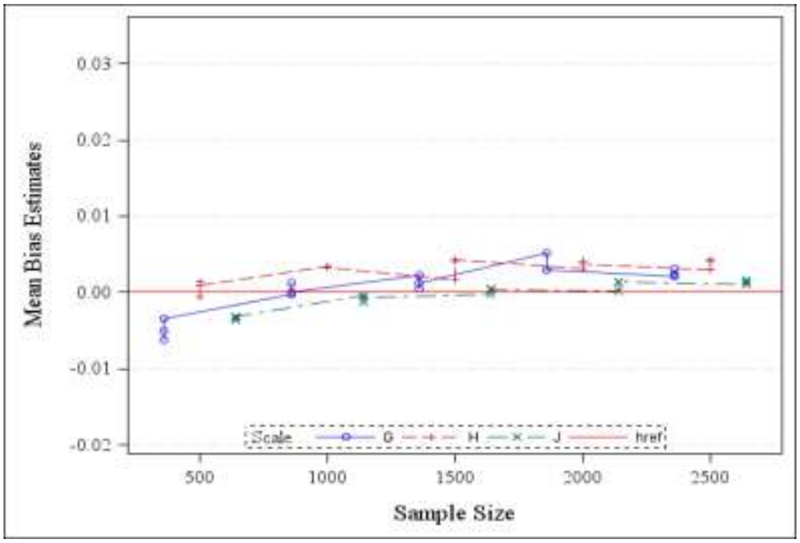


Figure C15. Mean bias estimates location parameter b_2 mean bias distributions by sample size, number of replications, and scale.

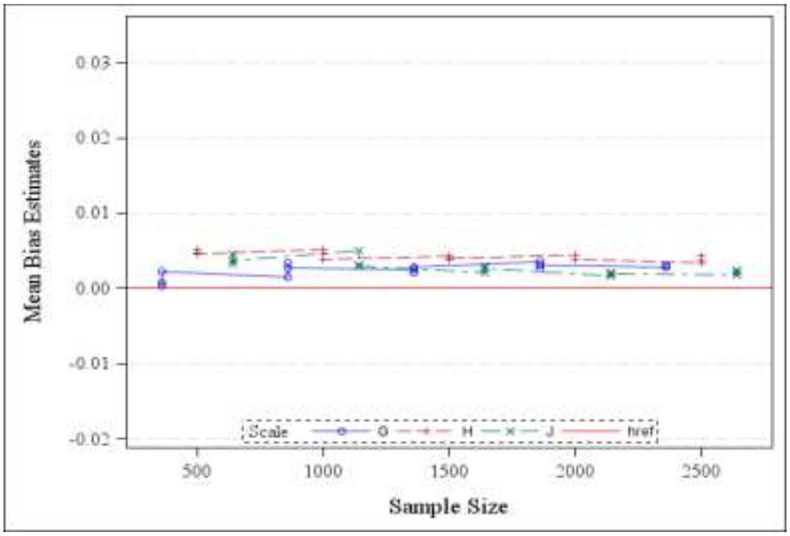


Figure C16. Mean bias location parameter b_3 mean bias distributions of by sample size, number of replications, and scale.

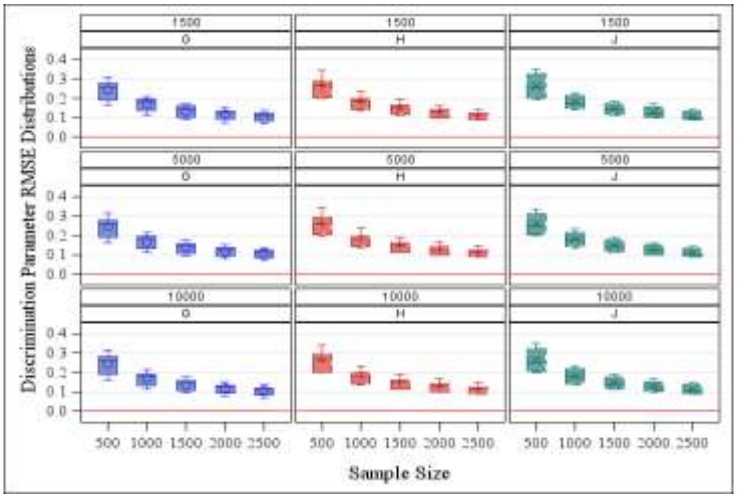


Figure C17. Box plots RMSE discrimination parameter (*a*) by sample size, number of replications, and scale.

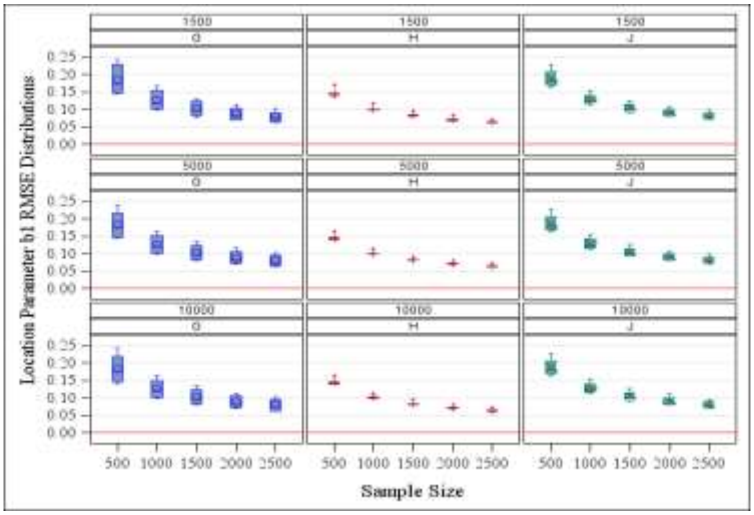


Figure C18. Box plots RMSE location parameter (b_1) by sample size, number of replications, and scale.

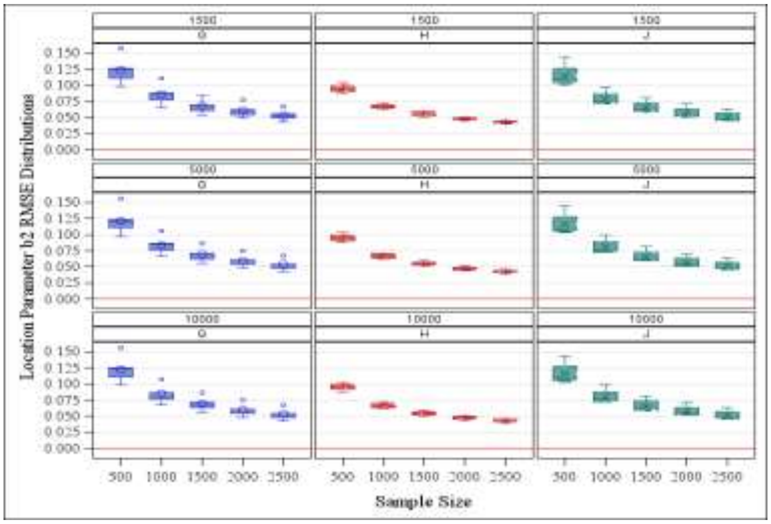


Figure C19. Box plots RMSE location parameter (b_2) by sample size, number of replications, and scale.

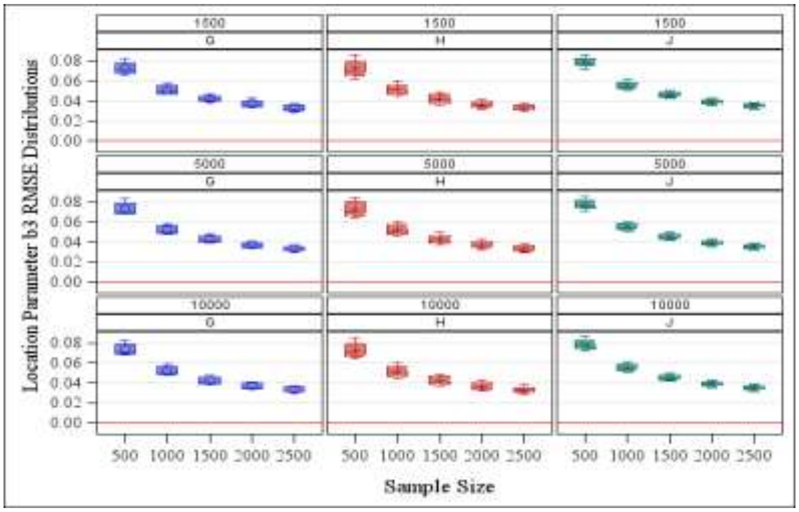


Figure C20. Box plots RMSE location parameter (b_3) RMSE by sample size, number of replications, and scale.

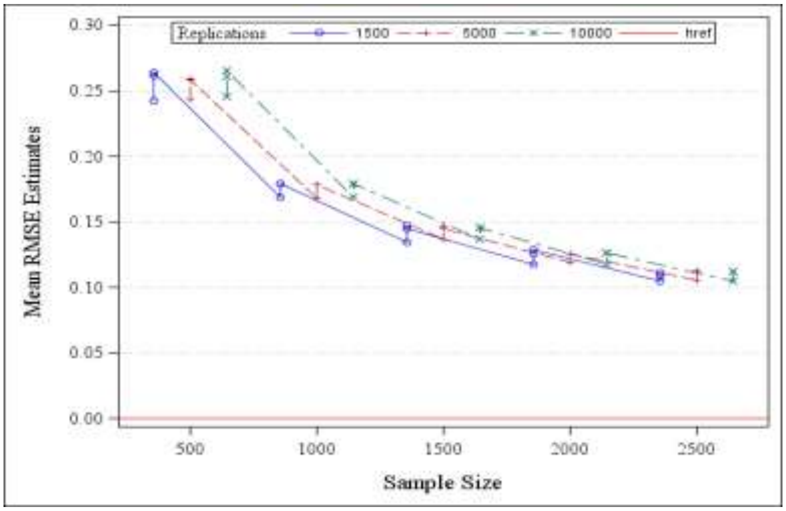


Figure C21. Mean RMSE discrimination parameter (*a*) by sample size, number of replications, and scale.

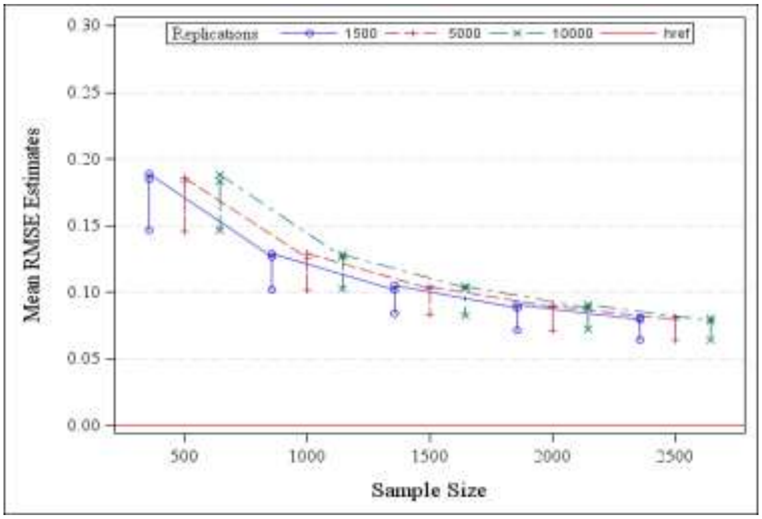


Figure C22. Mean RMSE discrimination parameter (b_1) by sample size, number of replications, and scale.

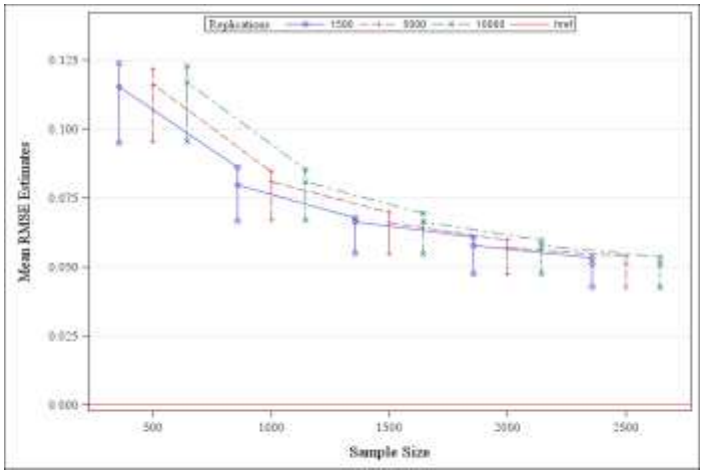


Figure C23. Mean RMSE discrimination parameter (b_2) by sample size, number of replications, and scale.

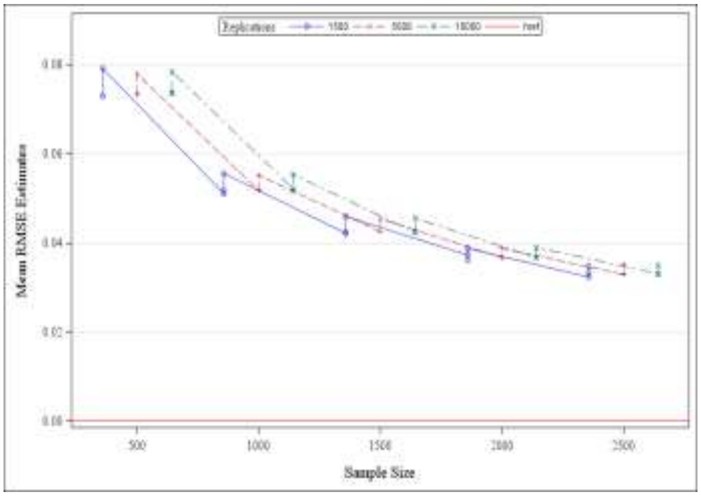


Figure C24. Mean RMSE discrimination parameter (b_3) by sample size, number of replications, and scale.

Table C4

Mean Bias Estimates by Sample Size across Replications

		Scale G – 6 items				Scale H– 5 items				Scale J – 4 items			
N	Rep	*a*	b_1	b_2	b_3	*a*	b_1	b_2	b_3	*a*	b_1	b_2	b_3
500	1500	.023	-.017	-.006	.000	.026	-.007	-.001	.005	.027	-.016	-.004	.005
	5000	.025	-.014	-.005	.001	.031	-.006	.001	.005	.029	-.013	-.003	.003
	10000	.029	-.012	-.003	.002	.031	-.006	.001	.005	.032	-.014	-.003	.004
1000	1500	.015	-.004	.000	.002	.016	.001	.003	.005	.011	-.008	.000	.005
	5000	.014	-.003	.001	.003	.016	.000	.003	.005	.015	-.007	-.001	.003
	10000	.013	-.004	.000	.003	.017	.000	.003	.004	.016	-.005	-.001	.003
1500	1500	.009	-.001	.002	.002	.008	-.001	.002	.004	.011	-.003	.000	.002
	5000	.010	-.002	.001	.002	.013	.003	.004	.004	.009	-.003	.000	.003
	10000	.009	-.002	.001	.003	.012	.002	.004	.004	.010	-.003	.000	.003
2000	1500	.010	.005	.005	.004	.008	.002	.003	.004	.007	-.003	.000	.002
	5000	.009	.002	.003	.003	.009	.003	.004	.004	.008	-.001	.001	.002
	10000	.008	.001	.003	.003	.010	.002	.004	.004	.009	-.001	.001	.002
2500	1500	.006	.001	.002	.003	.006	.001	.003	.003	.006	-.001	.001	.002
	5000	.007	.001	.002	.003	.009	.004	.004	.004	.006	.000	.002	.002
	10000	.007	.002	.003	.003	.007	.003	.004	.004	.006	.000	.001	.002

Table C5

Mean RMSE Estimates by Sample Size across Replications

		Scale G – 6 items				Scale H– 5 items				Scale J – 4 items			
N	Reps	*a*	b_1	b2	b_3	*a*	b_1	b_2	b_3	*a*	b_1	b_2	b_3
500	1500	.243	.185	.124	.073	.262	.147	.095	.073	.264	.189	.115	.079
	5000	.244	.183	.122	.073	.259	.145	.096	.073	.258	.186	.116	.078
	1000	.246	.183	.123	.074	.261	.147	.096	.073	.265	.188	.117	.078
1000	1500	.169	.127	.086	.051	.179	.101	.067	.052	.179	.129	.080	.056
	5000	.168	.126	.085	.052	.178	.102	.067	.052	.179	.129	.081	.055
	10000	.169	.126	.085	.052	.179	.102	.067	.052	.180	.128	.081	.055
1500	1500	.134	.102	.068	.042	.148	.084	.055	.042	.145	.105	.066	.046
	5000	.137	.103	.070	.043	.146	.083	.055	.043	.145	.104	.066	.045
	10000	.137	.103	.070	.042	.146	.083	.055	.043	.146	.104	.066	.046
2000	1500	.118	.088	.061	.037	.126	.072	.047	.036	.129	.091	.058	.039
	5000	.119	.089	.060	.037	.126	.072	.047	.037	.126	.090	.057	.039
	10000	.118	.088	.060	.037	.126	.072	.048	.037	.127	.091	.058	.039
2500	1500	.105	.079	.053	.032	.111	.064	.043	.033	.109	.081	.051	.035
	5000	.106	.080	.054	.033	.113	.064	.043	.033	.112	.081	.051	.035
	10000	.105	.079	.054	.033	.112	.064	.042	.033	.112	.080	.051	.035

APPENDIX D

SUMMARY TABLES: EFFECT SIZE ESTIMATES FOR TYPE I ERROR AND POWER

Table D1

Main and First-Order Interaction effects on Familywise Error Rate Estimates by Missing Data Method (α = .05 and *α = .05)*

	Complete Data		FIML		Multiple Imputation		Person Mean Substitution		Single Regression Substitution		Relative Mean Substitution		Listwise	
Source	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
Ability Distribution	.00	.01	.01	.00	.00	.00	.01	.01	.00	.00	.00	.00	.01	.00
Items	.01	.03	.02	.02	.06	.04	.02	.05	.14	.16	.13	.14	**.07**	.04
Sample Size	.00	.02	.01	.00	.00	.00	.01	.01	.00	.00	.00	.00	.04	.06
Missing Observations	.02	.01	.01	.01	.55	.59	.03	.03	.21	.23	.12	.12	.03	.07
Missing Items	.00	.00	.01	.00	.15	.18	.00	.02	.15	.18	.08	.09	.01	.00
Distribution / Items	.01	.04	.00	.01	.00	.00	.01	**.05**	.00	.00	.01	.00	.01	.00
Distribution / Sample Size	.00	.01	.00	.01	.00	.00	.01	.02	.00	.00	.00	.00	.02	.03
Distribution / Missing Observations	.02	.00	**.05**	.02	.00	.00	.01	.00	.00	.00	.00	.00	.02	.00
Distribution / Missing Items	.00	.01	.01	.00	.00	.00	.03	.00	.00	.00	.00	.00	.01	.01
Items / Sample Size	.02	.02	.01	.02	.00	.00	.02	.01	.00	.00	.01	.00	.03	.04
Items / Missing Observations	.02	.01	.02	.01	.03	.02	.02	.01	**.11**	**.09**	**.07**	**.11**	.04	.01
Items / Missing items	.00	.01	.00	.01	.00	.00	.00	.00	**.10**	**.10**	**.15**	**.16**	.01	.01
Sample Size / Missing Observatio ns	.04	.04	.03	.04	.00	.01	.02	.03	.00	.00	.00	.01	.02	**.07**
Sample Size / Missing Items	.01	.02	.01	.00	.00	.00	.00	.00	.00	.00	.01	.00	.03	.02
Missing Observations/ Missing Items	.01	.00	.01	.00	**.12**	**.11**	.02	.02	**.13**	**.13**	**.07**	**.08**	.01	.00

Table D2

Main and First-Order Interaction effects on Statistical Power Estimates by Missing Data Method

(*α = .05* and *α = .05)*

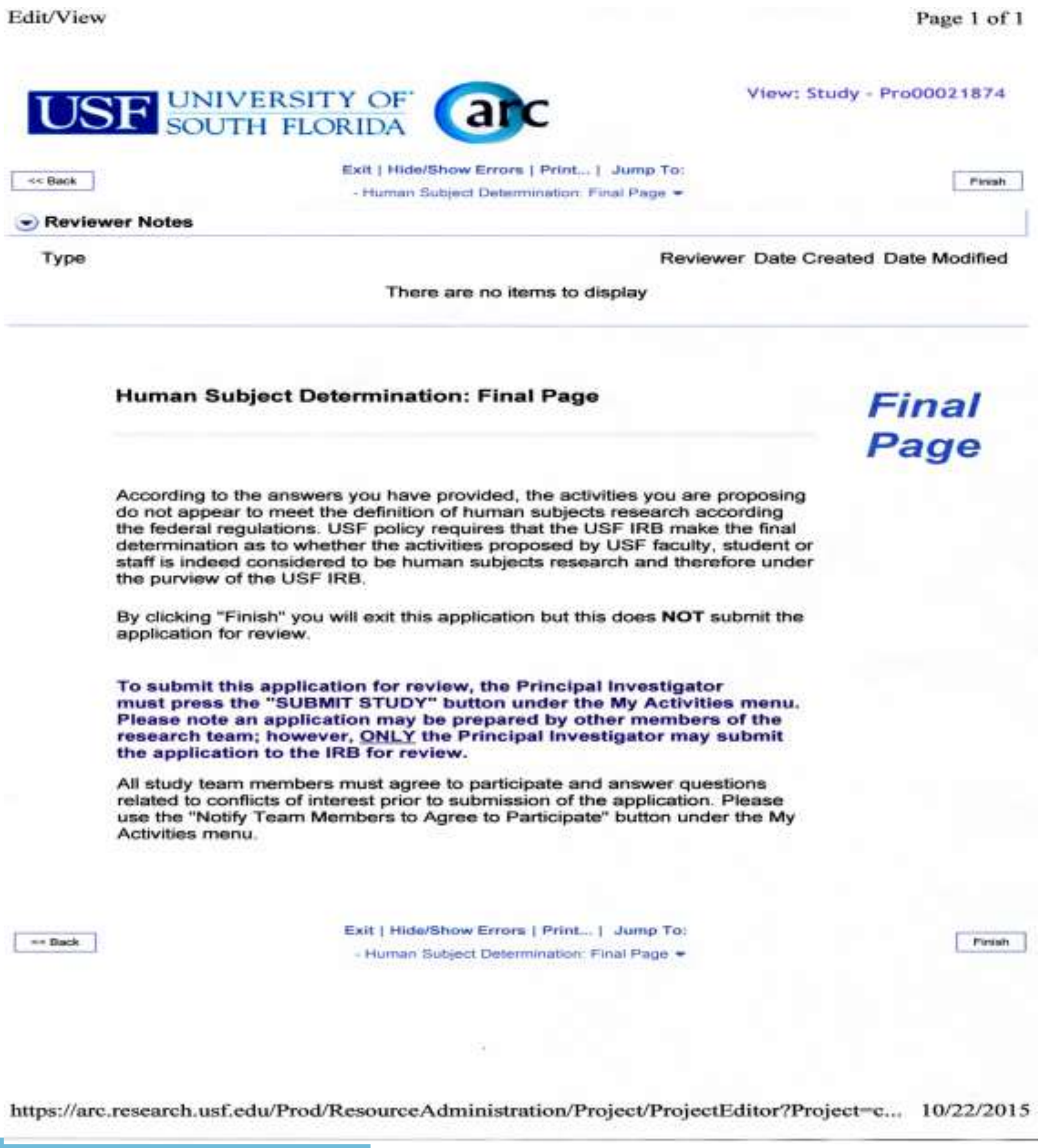
	Complete Data		FIML		Multiple Imputation		Person Mean Substitution		Single Regression Substitution		Relative Mean Substitution		Listwise	
Source	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
Ability Distribution	.01	00	.01	.00	--	.00	.01	.01	--	.00	.00	.00	.00	.00
Items	**.17**	.09	**.18**	**.10**	--	**.10**	**.23**	.10	--	**.10**	**.09**	**.09**	**.15**	**.08**
Sample Size	**.70**	32	**.71**	**.34**	--	**.34**	**.72**	.33	--	**.33**	**.72**	**.33**	**.67**	**.33**
Missing Observations	.00	.00	.02	.00	--	.00	.03	.00	--	.00	.01	.00	**.12**	.03
Missing Items	.00	.00	.01	.00	--	.00	.01	.00	--	.00	.00	.00	.00	.00
Ability Distribution / Items	.00	.00	.00	.00	--	.00	.00	.00	--	.00	.00	.00	.00	.00
Ability Distribution / Sample Size	.00	.00	.00	.00	--	.00	.00	.00	--	.00	.00	.00	.00	.00
Ability Distribution / Missing Observations	.02	.00	.00	.00	--	.00	.00	.00	--	.00	.00	.00	.00	.00
Ability Distribution / Missing Items	.00	.00	.01	.00	--	.00	.00	.00	--	.00	.00	.00	.00	.00
Items / Sample Size	**.08**	.12	**.05**	**.10**	--	**.11**	**.05**	**.10**	--	**.10**	**.05**	**.11**	.02	**.06**
Items / Missing Observations	.01	.00	.00	.00	--	.00	.01	.00	--	.00	.00	.00	.00	.00
Items / Missing items	.00	.00	.00	.00	--	.00	.00	.00	--	.00	.01	.00	.00	.00
Sample Size / Missing Observations	.00	.00	.00	.00	--	.00	.00	.00	--	.00	.01	.00	.01	.00
Sample Size / Missing Items	.00	.00	.00	.00	--	.00	.00	.00	--	.00	.00	.00	.00	.00
Missing Observations/ Missing Items	.00	.00	.00	.00	--	.00	.00	.00	--	.00	.00	.00	.00	.00

Note. If present, significant effect sizes ($\eta^2 \geq .05$) for first-order were reported. Main effects were reported for those methods with no significant interactions.

APPENDIX E

IRB HUMAN SUBJECTS DETERMINATION AND IRB RESEARCHER TRAINING RECORDS

E1. IRB Human subjects determination



Edit/View

Page 1 of 1

USF UNIVERSITY OF SOUTH FLORIDA
arc
View: Study - Pro00021874
<< Back
Exit | Hide/Show Errors | Print... | Jump To:
- Human Subject Determination: Final Page
Finish
Reviewer Notes
Type
Reviewer
Date Created
Date Modified

There are no items to display

Human Subject Determination: Final Page

Final Page

According to the answers you have provided, the activities you are proposing do not appear to meet the definition of human subjects research according the federal regulations. USF policy requires that the USF IRB make the final determination as to whether the activities proposed by USF faculty, student or staff is indeed considered to be human subjects research and therefore under the purview of the USF IRB.

By clicking "Finish" you will exit this application but this does **NOT** submit the application for review.

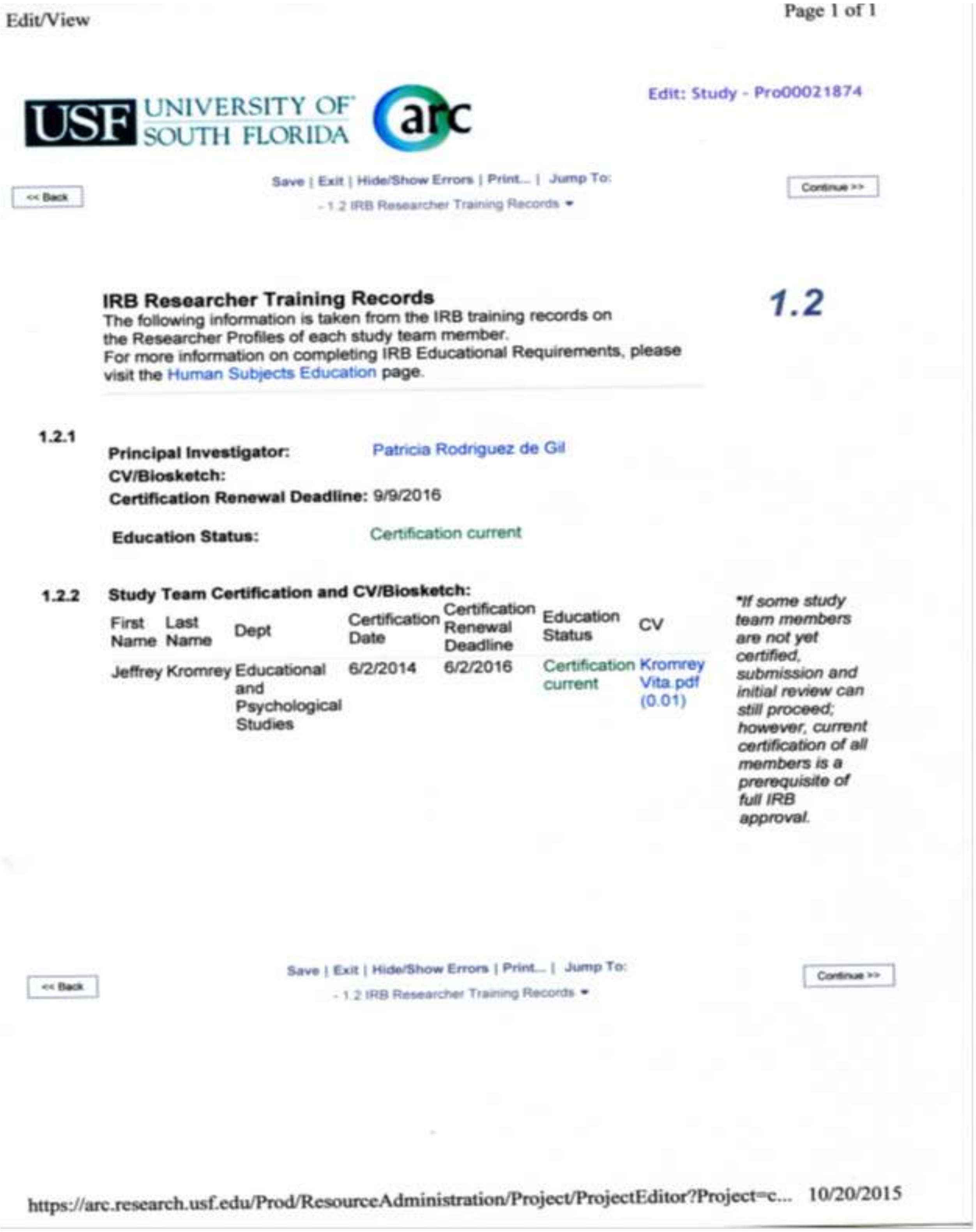
To submit this application for review, the Principal Investigator must press the "SUBMIT STUDY" button under the My Activities menu. Please note an application may be prepared by other members of the research team; however, ONLY the Principal Investigator may submit the application to the IRB for review.

All study team members must agree to participate and answer questions related to conflicts of interest prior to submission of the application. Please use the "Notify Team Members to Agree to Participate" button under the My Activities menu.

<< Back
Exit | Hide/Show Errors | Print... | Jump To:
- Human Subject Determination: Final Page
Finish

https://arc.research.usf.edu/Prod/ResourceAdministration/Project/ProjectEditor?Project=c... 10/22/2015

E2 IRB Researcher Training Records



Edit/View

Page 1 of 1

USF UNIVERSITY OF SOUTH FLORIDA
arc

Edit: Study - Pro00021874

<< Back
Save | Exit | Hide/Show Errors | Print... | Jump To:
- 1.2 IRB Researcher Training Records ▾
Continue >>

IRB Researcher Training Records

1.2

The following information is taken from the IRB training records on the Researcher Profiles of each study team member.
For more information on completing IRB Educational Requirements, please visit the Human Subjects Education page.

1.2.1

Principal Investigator: Patricia Rodriguez de Gil
CV/Biosketch:
Certification Renewal Deadline: 9/9/2016

Education Status: Certification current

1.2.2 **Study Team Certification and CV/Biosketch:**

First Name	Last Name	Dept	Certification Date	Certification Renewal Deadline	Education Status	CV
Jeffrey	Kromrey	Educational and Psychological Studies	6/2/2014	6/2/2016	Certification current	Kromrey Vita.pdf (0.01)

**If some study team members are not yet certified, submission and initial review can still proceed; however, current certification of all members is a prerequisite of full IRB approval.*

<< Back
Save | Exit | Hide/Show Errors | Print... | Jump To:
- 1.2 IRB Researcher Training Records ▾
Continue >>

https://arc.research.usf.edu/Prod/ResourceAdministration/Project/ProjectEditor?Project=c... 10/20/2015