Controlling Type I Errors in Moderated Multiple Regression: An Application of Item Response Theory for Applied Psychological Research

A dissertation presented to

the faculty of

the College of Arts and Sciences of Ohio University

In partial fulfillment

of the requirements for the degree

Doctor of Philosophy

Brendan J. Morse

August 2009

This dissertation titled

Controlling Type I Errors in Moderated Multiple Regression: An Application of Item Response Theory for Applied Psychological Research

by

BRENDAN J. MORSE

has been approved for

the Department of Psychology

and the College of Arts and Sciences by

Rodger W. Griffeth

Professor of Industrial/Organizational Psychology

Benjamin M. Ogles

Dean, College of Arts and Sciences

ABSTRACT

MORSE, BRENDAN J., Ph.D., August 2009, Industrial/Organizational Psychology

Controlling Type I Errors in Moderated Multiple Regression: An Application of Item Response Theory for Applied Psychological Research (246 pp.)

Director of Dissertation: Rodger W. Griffeth

Applied psychologists have long recognized the importance of measurement as a key component of research quality, but the use of psychometrically sound measurement practices has not kept pace. Recent evidence has emerged to suggest that weak measurement practices can have serious implications for the accuracy of parametric statistics. Two simulation studies (Embretson, 1996; Kang & Waller, 2005) have identified that response score scaling and assessment appropriateness heavily influence the Type I error rate for interaction effects in moderated statistical models when simple raw scores are used to operationalize a latent construct. However, the use of item response theory (IRT) models to rescale the raw data into estimated theta scores was found to mitigate these effects. The purpose of this dissertation was to generalize these results to polytomous data that is commonly found in applied psychological research using a Monte Carlo simulation. Consistent with the previous studies, inflated Type I error rates for the interaction effect in a moderated multiple regression model were observed when raw scores were used to operationalize a latent construct. In the most extreme cases, this inflation approached 85%. Also consistent with previous studies, psychometric factors were found to have a greater impact on raw scores than on estimated theta scores, and assessment appropriateness was found to be the most

influential factor on the empirical Type I error rate. Inconsistent with previous studies, an inflated Type I error rate was also observed under some conditions for the estimated theta scores suggesting that the graded response model (GRM) may not have provided a sufficiently equal-interval metric. Additionally, the expected interaction between assessment appropriateness and assessment fidelity was not found to be significant. Overall, these results suggest that the IRT-derived scores were more robust to spurious interactions than simple raw scores, but may still result in inflated Type I error rates under some conditions. The implications of these results are discussed from two perspectives. The performance of the GRM under the simulated conditions is emphasized for measurement researchers, and the usefulness of model-based measurement practices for improving research quality is emphasized for applied psychologists.

Approved: _____

Rodger W. Griffeth

Professor of Industrial/Organizational Psychology

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge my family, Deborah and Gary Morse, for their unwavering support throughout my long educational pursuits. Everything that I have achieved has been enabled by your love, encouragement, and belief in me. This simply would not have been possible without you.

I would also like to thank my co-advisors, Dr. Rodger Griffeth and Dr. George Johanson, for their enthusiastic support of this dissertation. The success of this project is due in large part to your collegiality and willingness to work together from very different perspectives. I have the utmost respect for you both, and I will begin my career as a better person and scholar due to your wisdom and advice, both academic and otherwise.

Finally, I would like to extend warm thanks to the additional members of my dissertation committee, Dr. Paula Popovich, Dr. Jeffrey Vancouver, and Dr. Victor Heh, for their time and expertise on these topics. Your insights were greatly appreciated.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1: INTRODUCTION

Experimental design and the measurement of psychological constructs represent two important facets of the overall quality of research in the behavioral sciences. Deficiencies in methodology and/or measurement can result in misleading and inaccurate findings with regard to a phenomenon of interest. Most researchers are well aware of the consequences of weak experimental design, but there is often a lack of discussion regarding the consequences of poor measurement practices.

Summative reviews have revealed troubling trends for the reporting and understanding of psychometric data in applied psychological research. Scandura and Williams (2000) found that less than half of the studies conducted in applied psychology identified any reliability information for the assessments that were used in the studies. Additionally, Podsakoff and Dalton (1987) found that less than four percent report any validity information. Perhaps eve more troubling, Cortina (1993) and Phillips and Lord (1986) found that many researchers misinterpret the definition and nature of psychometric indicators, such as reporting measures of reliability (Cronbach's alpha) as evidence for construct validity or unidimensionality. Finally, Austin, Scherbaum, and Mahlman (2002) concluded that regardless of the availability of validated instruments, researchers in Industrial and Organizational (I/O) psychology tend to use "garden variety" scales that have little if any psychometric evaluation.

Although the importance of measurement is often deemphasized in the overall discussion of experimental quality, the implications of such practices are hardly insignificant. Stone-Romero (1994) cogently argued that the use of measures lacking

psychometric quality is a systemic flaw for the interpretability and applicability of the research itself. From a more reflective position, Smith and Stanton (1998) iterated that there is a need for a better "measurement culture" in applied psychology. "Perhaps journal editors and reviewers could, in greater numbers, join an effort to stop rewarding "adhocracy" and start rewarding a deliberate, thoughtful, and long-term approach to measurement" (Smith & Stanton, 1998, p. 381). Clearly, there is a need to increase our understanding and communication of the quality of our measurement practices.

When measurement issues are discussed, the primary emphasis is often on the construct validity of a particular assessment. However, the psychometric characteristics of an assessment have also been found to influence the results of parametric statistical tests (Busemeyer, 1980; Davison & Sharma, 1988; 1990). One such parametric test, the identification of a significant interaction effect in analysis of variance (ANOVA) or moderated multiple regression (MMR), is a popular analytic procedure in applied psychological research (Stone, 1988). In these analyses, the interaction term, or non-additive moderator, is the effect of one independent variable on a dependent variable that varies as a function of (at least) one other independent variable (Aiken & West, 1991). Decisions about the statistical significance of these effects are made on a probabilistic basis, and it is important for researchers to be aware of the potential decision errors associated with null hypothesis statistical testing.

The decision errors related to evaluations of a null hypothesis come in two forms, Type I and Type II errors. A researcher who commits a Type I error erroneously rejects a null hypothesis, concluding that a statistically significant effect is present when no effect

actually exists at the population level. A researcher who commits a Type II error erroneously fails to reject a null hypothesis, concluding that no statistically significant effect is present when an effect actually does exist at the population level. Rosenthal and Rosnow (2008) cleverly refer to these as errors of gullibility and errors of blindness respectively. A commonly held convention in psychological research is that statistical significance is determined when effects are observed at an alpha level of $p < .05$. This criterion translates into a 5% probability of committing a Type I error. Therefore, the Type I error rate typically refers to the overall accuracy of a test. Likewise, the probability of correctly rejecting a false null hypothesis is defined as the statistical power of a test. Statistical power is the inverse of Type II errors and refers to the sensitivity of a test. Together, these decision errors represent a balancing act that is directly tied to the overall quality of our research.

Often in applied research, the sensitivity of a test for moderators, or the prevalence of Type II errors, is the most salient concern. A variety of issues such as the relationship between predictors (multicollinearity) and population level effect sizes contribute heavily to Type II errors (Aguinis & Stone-Romero, 1997; McClelland & Judd, 1993; Stone, 1988; Zedeck, 1971). However, recent theoretical and simulation research has found that measurement factors such as assessment appropriateness and response score scaling can influence the prevalence of Type I errors for moderator effects (Davison & Sharma, 1990; Embretson, 1996; Kang & Waller, 2005; Maxwell & Delaney, 1985). Kang and Waller (2005) found, in some conditions, that the Type I error rate for the moderator term in a moderated multiple regression analysis exceeded 40% depending

on various psychometric factors. These findings suggest that understanding the specific psychometric qualities of our assessments is also an important aspect of accurate statistical decisions.

An increasingly championed approach for facilitating our understanding of psychometric qualities is a modern measurement methodology known as item response theory (Embretson & Reise, 2000; Harwell & Gatti, 2001). Item response theory (IRT) is an item-level evaluation of the psychometric properties of a latent construct assessment, as well as a means of operationalizing the latent construct itself (Drasgow & Hulin, 1990; Embretson & Reise, 2000; Zickar, 1998). Derived from the seminal work of Frederick Lord (Lord, 1953b; Lord & Novick, 1968), and research conducted in the U.S. Air Force during the 1940s and 1950s, IRT has several characteristics that make it unique from classical test theory approaches to measurement. Specifically, the invariance of item and person parameters, the computation of a variable standard error of measurement, and the scaling of the response scores allow IRT models to identify information that is inaccessible by other approaches to measurement. These characteristics have also been found to have beneficial effects for the accuracy of parametric analyses such as factorial ANOVA (Embretson, 1996) and MMR (Kang & Waller, 2005). As a result of these advantages, many researchers have advocated that IRT be considered standard psychometric procedure in both general and applied psychology (Borsboom, 2008; Drasgow & Hulin, 1990; Embretson & Reise, 2000; Thissen & Steinberg, 1988; Zickar, 1998).

My goal for this dissertation will be to extend our understanding of the psychometric conditions that may contribute to an increased risk of Type I errors in moderated statistical models. A Monte Carlo simulation will be conducted to investigate the relationship between response score scaling, assessment appropriateness, and the potential for inflated Type I errors for interaction effects in moderated multiple regression when a Likert-type scale is used to measure a latent construct. Embretson (1996) and Kang and Waller (2005) conducted similar simulation studies and found promising results for the benefits of IRT. However, these studies simulated dichotomous response data and applied restrictive IRT models, thus reducing their generalizability for general and applied psychological research. Simulating latent construct assessments with multi-category response formats and applying a polytomous IRT model will be an important extension of this work. Such scales are more representative of the latent construct assessments utilized in general and applied psychological research (Aguinis Pierce, Bosco, & Muslin, 2009; Austin, et al., 2002; Fields, 2002), and polytomous IRT models are available for these assessments (Embretson & Reise, 2000; Ostini & Nering, 2006).

An additional goal in this dissertation will be to demonstrate IRT as a data scaling technique that has the potential to help increase the accuracy of parametric models in applied psychological research. For example, in I/O psychology, the analysis of moderators is among the most popular goals of research (Aguinis, 2004; Stone, 1988), and self-report scales for latent construct assessments comprise a majority of the data that is collected (Aguinis et al., 2009; Austin et al., 2002). Additionally, an overwhelming majority of these self-report scales use multi-category, Likert-type response metrics

(Fields, 2002) which result in scores with long debated mathematical properties (Stine, 1989). An obvious aim in this field is to accurately test hypotheses of interaction effects with Likert-type data. The unique qualities of IRT can provide methods by which I/O psychologists can increase the overall accuracy of their research and practice. Specifically, the robustness of IRT scoring procedures to Type I errors in moderated analyses could prove to be very beneficial for basic researchers evaluating I/O theory, as well as for practitioners making recommendations to organizations and legal entities.

CHAPTER 2: LITERATURE REVIEW

Previous authors in the field of I/O psychology have championed the potential benefits of IRT (Drasgow & Hulin, 1990), but it has remained largely underrepresented in the applied literature (Aguinis et al., 2009; Austin et al., 2002). In the past 25 years, fewer than 50 studies have been published in top I/O research outlets compared to over 500 in educational research and measurement journals. Embretson and Reise (2000) provide a fascinating historical lineage for IRT that explores its largely intradisciplinary nature within the field of educational testing. The reasons for this have largely been attributed to early barriers of computational complexity and practicality for applied research as well as a general lack of interdisciplinary research training. The implementation of IRT in future studies in applied psychology will be an important step forward for the quality of measurement in our field.

Researchers that have utilized IRT in I/O psychology often have a very specific focus such as using differential item functioning to identify faking motivations in job applicants (Zickar, Gibby, & Robie, 2004) or to evaluate the quality of performance appraisal raters (Facteau & Craig, 2001). Other studies have used IRT to investigate the specific psychometric properties of popular I/O assessments (Hulin & Mayer, 1986; Reeve & Smith, 2001; Zagorsek, Stough, & Jaklic, 2006). Although these applications are certainly informative, a broader integration of the benefits of IRT in applied psychological research may be achieved by widening the net of applicability. Therefore, I will aim to demonstrate a capability of IRT as an analytical tool that is generalizable for a

diverse range of research in I/O psychology. In accomplishing this goal, a better case can be made for the adoption of this useful methodology by researchers in this field.

Moderator Analyses in Psychological Research

Moderator variables, or non-additive components of a predictor-criterion relationship, have been used substantially throughout the history of psychological research (Aguinis, 2004; Aiken & West, 1991). A moderator is typically defined as a variable that creates a conditional relationship between a predictor and a criterion. In other words, the relationship between the predictor and the criterion depends on the level of the moderator variable (Cohen, Cohen, West, & Aiken, 2003; McClelland & Judd, 1993; Stone, 1988). Two common analytical models for moderator detection are factorial analysis of variance (ANOVA) and moderated multiple regression (MMR). Factorial ANOVA is an appropriate methodology for manipulated experimental designs in which the manipulated conditions represent categorical independent variables, or fixed effects. Likewise, MMR is often used in quasi and non-experimental research and can analyze categorical independent variables (fixed effects), as well as continuous independent variables, or random effects, although this final condition has been met with some debate (c.f., Fisicaro & Tisak, 1994; Sockloff, 1976). It is important to note that both types of variables (categorical or continuous) can be used in either analysis (ANOVA or MMR), but it is considered to be the least desirable to introduce continuous variables as independent variables into ANOVA, as some artificial dichotomization must occur such as a median split. This practice is often associated with a loss of variance information (Aiken & West, 1991; Cohen et al., 2003).

Of these two analyses, MMR is perhaps the most common technique in applied psychological research (Aguinis, 2004; Aguinis, Beaty, Boik, & Pierce, 2005; Aguinis & Stone-Romero, 1997; Stone-Romero, Alliger, & Aguinis, 1994). Aguinis (2004) indicates that MMR has existed as an analytic technique for well over 50 years, and has proved to be robust to mathematical as well as conceptual shortcomings of alternative procedures proposed during that time. In the recent history of applied psychological research, Aguinis et al. (2005) indicate that MMR has been used in approximately 20 to 40 articles per year in the *Journal of Applied Psychology, Academy of Management Journal,* and *Personnel Psychology* over the past 30 years. In these studies, MMR has been applied to a wide variety of topic areas such as "job performance, job satisfaction, training and development, turnover, pre-employment testing, performance appraisal, compensation, organizational citizenship behaviors, team effectiveness, perceived fairness of organizational practices, self-efficacy, job stress, and career development" (Aguinis et al., 2005, p. 94). Additionally, MMR can be used in personnel selection, such that the existence of significant moderators like minority group status can imply assessment bias (Bartlett et al., 1978; Society for Industrial and Organizational Psychology, 1987), and these findings are regarded as key indicators in legal challenges to employment policies (Cascio & Aguinis, 2005). In fact, Lubinski and Humphreys (1990) attribute this use of moderator analyses as the initial impetus for the technique. Indeed, MMR is a pervasive and useful procedure for a variety of different types of variables in applied psychological research.

In MMR, a predictor – criterion relationship is composed of at least two independent variables (predictors) that each produce a main effect and collectively produce an interaction (moderator) effect on the dependent variable (criterion). One independent variable is determined *a priori* to be the moderator variable upon which the relationship between the other independent variable and the dependent variable vary. The MMR model is typically expressed in a hierarchical structure such that the first model contains the additive main effects of the predictor and the moderator, and the second model contains the additive main effects plus the multiplicative interaction effect. This relationship can be given in the general form in Equations 1a and 1b.

$$y = a + b_1 x + b_2 z + \varepsilon \tag{1a}$$

$$y = a + b_1 x + b_2 z + b_3 x \cdot z + \varepsilon \tag{1b}$$

In the additive model expressed in Equation 1a, y is a dependent variable, x is a predictor variable, z is the moderator variable, a is an intercept term, b_i are regression weights, and ε represents residual error. In a regression model fitting this form, x and z are predicted to be both additively related to y and represent the main effects of x and z. In Equation 1b, the interaction term $b_3 x \cdot z$ has been added to the model to represent the multiplicative effect of x and z on y. The key feature of a significant moderator in this form is that the relationship between x and y varies as a function of z. The significance and impact of a moderator is typically identified by examining the change in the amount of variance

accounted for in the criterion (ΔR^2) when the interaction term is introduced in the second step of the hierarchical model (Aguinis, 2004; Aiken & West, 1991; Cohen et al., 2003).

Job Satisfaction as an Illustrative Construct

Job satisfaction is perhaps one of the most commonly researched constructs in applied psychology, and it has played a ubiquitous role in theories of organizational behavior and in job attitudes measurement initiatives (Smith & Stanton, 1998). As such, the measurement and implementation of job satisfaction in applied psychological research has several key features that are related to the research goals presented in this dissertation. First, job satisfaction is a latent construct that is often assessed with multicategory, or polytomous, assessments. Second, job satisfaction has a variety of long-standing assessments devoted to its measurement that may have psychometric features that can influence statistical analyses (c.f., Morse & Griffeth, 2009). Finally, job satisfaction is a popular component in a variety of moderated theoretical models and as such, it is subject to the aforementioned statistical influences.

Job satisfaction is typically defined as a cognitive and affective evaluation of an individual's job (Hulin & Judge, 2003). Job satisfaction has been defined as both a global and facet construct, with popular assessments designed from each perspective such as the *Job in General Scale* (global measure) and the *Job Descriptive Index* or the *Minnesota Satisfaction Questionnaire* (facet measures). Hulin and Judge (2003) identify a variety of theoretical models of job satisfaction such as the Job Characteristics Model (Hackman & Oldham, 1976) in which characteristics of job tasks: task identity, task significance, skill variety, autonomy, and feedback, are moderated by an individual construct (growth need

strength) to influence various attitudinal outcomes such as motivation and satisfaction. This model can be broadly classified into a group of theories known as person by environment (PxE) interactions.

The underlying structure of a person by environment interaction for job attitudes such as that seen in the Job Characteristics Model continues to be popular in applied psychological research. Indeed, more contemporary theoretical models of job satisfaction champion the person–environment fit perspective (Dawis, 1992), in which the focus is on fully moderated relationships. However, these complex multivariate models are in need of more research (Brief, 1998; Hulin & Judge, 2003). Heeding this call for research is essentially asking for a moderation analysis in which job satisfaction is a dependent variable that is measured and analyzed as a latent construct. It would be prudent then to closely examine the measurement properties of job satisfaction assessments and the operationalization of the job satisfaction construct when evaluating theoretical frameworks that rely on the existence and interpretation of interactions.

Research that uses a construct such as job satisfaction as a central theme in a larger model is also subject to the effects of measurement artifacts. As an example, consider classical and modern theories of voluntary employee turnover. Job satisfaction has historically been a central mechanism in models of organizational behavior (Brayfield & Crockett, 1955; March & Simon, 1958), and research on employee turnover has touted job satisfaction as a key construct (Lee & Mitchell, 1994; Mobley, Griffeth, Hand, & Meglino, 1979). Often, job satisfaction is considered to be central (in terms of location in a theory/model), such that there are antecedents and consequences of

satisfaction that have important implications for predicting behavior. Additionally, many of these relationships rely on the moderating effects of various antecedents on constructs such as job satisfaction. Understanding the situations in which moderators are assessed can lead one to explore the statistical issues in moderator detection. Because inferential statistics rely on probability-based decision making, of particular concern are the factors that influence the prevalence of Type I and Type II errors related to decisions about the null hypothesis for the moderator analysis.

Statistical Issues in Moderator Detection

As previously demonstrated, the detection of significant moderators and interactions in applied psychology has occupied an important research goal. Aguinis (2004) argues further that this technique is only becoming increasingly popular. However, the accurate detection of moderators has been fraught with methodological and statistical difficulties. A particularly important question in this area has focused on the barriers to accurate moderator detection. That is, what predispositions exist that increase the likelihood of Type II errors for moderator effects? In response to this general query, a variety of simulated (e.g., Monte Carlo) and empirical studies have identified overall sample size, the inequality of moderator subgroup sample sizes (for categorical variables), multicollinearity of predictors, range restriction, artificial dichotomizations, and population level moderator effect sizes as some of the risk factors for Type II errors in moderator detection (Aguinis & Stone-Romero, 1997, Stone-Romero et al., 1994; Zedeck, 1971). The general consensus has been that any number of the aforementioned

factors will result in low statistical power in MMR analyses and increase the likelihood of Type II errors for the moderator term.

A final effect impacting the detection of significant moderator effects that warrants mention is the predictive strength of the additive model. Specifically, Rogers (2002) identified substantial mathematical constraints imposed by the variance in the criterion that is accounted for by the main effects in the model. For example, when the independent variables are continuous, the effect size of an interaction varied as a function of the strength of the additive model and the correlation between the independent variables, or multicollinearity. Ironically, under simulated conditions of no multicollinearity which is regarded as methodologically and statistically ideal, the maximum effect size of the interaction term was less than .4 for in the strongest additive model ($R^2 = .70$). As the predictive ability of the additive model decreased, the interaction effect size further plummeted to levels below .15 as the multicollinearity approached and exceeded 0. Rogers (2002) concluded his study with a basic, yet powerful implication, "Simply stated, to have strong ordinal moderation, there must be a strong effect to be moderated" (p. 223).

These factors are certainly worthy of attention for any researcher interested in these analyses, and a large number of publications are available to specify the behavior of these statistics (c.f., Aguinis & Stone-Romero, 1997; Aiken & West, 1991 chapter 3; Jaccard & Turrisi, 2003 chapter 4; Morris, Sherman, & Mansfield, 1986; Paunonen & Jackson, 1988; Stone-Romero et al., 1994; Zedeck, 1971). However, recent studies have demonstrated that an inflated Type I error rate for decisions related to moderators is also

a concern. Specifically, theoretical elaborations (Davison & Sharma, 1988; 1990; Maxwell & Delaney, 1985) and simulation studies (Embretson, 1996; Kang & Waller, 2005) have demonstrated that the psychometric properties of an assessment used to measure latent constructs, and the operationalization of the construct itself can lead to an increased risk for spurious interaction effects.

Specifically, two conditions lend greater opportunity for this effect to occur. First, the use of raw scores to quantify the dependent variable may well violate assumptions about the scale of measurement on which the latent construct is operationalized. Specifically, latent constructs are thought to exist at least at the interval level, but raw sum scores calculated under the classical test theory model likely do not exceed ordinal level properties (Borsboom, 2008; Harwell & Gatti, 2001; Maxwell & Delaney, 1985). Second, when the reliability of the assessment and the distribution of individuals' construct scores are poorly matched, a condition arises known as "assessment inappropriateness". Assessment inappropriateness can be thought of as arising when a group of individuals take an assessment that is either very difficult or very easy based on their abilities. The results of such an assessment will typically lead to restricted scores due to floor and ceiling effects caused by the relative difficulty of the items. Response score scaling and assessment inappropriateness have both been demonstrated to heavily influence the occurrence of Type I errors in moderator analyses (Embretson, 1996; Kang & Waller, 2005). These two conditions will be described in detail below.

Response Score Scaling

A central issue to the identification of spurious interaction effects is the scale of measurement on which the observed, or manifest, variable can be classified. Stevens (1946) presents the classic scales of measurement in psychological research as nominal, ordinal, interval, and ratio. These scales are classified by the assumed relationships between the data values as well as the admissible transformations that can be performed on the data. Nominal data simply represents group categorizations and allow only one-to-one transformations. Ordinal data suggests a successive order in group categories, but assumes no information about the magnitude of the inter-category differences. Ordinal data can be subjected to monotonic transformations that preserve the ordering of the categories. Interval data suggests order and meaningful differences between the data points, but no true zero point. Interval-level data permits linear, or affine, transformations such as in the structure of a linear regression equation with a multiplicative component and additive constant. Ratio data has both meaningful intervals and a true zero point that allows for similarity transformations. Several authors have advocated that the failure to preserve the scale of measurement for a particular variable can result in misleading statistical results (Borsboom, 2008; Harwell & Gatti, 2001; Maxwell & Delaney, 1985; Stine, 1989; Stevens, 1946).

Stevens' (1946) scales of measurement were derived primarily to provide a taxonomy of admissible calculations and transformations with regard to psychological data. In an influential text approximately twenty years earlier, the physicist N. R. Campbell (Campbell, 1928), argued that a numerical structure used to perform

mathematical operations must appropriately represent an empirical structure of the object(s) under inquiry. Campbell (1928) referred to this process as extensive, or fundamental, measurement. Stine (1989) succinctly summarized the basis of fundamental measurement in the following passage:

> … a given set of empirical relations could be represented by several equivalent sets of numerical relations. Given that each of these numerical representations (numerical structures) is of a common empirical structure, they should be related to one another. Equivalently, one can convert a given numerical structure into one of the other structures without changing the nature of the empirical structure that is represented. Certain changes, or *transformations*, of the numerical structure are, therefore, admissible in the sense that the empirical phenomena that are being described are invariant with respect to the relationships that define the transformation. (Stine, 1989, p. 147)

The properties and rules surrounding measurement practices specify that the scale of measurement and admissible transformations define the appropriateness of the mathematical or statistical functions that are used. In parametric statistics, it is assumed that the variables being measured attain at least an interval scale of measurement for deriving meaningful conclusions. However, raw scores from many psychological assessments are thought to be limited to an ordinal scale of measurement (Borsboom, 2008; Harwell & Gatti, 2001; Maxwell & Delaney, 1985; Stevens, 1946). Other authors will slightly relax this distinction and suggest that the majority of measurement in

psychological research takes place in a gray area in between the ordinal and interval scales (Gardner, 1975; Stine, 1989).

In the years since the initial publication of Stevens' scales, a fervent debate has ensued over the validity and importance of measurement scale as well as the appropriate statistical procedures that can be applied (Stine, 1989). This debate has primarily concentrated on the importance of the distinction between the ordinal and interval scales of measurement and the resulting implications for parametric statistics. One solution that has been pursued is the development of non-parametric, or distribution-free, statistics (Clogg & Shihadeh, 1994; Gibbons, 1993). Non-parametric analogues have been developed for nearly all parametric procedures and benefit from relaxing two primary assumptions of parametric procedures namely, that the scores in the population are normally distributed and the scale of measurement is at least at the interval level. Thus, data that are based on scales that are, at best, ordinal levels of measurement can be subjected to non-parametric procedures that preserve data rankings and do not violate measurement rules (Gibbons, 1993). However, many researchers are reluctant to use non-parametric techniques, as many parametric statistics have demonstrated adequate robustness to violations of these assumptions (Davison & Sharma, 1988), and non-parametric procedures are often associated with a loss of information pertaining to the nature of the variables (Gardner, 1975). One author articulated this point by saying, "Consequently, in using a non-parametric method as a short-cut, we are throwing away dollars in order to save pennies" (McNemar, 1969, p. 432). It is reasonably clear that

regardless of the pragmatic appropriateness of non-parametric tests, the preference for parametric statistics is unabashed.

In a review of the literature addressing the appropriateness of the scale of measurement on parametric statistics, Stine (1989) presents a variety of counterarguments for the practical importance of Stevens' scales of measurement. For example, some have argued that the statistics that are conducted with the numerical representations of psychological variables are "closed systems" (c.f., Anderson, 1961; Burke, 1953; Lord, 1953a). Those arguing from this perspective say that the scale of measurement is unimportant for the results of statistical analyses, as long as the distributional assumptions of the analyses themselves are met. Frederic Lord's popular quip, "The numbers don't remember where they came from," (Lord, 1953a, p. 751) represents this position. Another primary argument against Stevens are those of Gaito (1959; 1960) and Jenson (1980) who assert that a variable that is normally distributed can be assumed to have interval properties because the normal distribution can be subdivided into equal intervals. Finally, Gardner (1975) and Stine (1989) report the existence of simulation evidence pertaining to the invariance of statistical procedures when aspects of measurement scales are violated.

These counterarguments to Stevens (1946) would appear to suggest that the scale of measurement issue is of lesser importance than had been initially conceived. However, Stine (1989) provides an elegant rebuttal of these positions in terms of Stevens' original assumptions; ultimately concluding that the primary focus for researchers is the interpretability of statistical tests and the numerical – empirical structure relationship.

> The origin of the numbers used in a statistical analysis is central to the interpretation of the analysis. For the scientist, what counts is that there is an empirical analog to the numerical results, a situation that will occur only when the analog is viewed with the appropriate scale (Stine, 1989, p. 153).

Stevens would certainly agree with Stine's representation of these arguments, as he [Stevens] considered the application of inappropriate statistical procedures as "illegal statisticizing" (Stevens, 1946, p. 679).

At this point, one might ask what are the situations in which the scale of measurement of an observed variable at the ordinal level can be demonstrated to have misleading results? It appears to depend on the specific procedure. For instance, Davison and Sharma (1988) and Maxwell and Delaney (1985) demonstrated through mathematical derivations that there is little cause for concern regarding scaling in comparing mean group differences in the independent samples t-test. However, in another derivation, Davison and Sharma (1990) subsequently demonstrated that scaling-induced spurious interaction effects can occur with ordinal-level observed scores in multiple regression analyses.

To test these derivations, Embretson (1996) and Kang and Waller (2005) conducted simulation studies to demonstrate that characteristics of the assessment used to measure dependent variables can influence the detection of interaction effects in typical parametric analyses. Specifically, the researchers investigated whether a significant interaction effect would occur in a factorial ANOVA or MMR (respectively) when raw scores of a latent construct assessment were entered into the model. The results of these

simulations demonstrated that spurious interaction effects were identified at rates above the nominal Type I error rate of $\alpha = .05$ when raw score composites were used to operationalize latent variables. As an empirical follow-up to Davison and Sharma (1990), Embretson and Kang and Waller also considered these to be scaling-induced interactions. The issues of the scaling of latent and manifest variables as well as an interval-level rescaling solution will be further discussed in the section on item response theory.

Assessment Appropriateness

A second contributing factor to the prevalence of Type I error inflation in moderator analyses is the idea of assessment appropriateness. Assessment appropriateness refers to the congruence of the reliability of a particular assessment to the distribution of the individuals' construct scores who respond to an assessment. Traditional forms of reliability such as Cronbach's alpha make the assumption that an assessment is equally reliable across the entire construct range for the individuals who take the assessment (Crocker & Algina, 1986). However, this assumption is a (known) oversimplification and is limited to test/scale-level psychometric evaluations, such as those found in classical test theory models (Hambleton, Swaminathan, & Rogers, 1991). An alternative estimate of reliability can be calculated as a cumulative function of item-level information curves with modern measurement theory models such as IRT. This item-level, variable measure of reliability allows for the possibility that an assessment may have peak reliability at one point along the continuum of the construct that is being assessed, whereas the distribution of construct scores of the individuals responding to the assessment may peak at a different place (see Figure 1). This situation creates assessment

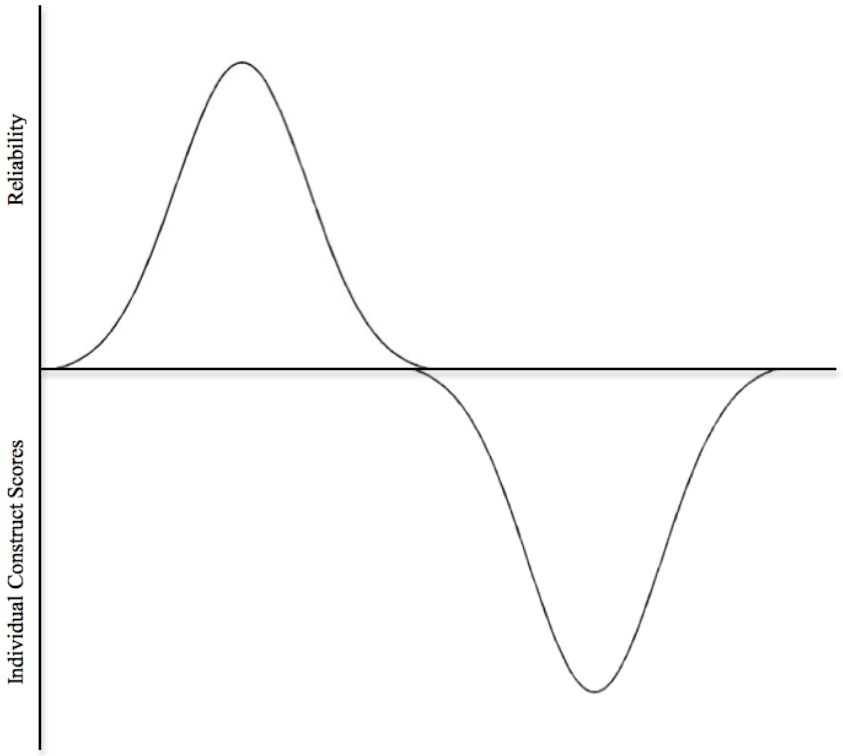

Figure 1. Assessment inappropriateness

inappropriateness, and has been shown to drastically inflate the occurrence of Type I errors in moderator detection (Embretson, 1996; Kang & Waller, 2005) as well as influence the validity of other statistics such as test-retest reliability (Fraley, Waller, & Brennan, 2000). To better understand this phenomenon and how to detect when an assessment is predisposed to it, it will be necessary to discuss the components of a modern measurement theory known as item response theory.

Item Response Theory

Item response theory is an item-level approach to the evaluation of the psychometric properties of a latent construct assessment, as well as the operationalization of the latent construct itself (Drasgow & Hulin, 1990; Embretson & Reise, 2000;

Hambleton et al., 1991). Derived from the more general term "latent trait theory" (Lord, 1953b), item response theory provides several unique and appreciable characteristics as a psychometric measurement model (Drasgow & Hulin, 1990; Embretson & DeBoeck, 1994, Embretson & Reise, 2000; Hambleton et al., 1991). A key feature of all item response theory models is the ability to examine item-level information such as the probability that an individual will respond in a particular response category for any item on an assessment given his or her standing on an underlying, latent construct. This item-level focus makes IRT conceptually and mathematically unique from the classical test theory (CTT) model that has dominated psychometric techniques since the early 20th century. The item-level focus of IRT is also beneficial for researchers seeking more specificity in their measurement data. In her reflection on 50 years of job attitude measurement research, Patricia Smith (Smith & Stanton, 1998) extolled the virtues of investigating data at the individual and item level. Although this general piece of advice was not specifically in reference to IRT, a measurement model from this perspective would likely be a welcomed tool. To fully appreciate the differences between IRT and CTT, it will be useful to examine the assumptions and general components that comprise most IRT models.

Latent Constructs

Item response theory models are particularly suited for evaluating assessments designed to measure latent constructs (Embretson & Reise, 2000). Latent constructs are unobserved, causal factors of human behavior (Bollen & Lennox, 1991; Borsboom, 2008; Maxwell & Delaney, 1985). In other words, the latent construct manifests behaviors that

can be measured. For example, an actual behavioral occurrence such as instances of inattentiveness in children suffering from attention deficit hyperactive disorder, or response patterns on a test such as five out of ten items correct on a math exam, or a certain pattern of responses on an attitude assessment are all instances of causal, latent constructs. Examples of latent constructs include (but are not limited to) individual characteristics such as general cognitive ability, an attitude such as job satisfaction, or a personality characteristic dimension such as conscientiousness. A latent construct assessment is any method that is used to measure the behavioral manifestations of the latent constructs within an individual. As noted above, these methods most commonly include ability tests such as a mathematics exam, self-report scales such as a personality or attitude inventory, or measures of actual behavioral occurrences. As a point of clarification, I will use the term "assessment" as a general reference to any type of construct measurement such as a test or scale. Additionally, I will use the term "construct" as a general reference to any type of latent characteristic such as an attitude, trait, or ability.

It is always prudent to appropriately specify the relationship between latent variables and their observed, or manifest, variables. Maxwell and Delaney (1985) describe this connection as a conditional relationship $f(Y|\theta)$ where Y represents a manifest variable that represents the latent variable theta (θ). Borsboom (2008) provides a succinct description of three different types of relationships between latent and manifest that specifies the intended scale of measurement for each. Specifically, these can be considered latent-continuous, observed-continuous; latent-continuous, observed-

categorical; and latent-categorical, observed categorical. It is important to note that "categorical" in this sense specifically refers to ordered categories. This distinguishes that the categorical representations are ordinal rather than nominal in nature. The commonality of all latent variable models is that the observed variables can be related to the latent construct with an appropriate regression function (Borsboom, 2008). For the purposes of CTT models, the appropriate relationship distinction for $f(Y|\theta)$ is latent-continuous, observed-categorical. This specifies that the latent construct exists at an interval scale and the observed score exists at an ordinal scale of measurement. In IRT models, this relationship is extended such that the manifest variable is measured at an ordinal scale but can be rescaled to an interval scale of measurement (Embretson & Reise, 2000; Harwell & Gatti, 2001; Reise, Ainsworth, & Haviland, 2005).

Assumptions

Item response theory models all assume three primary components that should be evaluated in order to achieve accurate results. First, with the exception of multidimensional models, IRT models assume that the latent construct is unidimensional in nature. Therefore, *a priori* steps must be taken to either verify that the construct being measured is indeed unidimensional, or multidimensional sub-factors must be analyzed separately as unidimensional scales (Embretson & Riese, 2000; Reckase, 1997). In early IRT work, this created a practical problem of instability in the parameter estimates due to dimensions with very few items (Drasgow & Hulin, 1990). Fortunately, improvements in the maximum likelihood estimation algorithms such as marginal maximum likelihood estimation and the expected a posteriori method allow IRT models to reliably handle

scales (or sub-factors) comprised of fewer items, thus enabling the latter strategy (Drasgow & Hulin, 1990). Additionally, valid arguments have been posed that question the existence of any truly unidimensional constructs (Drasgow & Hulin, 1990; Hambleton et al., 1991). It is now generally held that researchers should demonstrate that there is a dominant dimension present as evidence of practical unidimensionality.

There is substantial disagreement among researchers as to the best method of verifying the dimensionality of a construct assessment. Hambleton, et al. (1991) and Hattie (1985) provide very comprehensive lists of available methods for investigating dimensionality, although the authors implore that no single measure is thought to be definitive or appropriate for all situations. Additionally, there appears to be increasing support for the robustness of IRT models in mild to moderate violations of unidimensionality (Embretson & Reise, 2000). As an alternative, several IRT procedures currently exist for assessing multidimensional data such as "full-information factor analysis" (Reckase, 1997). However these procedures are still considered to be in the "infancy" stages of development. Additionally, commercially available software packages for these purposes have limited applications such as for dichotomous data only (Embretson & Reise, 2000).

A second component of IRT models related to dimensionality is that of local independence. Local independence is an IRT assumption stating that the relationships between items should be fully accounted for by the underlying latent construct (Embretson & Reise, 2000; Hambleton, et al., 1991). Another way of conceptualizing local independence is to say that the partial correlation between items should drop to

nearly zero in the presence of the underlying construct. Local independence is related to dimensionality such that local independence is obtained when the entire latent space has been identified and is accounted for in the model. Therefore, with a unidimensional construct, unidimensionality should imply local independence (Hambleton et al., 1991). It should be recognized however, that because strict unidimensionality is thought to be nonexistent in measures with more than three items (Drasgow & Hulin, 1990) this assumption is never explicitly satisfied. Finally, with multidimensional constructs, local independence can be obtained when all relevant construct dimensions are identified and accounted for.

One method that is suitable for checking both the dimensionality and the local independence of a particular assessment is to examine the item variance-covariance or correlation matrix. Specifically, this should be done within several homogenous subgroups along the construct continuum in order to partial out the relationships between items due to the latent construct. In the case where unidimensionality and local independence hold, the off-diagonal covariances or correlations should be very small (Hambleton et al., 1991).

Finally, IRT models are considered strong, or falsifiable, measurement models and are thus able to be evaluated for model-data fit. In a similar vein as the dimensionality arguments, there is ongoing debate as to the appropriateness of many IRT model-fit indices. A primary problem is that most measures of model fit use a chi-square based index. Similar to problems with other strong modeling methodologies such as structural equation modeling, increases in sample size will almost surely result in

rejection of model fit. In IRT, this is an especially strong paradox in that larger sample sizes are typically associated with better convergence in maximum likelihood estimations, but are also directly linked to a rejection of model fit (Drasgow & Hulin, 1990; Embretson & Reise, 2000). Currently, there are a variety of new model fit indices being evaluated. For example, the S-X^2 item fit index for polytomous models has been shown to be sufficiently robust to spurious violations of model fit when the an assessment has few items (Kang & Chen, 2007). However, most applied researchers will likely have to deal with some degree of model misfit.

The Invariance Property and the Theta Scale

Two final characteristics that are common to all IRT models are the invariance property and the theta scale. The invariance property is considered to be a defining feature of IRT that separates it from classical test theory (Drasgow & Hulin, 1990; Hambleton et al., 1991; Embretson & Reise, 2000). As a general prologue, the invariance property is an extension of the idea of specific objectivity (Rasch, 1977), which specifies that, "comparisons between objects must be generalizable beyond the specific conditions under which they were observed" (Embretson & Reise, 2000, p. 143). At the conceptual level, specific objectivity and invariance are akin to the idea of fundamental measurement promoted by Campbell (1928). At the practical level, this tenet of IRT suggests that item parameters derived from IRT models should generalize to other populations of individuals. However, Rupp and Zumbo (2006) appropriately caution that item invariance is often contrived as a "mysterious" property of IRT, justifying parameter appropriateness across an infinite range of populations when, in fact, it has defined limits. These limits,

primarily the indeterminacy of scale, will be discussed shortly, following the introduction of the theta scale.

As an example of the difference between IRT and CTT in terms of invariance, consider classical test theory approaches to evaluating an individual's score, item difficulty, and item discrimination. Under CTT, an individual's score is often reported as a raw number of correct responses (X) or a transformation of X such as a mean score. Further, item difficulty is measured as the proportion passing a particular item, and item discrimination is measured as the biserial correlation of each item to total score. These values are indelibly confounded in the CTT model such that true score is a function of the observed responses to a particular assessment. This creates a limiting factor in CTT of the sample used to evaluate the assessment (Embretson & Reise, 2000).

In IRT models, item parameters and individual construct scores are (nearly) invariant. When an IRT model is found to fit the data, a certain level of confidence in the invariance of item and person parameters can be achieved. This allows item properties to generalize to other populations, and it allows person parameters to be estimated with items from different assessments of the same construct that fit the same model (Embretson & Reise, 2000; Hambleton et al., 1991; Reise et al., 2005). However, one limiting factor is the issue of indeterminacy (Hambleton et al., 1991; Rupp & Zumbo, 2006). Indeterminacy stems from the fact that there is no "natural" scale for the latent construct. Therefore, because both person and item parameters are initially treated as unknown values in an IRT model, an arbitrary scale must be set initially in order to solve for the parameter estimates. Although indeterminacy is accepted as a conceptually

limiting factor of complete invariance (Rupp & Zumbo, 2006), it should be noted that mathematical derivations of popular IRT models such as the Rasch model have demonstrated the ability to achieve (nearly) fundamental measurement when the model fits the data (Embretson & Reise, 2000; Fischer, 1995; Perline, Wright, & Wainer, 1979).

The theta (θ) scale is the IRT measure of latent construct scores. In more technically appropriate terms, the theta scale is the numerical structure representing the empirical structure of the latent construct. Theta scores vary along a continuum from $-\infty$ to $+\infty$ although they typically fall within the range of -4.0 to 4.0. The theta scale has no natural scale, but is typically anchored at zero so as to represent a standardized scale in which a score of zero represents a moderate level of the latent construct, negative values represent low levels and positive values represent high levels (Hambleton et al., 1991; Embretson & Reise, 2000). Additionally, the probability of a particular response for person *j* on item *i* can be given as $P_j(\theta_i)$. For some IRT models such as the one-parameter logistic, or Rasch[1] model, the points along the theta scale simply represent a log-odds function. This implies that any particular theta value that is twice that of another theta value means that one individual is twice as likely to respond correctly (to a dichotomously scored item) than the latter individual. This property also holds for items. That is, any particular item that is twice as difficult as another item will have half the

[1] Georg Rasch (1901-1980) was a Danish statistician whose research is considered to be the seminal work for the one-parameter logistic IRT model and the concept of *specific objectivity* in scientific research (see Embretson & Reise, 2000 for a historical account of the influence of Rasch to the field of modern psychometrics). Proponents of the Rasch model consider it to be the only true objective measurement model, however, it is limited in that it only estimates a single parameter (item difficulty) and is thus too restrictive for many types of data.

probability of success for the same individual responding to the items (Embretson & Reise, 2000; Hambleton, et al., 1991; Reise et al., 2005).

The invariance property and the scale of measurement for theta are often cited as the reason that theta estimates, as an operationalization of the latent construct, are more appropriate than raw scores for use in some parametric analyses (Borsboom, 2008; Embretson, 1996; Embretson & DeBoeck, 1994; Kang & Waller, 2005; Wainer, 1982). Reise et al. (2005) state that, "Trait-level estimates in IRT are superior to raw total scores because (a) they are *optimal* scalings of individual differences (i.e., no scaling can be more precise or reliable) and (b) latent-trait scales have relatively better (i.e., closer to interval) scaling properties" (p. 98). Reise and Haviland (2005) give an elegant treatment to this condition by demonstrating that the relationship between the log-odds of endorsing an item and the theta scale form a linearly increasing relationship. Specifically, the rate of change on the theta scale is preserved (for all levels of theta) in relation to the log-odds of item endorsement. Embretson and DeBoeck (1994) also iterate the position that theta scores achieve interval level scale properties when the IRT model adequately fits the data. "An interval scale is obtained when the score distances between persons have equal meaning for ability differences" (Embretson & DeBoeck, 1994; p. 645). This sentiment reflects the congruence of the numerical and empirical structures advocated by Stevens (1946) and Stine (1989) that is achieved by IRT scoring. Finally, Perline, Wright, and Wainer (1979) elaborated this idea by specifying the one-parameter logistic, or Rasch model, as an empirical instantiation of additive conjoint measurement (an analogue to fundamental measurement proposed by Campbell (1928)), which justifies the existence

of an interval scale of measurement. These properties allow the theta scale to retain interval-level scale of measurement characteristics in the Rasch model as well as in more complex models.

An important question to now ask is how to justify when these characteristics extend to more complex IRT models such as the two and three parameter logistic models (dichotomous models with a discrimination and guessing parameter, respectively) and polytomous models. There is substantial agreement among researchers that although the non-Rasch models do not retain the property of specific objectivity, the answer to the question of whether non-Rasch models achieve interval-level scaling properties is "yes" (Embretson & Reise, 2000; Hambleton, et al., 1991; Harwell & Gatti, 2001; Reise et al., 2005). Harwell and Gatti (2001) conducted a simulation study investigating the congruence of estimated construct scores and actual construct scores using a popular polytomous IRT model, the graded response model. In this study, the authors posited that if the estimated construct scores were sufficiently similar to the actual construct scores which have interval-level scaling properties, then the graded response model results in theta scores that are sufficiently interval level. The results of their study confirmed this relationship, demonstrating that the differences between the estimated and actual construct scores were "within a range attributable to sampling error" (Harwell & Gatti, 2001; p. 126). These preceding arguments and findings lend strong support to the interval-level scaling of IRT derived theta scores as estimates of a latent construct.

Nonetheless, there has traditionally been some question as to the benefit of calculating theta scores from IRT models in lieu of simply using raw scores from an

assessment. It is typically found that individuals' raw scores on an assessment and their IRT-derived theta estimates are correlated at $r = .90$ or above (Drasgow & Hulin, 1990). Therefore, some would argue that the added complexity of IRT is unnecessary in deriving estimates of an individual's standing on the underlying latent construct. However, simulation studies have demonstrated that the primary drawback of raw scores is that they do not achieve interval scales of measurement (Embretson, 1996; Harwell & Gatti, 2001; Kang & Waller, 2005). This level of scaling is technically required for valid applications of parametric statistics, although little attention is often given to violations of this practice. In some cases, this violation of scaling can increase the likelihood of Type I errors beyond the nominal rate of $\alpha = .05$ for certain analyses. When the conditions leading to this possibility are present, an argument can be made for the benefit of using theta scores instead of raw scores as the operationalization of a latent construct.

Estimating Item and Person Parameters in IRT

The estimation techniques for item and person parameters in IRT are functionally unique from other forms of psychometric evaluations (Baker & Kim, 2004). In IRT models, item and person parameters are initially unknown values and do not have a natural scale. Therefore, to make a particular model identifiable, an anchoring decision must be made (Embretson & DeBoeck, 1994; Reise & Haviland, 2005; Rupp & Zumbo, 2006). In most cases, the estimation procedure begins by anchoring the theta scale at a mean of zero and a standard deviation of one. This is considered person-anchoring (Reise & Haviland, 2005). A maximum likelihood estimation procedure is then used to estimate the item parameters from the initial anchor. The item parameters are then treated as real

values and used to re-estimate the person parameters. This process is iterated until a reasonable convergence is reached where further iterations do not result in changes in the values beyond some specified criterion. The overall goal is to maximize the likelihood (L) that an entire set of data of n item responses provided by N individuals was achieved given theta and item parameters (Drasgow & Hulin, 1990; Hambleton, et al., 1991; Embretson & Reise, 2000).

In early IRT models, joint maximum likelihood was used for the estimation procedure (Drasgow & Hulin, 1990). Joint maximum likelihood assumes consistency (convergence), asymptotic efficiency (the consistent estimator has the smallest standard error in large samples), and an asymptotic normal distribution. However, joint maximum likelihood requires a large sample size and item pool, and there is debate as to whether or not the estimates can actually converge (Drasgow & Hulin, 1990). Lord (1968) specified that a sample of 1,000 individuals and a pool of 50 items are necessary for convergence in joint maximum likelihood, which hampered the widespread development and use of IRT until the early 1980s when the estimation methods improved (Embretson & Reise, 2000).

Due to advances in computational power, a marginal maximum likelihood procedure is now used for most IRT models. In marginal maximum likelihood, values for theta are not treated as unknowns. Instead, they are assumed to be sampled from a known or specified distribution (Drasgow & Hulin, 1990; Embretson & Reise, 2000). Marginal maximum likelihood can still require fairly large sample sizes and item pools, but is often preferred because it is more consistent than joint maximum likelihood (i.e., it reaches

appropriate convergence). Many advances have been made in the algorithms implemented in IRT programs that allow for reasonable estimation with much less stringent requirements. Swaminathan and Gifford (1985) found that reasonable convergence can be reached with a sample size of 50 and a pool of 15 items, and Drasgow (1989) reached convergence with a sample size of 200 and a pool of 5 items. Drasgow and Hulin (1990) consider this improvement to be "one of the most successful areas of research in psychology in the last few years" (p. 602). Additionally, other estimation techniques such as the expected a posteriori (EAP) technique have been supported for conditions when assessments have few items (Mislevy & Stocking, 1989). Currently, most IRT models implement a marginal maximum likelihood or EAP procedure for generating item and person parameter estimates.

Dichotomous IRT Models

IRT models can be grouped into classes based on the nature of the assessment that they are designed to model. The most basic IRT models are designed for assessments that utilize dichotomous response scales. There are three dichotomous IRT models called the one- two- and three-parameter logistic models. Equation 2 represents the item response function (IRF) for the one-parameter logistic (or Rasch) model.

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad i = 1, 2, \ldots, n \tag{2}$$

Using equation 2, we can identify that $P_i(\theta)$ is the probability of a correct response on item i given a particular level of theta. The parameter b_i represents item

difficulty, and *e* represents a transcendental number with the value of 2.718. This equation can be observed graphically in Figure 2 as the nonlinear, monotonic (or ogive) IRF predicting an individual's probability of success on item *i* given their ability (θ) that is asymptotic at 0 and 1. As one would expect, the probability of success increases as

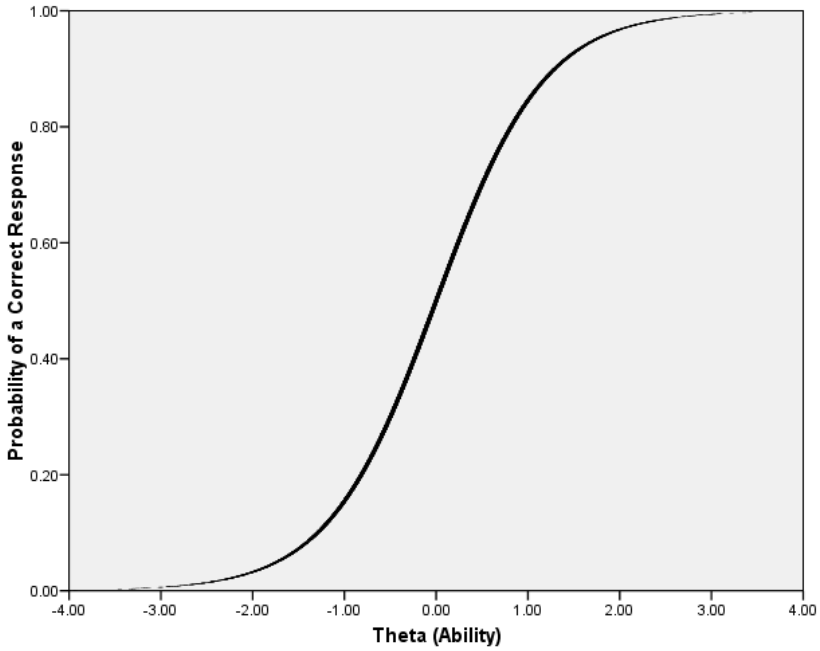

Figure 2. Item response function (IRF) for a one-parameter logistic model

examinee ability increases or as item difficulty decreases. The reverse is also true. This relationship represents the log odds function of item difficulty and examinee ability specified in the Rasch model (Embretson & Reise, 2000), as well as the invariance of item and person parameters. If the difficulty of this item were to be raised or lowered, the IRF would move laterally along the theta scale to the right or left (respectively).

The single-parameter logistic model can be expanded to also include a discrimination parameter (two-parameter logistic model) that would change the slope of the curve to indicate more or less item discrimination, and a guessing parameter (three-parameter logistic model) that would modify the lower asymptote to account for guessing. These models are very useful for assessments of unidimensional constructs using dichotomous scoring. Also, as the most basic type of IRT models, they are very useful for illustrating the advantageous properties of IRT modeling.

In psychological research however, many latent constructs include polytomous scoring schemes such as a Likert-type scale or a behavioral occurrence scale with three or more response categories. Wainer (1982) argued that the very nature of the polytomous response format violates assumptions of Gaussian normality, and scores from such assessments are likely better suited to be operationalized with latent trait theory models such as item response theory. The three aforementioned logistic IRT models can only be used with this type of data if the multiple response categories are collapsed into a two-category solution. However, doing so will result in a severe loss of information, and researchers are well advised to avoid this tactic (Embretson & Reise, 2000; Ostini & Nering, 2006). Instead, polytomous IRT models have been developed, tested, and are available for use in modeling multi-category response data.

Polytomous IRT Models

Polytomous IRT models are appropriate for assessing items with more than two response categories. A good example of this is a Likert-type scale where we can conceptualize the likelihood of endorsing a particular response option as being dependent

upon the individual's standing on the latent construct being measured. A primary distinction between dichotomous item responses and polytomous item responses deals with the probability of a response at a category boundary versus the probability of a response in a particular category. Ostini and Nering (2006) explain that with dichotomous responses, the probability of responding positively rather than negatively at a category boundary is the same as responding in the positive category. With only two response categories, there is only one boundary, and the boundary decision probability equals the category decision probability. However, in items that have more than two response categories, there is always at least one category that has two distinct boundaries. In this case, the probability of responding in a particular category is a cumulative function of the adjacent dichotomous category boundary decisions (Embretson & Reise, 2000; Ostini & Nering, 2006). The individual category boundary decisions are referred to as category boundary response functions (CBRFs), and each item has one less CBRF than it has response categories (see Figure 3).

The CBRFs in polytomous IRT models specify the nature of the decisions that are made with regards to that item. However, most researchers are interested in the probability of responding in a particular category for various levels of theta. Instead of an IRF that was used to model the probability of a response to an item with dichotomous scoring, polytomous items utilize category response functions (CRFs; Ostini & Nering, 2006). CRFs determine the likelihood of endorsing a particular response option given an individual's theta value in a very similar way that the IRF does with the logistic models discussed above. Here, θ again represents an individual's standing on the latent construct

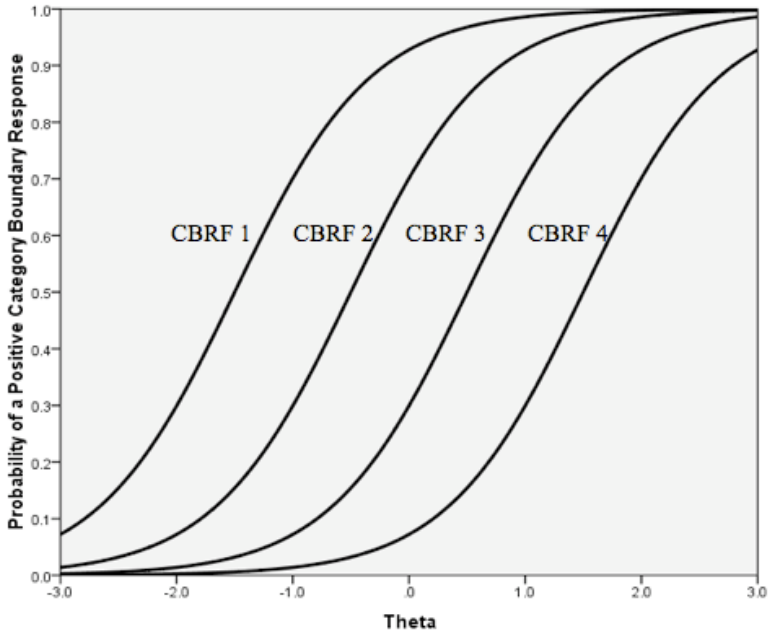

Figure 3. Category boundary response functions (CBRFs) for a 5-category polytomous model

being assessed. However, unlike the dichotomous IRF in which an individual's probability of a correct response (P_i) is indicated by a single ogive function, and the probability of an incorrect response is simply the inverse ($1 - P_i$); polytomous IRT models define as many CRFs as there are response options for each item. For example, an item rated on a 5-point Likert-type scale will have five CRFs with only the curves for the first and last response option represented as monotonically decreasing and increasing functions respectively (see Figure 4).

Polytomous IRT models that are suitable for common latent construct assessments can be classified as either divide-by-total or difference models (Ostini & Nering, 2006; Thissen & Steinberg, 1986). The primary distinction between the two is how the probability of responding in a particular category is operationalized at the category

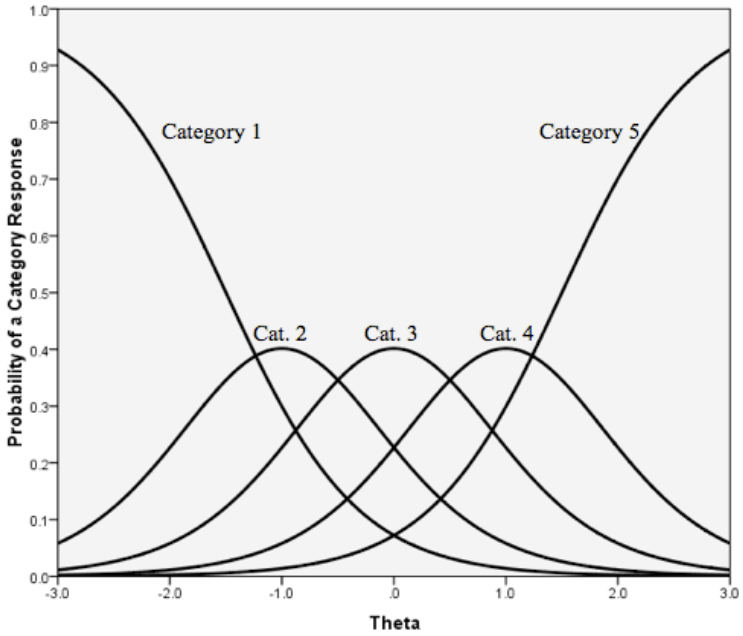

Figure 4. Category response functions (CRFs) for a 5-category polytomous model

boundary, or the nature of the CBRFs. In the divide by total family, the CBRFs represent the probability of responding positively as opposed to negatively between two adjacent categories only. These models are thought to represent category response decisions as a localized choice between two adjacent categories (Embretson & Reise, 2000; Ostini & Nering, 2006). In the difference family, the CBRFs represent the probability of responding above all of the preceding categories and below all of the subsequent categories. These models are commonly thought to retain the properties of Thurstone's scaling methods (Thurstone, 1927; 1928) in which an individual is theoretically attracted to each category (beginning with the first), and then subsequently attracted to the next category. A category is chosen when the attractiveness of the next category does not supersede the attractiveness of the current category (Embretson & Reise, 2000; Ostini & Nering, 2006).

Another important distinction between the divide by total and difference families is the manner in which the model was conceptually derived. Specifically, the divide-by-total models are Rasch-derivatives. In keeping with Rasch's goal of specific objectivity, models in the divide by total family such as the rating scale model (RSM; Andrich, 1978) are mathematically decomposable such that items invariance and interval-level scaling can be achieved (Embretson & DeBoeck, 1994; Embretson & Reise, 2000; Perline et al., 1979; Rupp & Zumbo, 2006). However, a drawback of the divide-by-total models is that they are considered to be more restrictive in their approach and thus may not be appropriate for some applications (Embretson & Reise, 2000). The difference models such as the graded response model (GRM; Samejima, 1969; 1996) focus on achieving conceptual response appropriateness based on the psychological processes that are likely to be at work when responding to an assessment (Embretson & Reise, 2000; Ostini & Nering, 2006). The importance of this approach for psychological measurement and research is evident in the measurement of job attitudes. Smith and Stanton (1998) presented a reflection of the construction of the Job Descriptive Index, a popular job satisfaction measure, in which they argued that an important aspect of measurement quality is keeping cognitive decision processes in mind during the development and validation of the assessment.

It is also important to note that several authors have advocated for the invariance and interval-level scaling properties in the non-Rasch difference models as well (Embretson & Reise, 2000). Indeed, Fraley et al. (2000), Harwell and Gatti (2001), and Kang and Waller (2005) have all identified simulated and empirical evidence to support

these claims. These findings suggest that the difference model family is both conceptually and mathematically appealing for use in psychological applications.

Reliability in IRT

A further benefit of IRT measurement models is that they allow researchers to calculate a variable standard error of measurement (SE_m) and its inverse function, information. In CTT, the assumption is made that the SE_m for scores on an assessment are uniform and restricted to the population that is being measured (Embretson & Reise, 2000). However, IRT models estimate SE_m as an item-level function that is optimally minimized when the characteristics of the item (difficulty) match the characteristics of the examinee (theta score). Thus, in the application of IRT to psychological assessments, we can optimize measurement by addressing the appropriateness of an item for a given individual based on their level of the underlying construct.

The variable SE_m property is the basis of item and assessment "information". In both dichotomous and polytomous IRT models, information and SE_m curves can be calculated for each item on an assessment as well as for the overall assessment. Item information is the precision of measurement (assessed as the minimization of the standard error of measurement) for item *i* across the ability continuum. Overall assessment information (scale reliability) is simply a cumulative function of the individual item information curves (Embretson & Reise, 2000; Ostini & Nering, 2006). This relationship between item and assessment information is conceptually and practically beneficial such that individual items can be selected that provide maximal information at various points

along the theta continuum. Indeed, this is the mechanism that enables computer adaptive testing.

In dichotomous models, item information is highest at the point at which the probability of a correct response is .50 (item difficulty matches examinee ability). Therefore, item *i* is most precise in individual's who have moderate levels of ability on this latent construct, but less precise for individuals with exceedingly high or low abilities (see Figure 5). Mathematically, item information for dichotomously scored items is

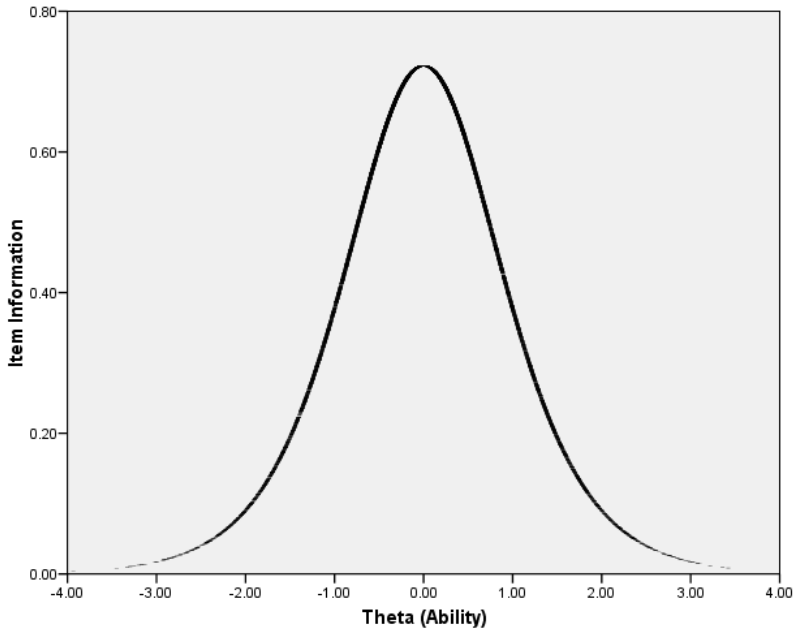

Figure 5. Information function

calculated as the largest first derivative of the item curve divided by the product of the probability of success and failure on that item (Hambleton et al., 1991).

In polytomous models, calculating information is slightly more complex due to the fact that polytomous items have as many CRFs as there are response options. There

are three primary methods of calculating information in these cases, namely the derivative approach, conditional expectations, or component elements (for a full explication of these approaches, see Ostini & Nering, 2006). Although the calculations are slightly different, the end result is the same in the polytomous case as in the dichotomous case. Item information represents the point along the theta continuum at which any particular item has the lowest SE_m. Individual items can be additively combined to represent an overall assessment information profile.

This type of estimate of measurement precision is not available with classical procedures because of the lack of local independence (Hambleton et al., 1991). Each item's contribution was always relative and a determinate of other items' contributions to overall test or scale quality. Variable item and assessment information can be used as useful diagnostic tools when evaluating an assessment for the possibility of measurement inappropriateness. Recall that assessment inappropriateness occurs when the peak measurement precision of an assessment is poorly matched to the distribution of individual construct scores along the theta continuum (Figure 1). This characteristic has been found to exacerbate the problem of spurious interaction effects due to the scaling of the dependent variable (Embretson, 1996; Kang & Waller, 2005). Although some instances of variable measurement precision can be identified in classical approaches, this feature of IRT as a central tenet of modern measurement theory is an advantageous diagnostic tool.

IRT in Parametric Statistics

For the purposes of drawing accurate statistical conclusions when collecting data based on latent construct assessments, the use of IRT has two primary benefits. First, the responses to a latent construct assessment can be operationalized at the interval-scale of measurement as theta scores. This is an important improvement over the use of raw score composites that typically do not exceed the properties of an ordinal scale of measurement (Embretson & DeBoeck, 1994; Harwell & Gatti, 2001). Second, the ability for IRT models to generate a variable SE_m and, likewise, more precise estimates of reliability is an important improvement over classical approaches. Some researchers have advocated that this property alone is a crucial mechanism in the appropriate assessment of statistical change (Fraley et al., 2000; Mellenbergh, 1999; Reise & Haviland, 2005). For the purposes of research in applied psychology, these two aspects of IRT also represent improvements in the assessment of moderated relationships.

The scale of measurement issue was central to the studies conducted by Embretson (1996) and Kang and Waller (2005). These studies successfully identified that raw score composites from latent construct assessments resulted in Type I error rates inflated above the nominal rate of five percent for moderator terms in both factorial ANOVA and MMR. In the Kang and Waller (2005) study, rates as high as 53% were observed when raw scores were used to operationalize the latent constructs and assessment inappropriateness was present. However, theta estimates from the one- and two-parameter logistic models were found to be robust to spurious interaction results in most cases. For example, Kang and Waller (2005) reported that the corresponding Type I

error rate for estimated theta scores in the condition identified above (with 53% errors) was only 9%. Although above the nominal rate of 5%, a researcher would certainly be more comfortable with the latter result. These results argue against the position that the added complexity of IRT is unnecessary for calculating individual scores. Indeed, if theta estimates are resistant to inflated Type I errors in parametric analyses through the achievement of interval-level scaling, a promising case can be made for more extensive applications.

The second benefit of IRT for parametric analyses is the ability to derive more specific psychometric information about the assessment that is being utilized. Specifically and perhaps most importantly, IRT models do not operate under the assumption that an assessment is equally reliable for all members of the population. This perspective on variable measurement precision is perhaps one of the most useful features of IRT models for applied psychological researchers (Embretson & Reise, 2000). A variable reliability (and variable SE_m) allows researchers to identify assessments that may be more appropriate for particular segments of the population. This is based on the point at which the reliability of the assessment peaks along the construct continuum. In cases where an assessment that is optimally reliable at one point of the construct continuum and the distribution of individual construct scores peaks at another point of the construct continuum, degrees of assessment inappropriateness occur. This situation was found by both Embretson (1996) and Kang and Waller (2005) to exacerbate the problem of Type I errors in moderator detection for raw score composites. However, even in these cases, IRT-derived theta scores were more resistant to elevated risks of Type I errors.

In cases where assessment inappropriateness heightens the potential for spurious interaction effects, it would seem prudent to reexamine many of our more popular construct assessments to see if this is a possibility. Without fitting an appropriate IRT model to the data from any such assessment, there is no way to determine if it is demonstrating these characteristics. Indeed, Morse and Griffeth (2009) identified that the Minnesota Satisfaction Questionnaire short form (MSQ-S; Weiss, Dawis, England, & Lofquist, 1967) exhibited this property with reliability peaking near the lower end of the construct continuum. This would suggest that the MSQ-S has greater measurement precision for those with lower levels of job satisfaction. As the MSQ is a very popular assessment of job satisfaction in applied psychology, the risk of Type I errors in moderator analyses based on the psychometric properties of this scale may be a salient concern. However, without an IRT-based analysis, this useful diagnostic information would go unnoticed.

The Current Study

To summarize, the preceding discussion has explored the possibility that the use of raw score composites and the specific psychometric properties of a latent construct assessment may lead to an elevated risk of spurious moderator detection in factorial ANOVA and MMR. Specifically, the scale of measurement of the manifest variable as well as the assessment appropriateness of the measure itself was demonstrated to influence the results of parametric analyses (Embretson, 1996; Kang & Waller, 2005). Further, the use of IRT-based approaches for operationalizing individual construct scores has been conceptually, mathematically, and empirically supported to be robust to these

effects. This robustness rests on the argument that IRT-derived construct scores are likely to achieve, or more closely approximate, an interval-level scale of measurement that is appropriate for use in parametric analyses that are popularly employed in applied psychological research (Borsboom, 2008; Embretson, 1996; Embretson & DeBoeck, 1994; Harwell & Gatti, 2001; Kang & Waller, 2005; Reise & Haviland, 2005; Reise et al., 2005; Wainer, 1982).

Currently, simulation data that substantiates these findings have been limited such that only dichotomous response format data has been studied (Embretson, 1996; Kang & Waller, 2005). Assessments with multi-category or polytomous response formats are much more popular in applied psychological research (Aguinis et al., 2009; Austin, et al., 2002; Fields, 2002). Therefore, the purpose of this dissertation will be to extend our knowledge about the influence of response score scaling and the psychometric properties of an assessment in moderator analyses using polytomous data and a polytomous IRT model.

Research Questions and Hypotheses

The primary research question to be addressed in this dissertation concerns the specific psychometric conditions that may create an increased risk for spurious interaction effects in moderated multiple regression. Therefore, empirical Type I errors are of primary interest. Specifically, these conditions will be explored as they relate to the use of latent construct assessments for the measurement of the independent and dependent variables that may be utilized in an MMR analysis. As an omnibus exploration

of the relationship between methods of operationalization of latent construct scores for a polytomous construct assessment, the following hypotheses are posed:

Hypothesis 1. Under conditions in which no significant interaction is present, the use of raw scores to operationalize a latent construct will result in higher Type I error rates than the use of actual or estimated theta scores derived using an IRT approach.

Hypothesis 1a. Under conditions of assessment appropriateness, the Type I error rates for raw scores, actual, and estimated theta scores will not exceed the nominal criterion of $\alpha = .05$.

The preceding hypotheses posit that IRT derived construct scores will perform better than raw scores in moderated multiple regression analyses. However, based on the results of Kang and Waller (2005), it is not likely that the overall Type I error rate for any scoring condition will exceed the acceptable $p < .05$ criterion under circumstances of assessment appropriateness. Under conditions of assessment inappropriateness where the reliability of the assessment does not match the distribution of construct scores of the measured sample, previous research indicates that there is likely to be a divergence in the performance of the aforementioned scoring conditions (Embretson, 1996; Kang & Waller, 2005). Based on this evidence, the following hypotheses are posed.

Hypothesis 2. Assessment inappropriateness will influence the prevalence of Type I error rates for the interaction term in moderated multiple regression.

Hypothesis 2a. Under conditions of assessment inappropriateness, the use of raw scores to operationalize a latent construct will result in Type I error rates that exceed the nominal criterion of $\alpha = .05$.

Hypothesis 2b. Under conditions of assessment inappropriateness, the use of estimated or actual latent trait scores to operationalize a latent construct will not result in Type I error rates that exceed the nominal criterion of $\alpha = .05$.

The preceding hypotheses posit that assessment inappropriateness will influence the Type I error rate for the interaction term in moderated multiple regression when raw scores are used to operationalize a latent dependent variable. This effect can also be influenced by the degree to which assessment inappropriateness exists. Specifically, Kang and Waller (2005) identified that the greater the divergence in appropriateness, the higher the Type I error rate for raw score conditions. Based on this evidence, the following hypothesis is posed.

Hypothesis 3. Under conditions of extreme assessment inappropriateness, the use of raw scores to operationalize a latent construct will result in the highest prevalence of Type I error rates beyond the nominal criterion of $\alpha = .05$ for the interaction term in moderated multiple regression.

Additionally, Kang and Waller (2005) identified that item discrimination and regression coefficients had an impact on the Type I error rate for the interaction term in moderated multiple regression when raw scores were used to operationalize a latent dependent variable. Specifically, the researchers found that conditions with higher discrimination

and stronger regression coefficients resulted in higher occurrences of Type I errors. Based on these findings, the following hypotheses are posed.

Hypothesis 4. Simulated assessments with higher item discrimination scores and stronger regression coefficients will result in the highest occurrence of Type I errors for the interaction term in moderated multiple regression when raw scores are used to operationalize a latent construct.

CHAPTER 3: METHODOLOGY

A particularly useful methodology for addressing this type of research question is computational modeling. Among a variety of techniques, one form of computational modeling is the creation of purely mathematical "environments" in which parameters can be defined and manipulated in order to observe the resulting effect on a statistic of interest. Such studies are often referred to as Monte Carlo simulations (Mooney, 1997). Although underutilized in industrial and organizational psychology research, computational modeling is beginning to gain ground as a useful methodology when time, resources, or knowledge of the behavior of a statistic are in low supply (Hulin & Ilgen, 2000; Zickar & Slaughter, 2002).

Monte Carlo simulations are a class of (mathematical) computational modeling in which the behavior of a statistic is estimated through repeated random trials of a defined pseudo population (Harwell, Stone, Hsu, & Kirisci, 1996; Mooney, 1997). Monte Carlo simulations are useful for tracking the properties and behavior of a statistic that is poorly understood due to factors such as weak theory, inadequate access to sufficient data, and/or poorly defined population parameters (such as violations of Gaussian normality). As such, simulations can provide a great deal of precision in specifying population parameters and manipulations for use in inferential testing.

Item response theory researchers have fruitfully used Monte Carlo techniques to explore many aspects of the applicability and validity of IRT models for a variety of circumstances (Harwell et al., 1996). For example, Harwell and colleagues indicate that a majority of Monte Carlo studies in IRT focus on identifying the robustness of parameter

recovery and estimation techniques for IRT models when applied with small sample sizes, skewed ability distributions, and violations of unidimensionality. The information provided by these techniques would be unattainable by other methods due to distributional and dimensional assumptions that have to be made. However, by implementing a Monte Carlo technique where the nature of these parameters can be controlled, the resulting information about the robustness of the IRT model can be much better specified.

The Monte Carlo approach is well suited for the goals specified in this dissertation. Specifically, the primary research question pertains to the behavior of a statistic when a "true" population condition is known. In addressing the question of the prevalence of Type I errors in moderated multiple regression, we must be able to determine when an error occurs. The only way of identifying this condition is by controlling a set of contrived population parameters and observing the behavior of a statistic under varied conditions. The averaged results of multiple iterations can then be recommended to other researchers as a reference when conducting future empirical studies.

Although Monte Carlo studies are a useful methodological technique, Harwell et al. (1996), Mooney (1997), and other Monte Carlo advocates warn, "the popularity of MC studies should not be taken as evidence that these techniques are methodological panaceas" (Harwell et al., 1996, p. 103). The design, execution, and verification of a Monte Carlo study are paramount to the validity of the results that are obtained. Harwell et al. (1996) developed a guide for conducting high quality Monte Carlo studies for IRT

research. Specifically, the authors outlined four important steps for successful Monte Carlo studies adapted for IRT assessments including "(1) formulating the problem; (2) designing the study, which includes specification of the independent and dependent variables, the experimental design, the number of replications, and the IRT model; (3) writing or identifying and validating computer programs to generate item responses and to estimate parameters; and (4) analyzing the results" (p. 105). These four steps will be implemented in this dissertation as the (1) introduction, (2) methodology, (3) procedure, and (4) results and discussion sections respectively.

The purpose of the Monte Carlo simulation in this dissertation will be to identify the conditions that lead to an elevated risk of Type I errors for interaction effects in moderated multiple regression from a psychometric approach. This simulation will be structured upon the work of Kang and Waller (2005) who performed a similar study with dichotomous response format assessments, and extend the inquiry to polytomous scales indicative of those commonly used in applied psychological research.

The graded response model (GRM; Samejima, 1969; 1996) was used as the IRT model for deriving the raw scores as well as the estimated theta scores. All of the simulations for this dissertation were conducted in R (Ihaka & Gentleman, 1996; R Development Core Team, 2008). Parameter estimates for the GRM will be derived using PARSCALE 4.1 (Muraki & Bock, 2003).

IRT Model

The Graded Response Model

The GRM is an IRT model suitable for modeling data with ordered categories such as Likert-type scales. A primary assumption underlying the GRM is that the psychological distance between response categories is consistent within items (Embretson & Reise, 2000; Ostini & Nering, 2006). This assumption is usually satisfied by requiring the same response scale for all items on an assessment. A benefit of choosing the GRM to model polytomous data is that the model was developed specifically to represent the processes underlying multi-category decision-making. The GRM is in the difference family of models that were developed specifically to model psychological processes underlying multi-category responding (Ostini & Nering, 2006). Additionally, evidence exists that theta estimates derived using the GRM are able to retain interval level scaling properties (Harwell & Gatti, 2001). These characteristics make the GRM an attractive model for the purposes of this dissertation.

The GRM is considered an "indirect" IRT model because an individual's likelihood of responding in a particular category is derived using a two-step process (Baker & Kim, 2004; Embretson & Reise, 2000; Ostini & Nering, 2006). First, CBRFs are calculated to determine boundary decision probabilities for j-1 categories of each item (Figure 3). This is simply an extension of the two-parameter logistic model where an item response function is calculated for each category boundary including a discrimination, or slope, parameter and a difficulty, or location, parameter. The CBRFs in the GRM can be derived with equation 4.

$$P_{ix}^{*}(\theta) = \frac{e^{[a_i(\theta - b_{ij})]}}{1 + e^{[a_i(\theta - b_{ij})]}} \quad (4)$$

In Equation 4 (adapted from Embretson & Reise, 2000), P_{ix}*(θ) is the probability that an individual with a trait (construct) level θ will respond positively at the boundary of category *j* for item *i* where $x = j = 1 \ldots m_i$. Theta (θ) represents the individual's trait (construct) level, a_i represents the item discrimination or slope, and b_{ij} represents the category location or difficulty parameter with respect to that trait continuum. In this first step of the GRM, Equation 4 would be calculated for the number of CBRFs that exist which is one fewer than the number of categories in the item. The value of a_i will be the same for the CBRFs in a particular item. The values of b_{ij} will vary depending on the particular CBRF being calculated. In well-functioning items, these values should be successive integers reflecting increased difficulty in progressing through the response options. In other words, the values of b_{ij} should indicate that increases in scores on the theta continuum are associated with increased probability of responding positively at higher category boundaries. Violations of this pattern result in what is known as the reversal pattern and indicate that one particular category is never the most likely response for any theta level (Embretson & Reise, 2000).

Equation 4 determines category boundary decision-making, however, one is often more interested in the probability that an individual will respond in a particular category. In the second step of the GRM, the category response functions (CRFs) are derived by

subtracting P_{ix}*(θ) from the following category (Figure 4). This process is illustrated in Equation 5 (adapted from Embretson & Reise, 2000).

$$P_{ix}(\theta) = P_{ix}^{*}(\theta) - P_{i(x+1)}^{*}(\theta) \quad (5)$$

Determining the first category is done by simply subtracting P_{i1}*(θ) from one because the probability of responding in the lowest category or above is equal to 1.0. Determining the last category is simply P_{im}*(θ) because the probability of responding above the final category is equal to 0.0. Each intermediate category is defined by Equation 5 (Embretson & Reise, 2000).

It is important to note that for any value of theta, the CRFs for each response option will sum to 1.0. Additionally, the CBRFs can be thought of as cumulative probabilities representing the probability of responding in or below a particular category. The procedure of determining the individual category probabilities that is described above reflects this condition such that the probability of a response at or above the first category, and at or below the last category is always 100%. If one conceptualizes moving along the theta scale from left to right, the probability of responding at or above a particular category will decrease from 100% to 0, and the probability of responding at or below a particular category will increase from 0 to 100%. The speed at which this change happens is related to the difficulty and discrimination of the item.

Independent Variables

An important *a priori* component of successful Monte Carlo studies involves the specification of the independent and dependent variables (Harwell et al., 1996). The

independent variables are the simulation parameters to be systematically varied. In this dissertation, the independent variables were respondent sample size (*n*: two levels), scale length (*k*: two levels), item discrimination (a_i: two levels), item difficulty ($b_{i,1..j-1}$: three levels), scale bandwidth (fidelity: two levels), and the regression coefficients (β_1 and β_2: two levels). The structure of this dissertation was therefore a 2x2x2x3x2x2 design comprising 96 conditions. For purposes of clarity, the data was simulated and summarized into four tables based on sample size and scale appropriateness (fidelity). Therefore, each simulation included 24 separate conditions (see Figure 6). This allowed for a more parsimonious summarization and interpretation of the results.

Sample Size *(n)*

Two respondent sample sizes were simulated according to recent evidence of the stability of parameter estimates in polytomous IRT and actual sample sizes in MMR studies. Ostini and Nering (2006) have reported that stable estimates for polytomous IRT models can be obtained with as few as 250 individuals, but that samples between 500 and 1,000 are still considered to be desirable. Additionally, Aguinis et al. (2005) indicated that the average sample size for MMR studies in applied psychological research is $\bar{x}_n = 272$ with an average standard deviation of $s_n = 434$. These results indicate that the simulation outcomes for the $n = 250$ sample size will be the most relevant for the majority of applied psychological research, however, some studies do achieve sample sizes upwards of $n = 1{,}000$ (see for example, Witt, 1998; $n = 979$). Therefore, sample size included two levels of $n = 250$ and 750 respondents to maximize the generalizability for

the majority of empirical MMR studies in applied psychology (n=250) as well as for typical IRT studies (n=750).

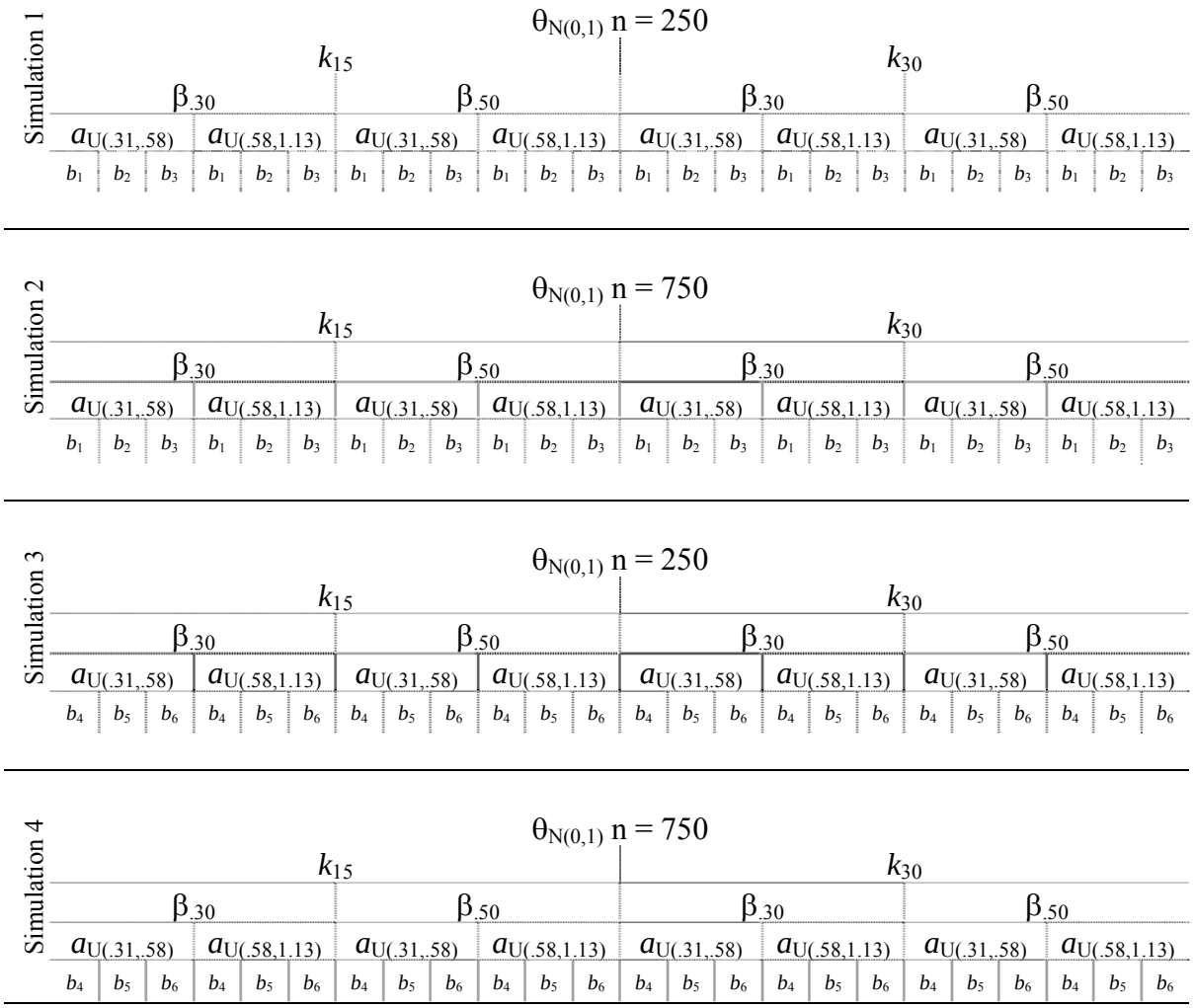

Figure 6. Graphical depiction of the simulation design

Scale Length (k)

In their study of dichotomous assessments, Kang and Waller (2005) utilized two levels of scale length at k=20 and 50 items. These two values were chosen to increase generalizability to common tests utilized in practice as well as to investigate the effects of increasing reliability with increasing assessment length in the classical test theory model. For the purposes of this dissertation, scale length was varied with two levels of k=15 and 30 items. These values served to investigate a similar phenomenon as in the Kang and Waller study, but with values that are more reflective of typical polytomous scales utilized in applied psychological research (Fields, 2002). Specifically, the summary data reported in Table 1 indicate a modal scale length of 15 items with a mean of 15.43 and a standard deviation of 10.43 for validated scales in applied psychology. The distribution

Table 1

*Summary statistics for validated construct assessments in applied psychology**

		Item Descriptive Statistics					Response Option Descriptive Statistics				
Response Category Type	n	min	max	mean (s.d.)	med	mode	min	max	mean (s.d.)	med	mode
Likert-type	122	3	56	15.4 (10.4)	14.0	15	3	9	5.8 (1.3)	5.0	5
Agreement	3	4	72	31.3 (35.9)	18.0	4	3	3	3.0 (0.0)	3.0	3
Mixed	3	3	21	9.3 (10.1)	4.0	3	- -	- -	- -	- -	- -
0-Polar	2	10	25	17.5 (10.6)	17.5	10	5	7	6.0 (1.4)	6.0	5
Semantic Diff.	2	14	21	17.5 (4.9)	17.5	14	5	7	6.0 (1.4)	6.0	5
Q-Sort	1	54	54	54.0 (- -)	54.0	54	- -	- -	- -	- -	- -
True/False	1	7	7	7.0 (- -)	7.0	7	2	2	2.0 (- -)	2.0	2

* The original source for this data is Fields (2002). The summarization and tabulation was conducted by the author of this dissertation.

related to these values is also slightly positively skewed indicating the existence of several very long scales. Therefore, the scale lengths of k=15 and 30 items will serve to satisfy the psychometric interest of the effect of increasing reliability as well as the generalizability of the results to applied psychological research.

Discrimination (a_i)

Discrimination in IRT refers to the extent to which a particular item is able to discriminate levels of the latent construct. In polytomous IRT, item discrimination can be thought of as a type of factor loading (Embretson & Reise, 2000; Takane & De Leeuw, 1987) with values ranging from 0.0 to $+\infty$ where higher values indicate better discrimination. As in other IRT models, item discrimination in the GRM is indicated by the relative steepness of the item curves with steeper curves reflecting higher discrimination. There is one item discrimination parameter for each item, denoted as a_i.

There are two approaches to identifying values for item parameters in simulated IRT studies. One is to set the item parameters to constants based on theoretically or empirically derived values, and the other is to randomly sample values from specified distributions (Harwell, 1996). A primary goal of a Monte Carlo simulation is to derive results that are generalizable for a variety of research applications. In light of this goal, it is advantageous to select item parameter values from specified distributions as opposed to using constant values. However, one drawback from this approach is the possibility of an uncommon combination of item parameters (Harwell et al, 1996), although a sufficient number of replications will adequately control for this problem. For the purposes of the

generalizability of this dissertation, all item parameter values will be randomly selected from specified distributions.

Following the structure of Kang and Waller (2005), item discrimination values were selected from a uniform distribution between the values of 0.31 to 0.58 for moderate discrimination and 0.58 to 1.13 for high discrimination. Estimating discrimination values from a uniform distribution has been demonstrated to appropriately represent empirically determined item discrimination values (Reise & Waller, 2003), and the particular cut-off values of .31, .58, and 1.13 were demonstrated to appropriately represent low, moderate, and high factor loadings for items (Kang & Waller, 2005; Takane & De Leeuw, 1987). Because the GRM is a polytomous extension of the two-parameter logistic model, these values can be deemed appropriate for use in this dissertation. Further, the decision to retain the values from the Kang and Waller (2005) study was made to maintain a basis of comparison for the extension to polytomous data.

Item Difficulty ($b_{i,1...j-1}$)

Item difficulty in IRT refers to the likelihood that an individual respondent would respond positively to a given item based on his or her score on the underlying construct. Binary IRT models have a single item difficulty parameter (b_i) that represents the point along the construct continuum where an individual with a particular θ value has a .5 probability of success on the item (Hambleton et al., 1991). For example, the item response function modeled in Figure 2 would have a difficulty parameter (b_i) approximately equal to 0.0, as this is the point of inflexion for the item response function. However, in polytomous IRT models, item difficulty is expressed as the point along the

construct continuum where an individual with a particular θ value has a .5 probability of responding positively at a category (*j*) boundary. Therefore, polytomous items have *j*-1 difficulty parameters that represent the category boundaries modeled within the item (Embretson & Reise, 2000). In the GRM, this interpretation is a simple extension of the item characteristic curve of the two-parameter logistic model. Specifically, there are *j*-1 item characteristic curves modeled in the GRM that are each centered at their own difficulty value ($b_{i,1..j-1}$) (see Equation 4).

An important aspect of the difficulty parameters in polytomous IRT models is that the difficulty parameter for each operating characteristic curve is sequentially ordered. For example, a polytomous item with five response categories ($j = 5$) will have four category response functions ($j-1$). These four curves will each have their own difficulty parameters ($b_{1,2,3,4}$). Appropriate modeled curves may have values $b_1 = -1.5$, $b_2 = -0.5$, $b_3 = 0.5$, and $b_4 = 1.5$. If the sequential order is not preserved, it would suggest a reversal pattern in which one category is never the most likely category response for any value of θ (Embretson & Reise, 2000). Therefore, difficulty parameters in this dissertation were modeled with the sequential ordering restriction imposed.

The characteristic that dichotomous and polytomous items share is that their difficulty parameters represent the point along the construct continuum at which the item is centered in difficulty. More difficult items are centered further to the right of the construct continuum, which is centered at zero with a standard deviation of one, and less difficult items are centered left of zero. The overall interpretation for the difficulty of a

polytomous assessment is the level of the underlying construct required to respond to higher categories of any particular item.

In accordance with Kang and Waller (2005), item difficulty values in this dissertation were simulated at three levels. The three levels will represent an "easy", "moderate", and "difficult" assessment by randomly selecting difficulty values from a normal distribution with a mean and a standard deviation of N(-1.50, 1.00), N(0.00, 1.00), and N(1.50, 1.00) respectively. It is important to note that in polytomous models such as the GRM, there are *j*-1 difficulty values for each item representing the relative difficulty of responding positively at a category boundary. Therefore, for this dissertation, there will be four difficulty values selected for each item. To achieve the appropriate ordering, the first difficulty value for the CBRF between options one and two will be randomly selected from a N(-2.5, 0.7), N(-1.0, 0.7), or N(0.5, 0.7) distribution for the easy, moderate, and difficult assessments respectively. In a similar approach as Meade, Lautenschlager, and Johnson (2007), a constant of 0.7 will then be added for the three subsequent difficulty parameters. The resulting four difficulty values will reflect random selection from approximately normal distributions centered at -1.5, 0, and 1.5 for the three levels of simulated difficulty. As with the discrimination parameters, these parameter values were chosen to maintain comparison with Kang and Waller (2005).

Finally, it should be noted that the meaning of the item difficulty values in the simulated assessments reflects the amount of the underlying construct that is required to respond in higher categories for each item. This also translates to the probability of receiving higher total scores on the assessment. However, the difficulty in this case does

not refer to the probability of answering an item correctly or incorrectly, although this meaning can exist in instances of partial credit models (Ostini & Nering, 2006). Therefore, in this study, item difficulty should be thought of as the overall probability of agreement with a particular set of items given an individual's theta score.

Scale Fidelity

An assessment's fidelity is measured as the inverse of variability, or bandwidth, in the difficulty of the items (Stocking, 1987). High fidelity assessments sacrifice bandwidth such that there is less variability in the item difficulty values. Fidelity contributes to assessment appropriateness by either restricting (high fidelity) or expanding (low fidelity) the width of the item difficulty distribution. A situation where this would be useful is with an assessment meant to provide the highest precision of information at a narrow score range such as those developed at or around a cut-off score (Hambleton et al., 1991). However, for the purposes of general construct assessment, a scale with high fidelity and low bandwidth may be in danger of being narrowly appropriate for particular groups of respondents. Assessments with these characteristics that are used for general construct measurement may be the most at-risk assessments for spurious interaction effects. This condition was simulated in this study by generating a second set of item difficulty values from more restricted normal distributions with a mean and standard deviation of N(-1.50, 0.50) for easy scales, N(0.00, 0.50) for moderate scales, and N(1.50, 0.50) for difficult scales. These restricted distributions will create the high fidelity and low bandwidth situation in which Kang and Waller (2005) observed the highest prevalence of Type I errors. As in the previous difficulty parameter selection, there were four (j-1) difficulty

values for each item sampled from within the specified distribution with the sequential ordering restriction imposed. To accomplish this in the high fidelity conditions, the first difficulty value for the CBRF between options one and two was randomly selected from a N(-2.0, 0.35), N(-0.5, 0.35), or N(1.0, 0.35) distribution for the easy, moderate, and difficult assessments respectively. A constant of 0.35 was be added for the three subsequent difficulty parameters. The resulting four difficulty values will reflect random selection from approximately normal distributions centered at -1.5, 0, and 1.5 for the three levels of simulated difficulty.

Regression Weights

In accordance with Kang and Waller (2005), regression weights were set at a value of 0.30 or 0.50 for both β_1 and β_2. An intercept of zero is used and therefore omitted from the regression models. It should be noted that these regression weights are fixed only for the purposes of simulating the dependent variables.

Fixed Effects

Item Response *Categories (j)*

The number of item response categories for this study was set at five to simulate a five-category Likert-type response scale. Fields (2002) identified 134 validated construct assessments that are utilized in applied psychological research of which five category Likert-type response scales were the most common (n = 57). Full descriptive statistics summarizing these 134 assessments can be found in Table 1.

Regression Models

The purpose of this dissertation was to observe the prevalence of Type I errors in moderated multiple regression in three different pairs of models. In the first regression model pair, actual latent trait scores θ will be analyzed (see Equations 6a and 6b). In the second regression model pair, raw scores (X) will be analyzed (see Equations 7a and 7b). In the third regression model pair, estimated theta scores $\hat{\theta}$ will be analyzed (see Equations 8a and 8b). These three model pairs will be expressed in accordance with Kang and Waller (2005) as follows:

$$\theta_3 = \beta_1\theta_1 + \beta_2\theta_2 + \varepsilon \tag{6a}$$

$$\theta_3 = \beta_1\theta_1 + \beta_2\theta_2 + \beta_3\theta_1\theta_2 + \varepsilon \tag{6b}$$

$$X_3 = \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{7a}$$

$$X_3 = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon \tag{7b}$$

$$\hat{\theta}_3 = \beta_1\hat{\theta}_1 + \beta_2\hat{\theta}_2 + \varepsilon \tag{8a}$$

$$\hat{\theta}_3 = \beta_1\hat{\theta}_1 + \beta_2\hat{\theta}_2 + \beta_3\hat{\theta}_1\hat{\theta}_2 + \varepsilon \tag{8b}$$

The first model of each pair is the additive model, and the second model in each pair contains a multiplicative, or interaction, term. As is typical in moderated multiple regression, each model pair will be structured as a hierarchical regression analysis where the interaction term is entered at the second step (Aiken & West, 1991; Cohen et al., 2003). A significant change in variance accounted for (ΔR^2) between the first and second model will indicate the existence of a spurious interaction effect.

Regression Main Effects

Two continuous predictor variables were simulated for each regression model specified in Equations 6 through 8. Predictor variables θ_1 and θ_2 were randomly selected for the number of observations (n) from normal distributions with a mean and standard deviation equal to N(0.00, 1.00). These variables served as the main effect scores in the regression models. It is important to note that θ_1 and θ_2 were sampled from identical but independent distributions, thus there was no correlation between the between the predictor variables. As such, no multicollinearity was modeled.

Regression Criterion Variables

One continuous criterion variable was calculated for each regression model specified in Equations 6 through 8. In accordance with Kang and Waller (2005), the general form of the criterion variables is given by the following equation, which represents a multiple regression model with two significant main effects and no interaction.

$$\theta_3 = \beta_1\theta_1 + \beta_2\theta_2 + \sqrt{1-(\beta_1^2+\beta_2^2)} \times \varepsilon \qquad (9)$$

In Equation 9, β_1 and β_2 are the simulated regression weights and ε is an error term. Note that the intercept term, β_0, was set to equal zero and thus omitted from the model. The term $\sqrt{1-\left(\beta_1^2+\beta_2^2\right)}$ was included to represent an appropriate error variance component for each level of β. This term can be derived in the following manner. First, the predictor variables θ_1 and θ_2 and the criterion variable θ_3 are normally distributed with a mean and standard deviation equal to N(0.00, 1.00). Because the standard deviation is simply the square root of the variance, the variance of the predictor and criterion variables is equal to one. Given these conditions, the following derivation gives the error term for the regression models.

$$\sigma_{\theta_3}^2 = \beta_1^2\sigma_{\theta_1}^2 + \beta_1^2\sigma_{\theta_2}^2 + \sigma_e^2 \tag{10a}$$

where $\sigma_{\theta_1}^2$, $\sigma_{\theta_2}^2$, and $\sigma_{\theta_3}^2$ are all equal to 1.00, therefore,

$$1 = \beta_1^2 + \beta_1^2 + \sigma_e^2 \tag{10b}$$

$$1 - \beta_1^2 - \beta_1^2 = \sigma_e^2 \tag{10c}$$

$$1 - \left(\beta_1^2 + \beta_1^2\right) = \sigma_e^2 \tag{10d}$$

$$\sqrt{1-\left(\beta_1^2+\beta_1^2\right)}=\sigma_e \tag{10e}$$

Finally, given that the two levels of regression weights β_1 and β_2 were simulated as .3 and .5, equation 9 can be further reduced to the following form.

$$\theta_3 = .5(\theta_1)+ .5(\theta_2)+ .7071(\varepsilon) \tag{11a}$$

$$\theta_3 = .3(\theta_1)+ .3(\theta_2)+ .9055(\varepsilon) \tag{11b}$$

In the simulation, the criterion variable θ_3 was operationalized with equations 11a and 11b for the two levels of the regression weights, .3 and .5, respectively. An alternative way of specifying this derivation would be to say that the error associated with each regression model is being sampled from a N(0.00, .7071) and N(0.00, .9055) distribution for each level of β.

Raw Scores

To generate the raw scores, X_1, X_2, and X_3, the values of the previously defined construct scores θ_1, θ_2, and θ_3 were entered into the GRM equation (Equations 4 and 5) for each simulated participant. It is pertinent to note that because IRT is a strong-modeling methodology, the responses are being simulated to fit the GRM. This is important because good model-data fit is a key assumption to be satisfied when drawing results based on IRT models (Embretson & Reise, 2000; Hambleton et al., 1990).

A matrix of response scores was generated by reporting the raw score (1, 2, 3, 4, or 5) corresponding to the highest category response likelihood for each simulated participant on each item. These values were derived using an algorithm written by the author based on response probabilities calculated in Equations 4 and 5. Actual raw score responses were generated by comparing a randomly selected value from a uniform distribution, U(0.0, 1.0), with the relative response probabilities that are generated for each level of theta (or individual) and each item. This process can be thought of as determining the relative likelihood of a category response given the item and person parameters with a realistic level of decision-making error (Kang & Waller, 2005; Stone, 1992). This integration of response error is important so as to not assume perfect responding by simulated individuals. A mean score for X_1, X_2, and X_3 for each simulated individual was calculated from the raw score response matrices for analysis in the regression models.

Estimated Theta Scores

Finally, the estimated theta scores $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ were simulated using PARSCALE 4.1 (Muraki & Bock, 2003). PARSCALE was set to derive the person (latent construct scores) and item parameters using the expected a posteriori (EAP) method and Bayesian priors. This method calculates $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ as the modal value of the posterior distribution which is the most likely value of theta for the observed response pattern (Baker & Kim, 2004), and is a preferred estimation method for assessments that are moderate to short in item length (Mislevy & Stocking, 1989). The syntax files for the PARSCALE integration can be found in Appendix F.

Iterations

In a Monte Carlo study, the number of iterations, or replications, that the simulation conducts is akin to sample size in a traditional, empirical study (Gagne, Furlow, & Ross, 2009; Harwell et al., 1996; Mooney, 2007). In IRT Monte Carlo studies, item and person parameters can be randomly selected from specified distributions for the purposes of generalizability. This sampling leads to the concern of sampling variance with regards to the parameters being estimated. A Monte Carlo study conducted with one or very few iterations is at serious risk for producing biased estimates, however, "aggregating results over replications produces more stable and reliable results" (Harwell et al., 1996, p. 110). Indeed, Harwell et al. (1996) indicated that increasing the number of iterations in IRT Monte Carlo studies will be a significant step forward and will allow for the testing of more complex and informative designs.

As an analogue for sample size, the number of iterations that are used in a Monte Carlo study is directly related to statistical power. Harwell et al. (1996) indicate that well developed Monte Carlo studies will use several hundred to several thousand iterations per condition, and that studies that exceed several hundred iterations have power approaching or equal to 1.0. For the purposes of estimating Type I error rates in Monte Carlo studies, Robey and Barcikowski (1992) specify that approximately 1,000 iterations will achieve a power equal to .90 when approximating an alpha level of $\alpha = .05$ and using the interval of $\alpha \pm \frac{1}{2}\alpha$ as a robustness interval. Therefore, 1,000 iterations per condition were conducted[2]. This allowed for adequate reduction in sampling variance for the IRT

[2] With 1,000 iterations per condition in a 2x2x2x3x2x2 study, 96,000 cases will be generated for aggregation into the final results.

parameter estimates (Harwell et al., 1996), achieves a power of .90 around the interval $.025 \leq \alpha \leq .075$ (Robey & Barcikowski, 1992), and doubles the number of iterations utilized by Kang and Waller (2005).

Simulation Dependent Variables

Type I Errors

The primary dependent variable for this study was the empirical Type I error rate (π) that is observed for the interaction term of the moderated multiple regression models. The specific value of π was identified in a three-step process. In each iteration of the simulation, the variance in θ_3 accounted for by θ_1 and θ_2 was recorded as the R^2 value for the additive and multiplicative regression models specified in Equations 6 through 8. Second, the significance of the change in variance accounted for, ΔR^2, between the respective additive and multiplicative models was tested at an alpha level of $p \leq .05$ and recorded as 1 for a significant result and 0 for a non-significant result. Finally, the empirical alpha level π was recorded as the proportion $\left(\frac{x}{1{,}000}\right)$ of iterations resulting in a significant ΔR^2 for the actual latent trait scores θ_3, the raw scores X_3, and the estimated theta scores $\hat{\theta}_3$.

Procedure

The simulation for the current study was conducted in the R environment using a mixture of functions written by the author of this dissertation, as well as one external program. R (Ihaka & Gentleman, 1996; R Core Development Team, 2008) is an open-source platform for statistical computing that allows a high degree of flexibility in the

design and implementation of a variety of statistical techniques. R is also particularly equipped for handling programmed loops and pseudo-random number generation as is required by Monte Carlo studies. An external program, PARSCALE 4.1 (Muraki & Bock, 2003), was employed as the estimation engine to derive item and person parameters based on the GRM. All of these programs were integrated into self-contained loops in the R environment for the purposes of automated execution of the Monte Carlo simulation.

For purposes of ease of interpretation, four separate simulations were conducted. The syntax for these simulations can be found in Appendices A, B, C, and D. The four simulations were separated based on sample size (n=250, 750) and scale fidelity (normal, high). In each simulation, the independent variables of scale length, regression weights, discrimination, and difficulty will be systematically varied. Therefore, the summary statistics for each simulation will be included in four tables, each with 24 rows.

Each simulation was run in the following process. First, using the pseudo-random number generator in R, theta vectors were estimated from a standard normal distribution N(0.0, 1.0) for θ_1 and θ_2. Next, corresponding vectors for θ_3 were calculated using Equation 9. These vectors were saved as the actual latent construct scores. To calculate the raw score matrices, X_1, X_2, and X_3, each of these three score vectors were evaluated in an algorithm written by the author (see lines 21-59 & 481-519 of Appendix A, B, C, & D) that implements Equation 4 and Equation 5 to determine the probability of a category response. Final raw score values were determined by the comparison of a randomly selected value from a uniform distribution as previously described. Finally, the estimated

theta scores $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ were derived using PARSCALE 4.1 (Muraki & Bock, 2003). To accomplish this task, raw scores matrices were "batched" out to PARSCALE with an accompanying syntax file (see Appendix E and Appendix F) following the structure identified by Gagne, Furlow, and Ross (2009). The estimated theta scores that are returned by PARSCALE were then returned to R as the vectors $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ (see lines 161-191 & 621-651 of Appendix A, B, C, & D).

The reliability of each simulated assessment was calculated (see lines 152-159 & 612-619 of Appendix A, B, C, & D) using the Cronbach's alpha function of the Latent Trait Modeling package in R (Rizopoulos, 2006). Other summary statistics such as the skewness and kurtosis of each simulated assessment were calculated as well as the Shapiro-Wilk test for normality (see lines 201-213, 229-237 & 661-673, 689-697 of Appendix A, B, C, & D). Finally, the nine score vectors to be entered into the corresponding additive and multiplicative regression models specified in Equations 6 through 8 (see lines 215-227 & 675-687 of Appendix A, B, C, & D), and the change in variance accounted for between the two corresponding models was recorded.

This procedure was repeated for each condition in the study over the total number of iterations (see lines 307-471 & 767-931 of Appendix A, B, C, & D). The final summary statistics and tables were generated in the R environment using portions of code provided, with permission, by Niels Waller and used in the Kang and Waller (2005) study (see lines 933-1,083 of Appendix A, B, C, & D). The simulation was conducted on an Apple iMac with a 2.80 GHz Intel Core2 Duo processor, 3 gigabytes of RAM, and

running Microsoft Windows XP as a native operating system on a harddrive partition. The version of R that was used was R 2.9.0.

Verification

An important aspect of any Monte Carlo study is to verify the values that are being calculated for the primary purposes of the simulation (Harwell et al., 1996). Portions of the code used in this dissertation were modeled off of the code used in the Kang and Waller (2005) study. However, the Kang and Waller code was missing important details, such as the IRT model estimation engine. Therefore, a pilot study was undertaken to replicate and verify the structure of the Kang and Waller simulation code. To replace the missing code components, the "irtoys" package (Partchev, 2007) for R was used to estimate the latent construct scores for dichotomous data with the two-parameter logistic model. The overall structure of the code for this simulation was verified by a complete replication of the Kang and Waller (2005) study. Table 2 contains the results of this replication. All of the values in Table 2 were calculated by the author of this dissertation. Some divergence in the results of the estimated theta scores was expected due to the differences in the estimation engines, however, the divergence was minimal and never exceeded 0.1 for the prevalence of Type I errors based on estimated latent construct scoring. In most cases, the divergence between the two studies did not exceed .01 – .02 in Type I error rates. Other scoring methods such as actual latent construct scores and raw scores, as well as measures of reliability for the raw score matrices were replicated to an exact value at two decimal points. This evidence suggests

that the overall structure of the simulation code that will be used in this dissertation is reliable.

An additional question of verification comes with the implementation of the PARSCALE program to generate parameter estimates. In this dissertation, the estimated theta scores were of primary interest. As a commercially published software package, one can often rely on the parameter estimation procedures implemented in programs such as PARSCALE. However, to satisfy the skeptic, DeMars (2002) and Childs and Chen (1999) both conducted parameter recovery studies for the GRM in PARSCALE. In each

Table 2

*Kang and Waller (2005) Table 1 Replication**

c	b_i	a_i	β	k	π_θ	π_x	$\pi_{\hat{\theta}}$	$\bar{p}$ value	KR20	SW_θ	SW_x	$SW_{\hat{\theta}}$	sk_x	kt_x	$sk_{\hat{\theta}}$	$kt_{\hat{\theta}}$
1	Difficult	Low	.3	20	0.06	0.06	0.05	0.71(.13)	0.65	0.96	0.06	0.14	0.57	0.02	0.45	0.17
2	Difficult	Low	.3	50	0.06	0.07	0.05	0.71(.13)	0.83	0.96	0.06	0.70	0.60	0.08	0.27	0.23
3	Difficult	Low	.5	20	0.06	0.09	0.07	0.71(.13)	0.65	0.96	0.20	0.40	0.56	0.01	0.44	0.18
4	Difficult	Low	.5	50	0.06	0.13	0.07	0.71(.13)	0.83	0.96	0.27	0.86	0.60	0.08	0.27	0.22
5	Difficult	High	.3	20	0.06	0.10	0.06	0.79(.16)	0.81	0.96	0.00	0.02	1.07	0.92	0.69	0.01
6	Difficult	High	.3	50	0.06	0.11	0.06	0.79(.16)	0.92	0.96	0.00	0.16	1.09	0.93	0.46	0.25
7	Difficult	High	.5	20	0.06	0.27	0.17	0.79(.16)	0.81	0.96	0.02	0.21	1.07	0.91	0.68	0.02
8	Difficult	High	.5	50	0.06	0.39	0.14	0.79(.16)	0.92	0.96	0.01	0.62	1.09	0.93	0.46	0.26
9	Moderate	Low	.3	20	0.06	0.04	0.04	0.50(.15)	0.69	0.96	0.79	0.83	0.00	0.49	0.00	0.45
10	Moderate	Low	.3	50	0.06	0.04	0.05	0.50(.16)	0.85	0.96	0.84	0.95	0.00	0.51	0.00	0.29
11	Moderate	Low	.5	20	0.06	0.04	0.04	0.50(.15)	0.69	0.96	0.89	0.91	0.00	0.49	0.00	0.45
12	Moderate	Low	.5	50	0.06	0.04	0.05	0.50(.16)	0.85	0.96	0.93	0.96	0.00	0.51	0.00	0.29
13	Moderate	High	.3	20	0.06	0.04	0.04	0.50(.22)	0.85	0.96	0.36	0.73	0.00	0.79	0.00	0.61
14	Moderate	High	.3	50	0.06	0.04	0.05	0.50(.22)	0.94	0.96	0.40	0.90	0.00	0.82	0.00	0.43
15	Moderate	High	.5	20	0.06	0.02	0.03	0.50(.22)	0.85	0.96	0.82	0.90	0.00	0.79	0.00	0.61
16	Moderate	High	.5	50	0.06	0.03	0.03	0.50(.22)	0.94	0.96	0.88	0.95	0.00	0.82	0.00	0.44
17	Easy	Low	.3	20	0.06	0.05	0.04	0.29(.13)	0.66	0.97	0.06	0.16	0.57	0.06	0.45	0.14
18	Easy	Low	.3	50	0.06	0.05	0.04	0.29(.13)	0.83	0.97	0.04	0.68	0.61	0.11	0.28	0.21
19	Easy	Low	.5	20	0.06	0.07	0.06	0.29(.13)	0.65	0.97	0.17	0.39	0.57	0.05	0.45	0.15
20	Easy	Low	.5	50	0.06	0.15	0.07	0.29(.13)	0.83	0.97	0.24	0.85	0.61	0.11	0.28	0.21
21	Easy	High	.3	20	0.06	0.11	0.09	0.21(.16)	0.81	0.97	0.00	0.02	1.07	0.90	0.68	0.02
22	Easy	High	.3	50	0.06	0.12	0.06	0.21(.16)	0.92	0.97	0.00	0.17	1.10	0.95	0.46	0.24
23	Easy	High	.5	20	0.06	0.31	0.20	0.21(.16)	0.81	0.97	0.01	0.19	1.07	0.92	0.68	0.01
24	Easy	High	.5	50	0.06	0.41	0.15	0.21(.16)	0.92	0.97	0.01	0.61	1.10	0.97	0.47	0.23

*All values in this table were calculated by the author of this dissertation; Iterations per condition = 500

Table Key: c = condition; b_i = item difficulty distribution, Difficult = N(-1.5,1), Moderate = N(0,1), Easy = N(1.5,1); a_i = item discrimination distribution, Low = U(.31,.58), High = U(.58,1.13); β = regression weight; k = number of items; π_θ = empirical Type I error rate for actual theta scores; π_x = empirical Type I error rate for raw scores; $\pi_{\hat{\theta}}$ = empirical Type I error rate for estimated theta scores; $\bar{p}$ value = item difficulty mean and standard deviation; KR20 = average internal consistency for the simulated raw scores; SW_θ = proportion of n.s. Shapiro-Wilk tests for the actual theta scores; SW_x = proportion of n.s. Shapiro-Wilk tests for the raw scores; $SW_{\hat{\theta}}$ = proportion of n.s. Shapiro-Wilk tests for the estimated theta scores; sk_x = skewness for the raw scores; kt_x = kurtosis for the raw scores; $sk_{\hat{\theta}}$ = skewness for the estimated theta scores; $kt_{\hat{\theta}}$ = kurtosis for the estimated theta scores

study, the results indicated that the parameter estimates for the GRM in PARSCALE were unbiased even when estimated from initially skewed distributions, and the root mean square error between the estimated and actual latent construct scores were low (all below .70 for the GRM), and comparable with another validated software package, MULTILOG. Additionally, Childs and Chen (1999) found parameter estimates between the two programs to be correlated at approximately .99. Given these results, the scores generated by PARSCALE can be considered accurate estimates of the latent construct being assessed.

Finally, a verification of the simulated raw scores is in order. Recall that the raw scores are being calculated using an algorithm derived by the author of this dissertation that is modeled from the GRM equations and an appropriately modeled response error term. An appropriate method of verification could be an examination of internal consistency, or Cronbach's alpha, for the scores. In his eloquent treatment of the meaning of alpha, Cortina (1993) indicates that larger alpha values reflect variance that is attributable to general factors and not item-specific variance. As such, one would expect item-specific variance to be high in randomly assembled response matrices with no underlying structure. However, if the simulated raw scores are indeed reflecting a latent construct and the specified model properties, acceptable alpha levels should be observed. Results of a single iteration simulation of four different conditions demonstrated that acceptable alpha levels were achieved with the raw score algorithm (see Table 3).

The preceding evidence lends strong support to the structure and content of the simulation that will be employed in this study. In verifying the nature of the programs and algorithms used for this simulation, specifying the experimental design and

Table 3

Results of a verification test for the simulated raw scores

Number of Individuals	Number of Items	a	b	theta	alpha
250	15	U(0.31, 0.58)	N(0.0, 1.0)	N(0.0, 1.0)	.729
250	30	U(0.31, 0.58)	N(0.0, 1.0)	N(0.0, 1.0)	.819
750	15	U(0.31, 0.58)	N(0.0, 1.0)	N(0.0, 1.0)	.719
750	30	U(0.31, 0.58)	N(0.0, 1.0)	N(0.0, 1.0)	.814

procedure, and conducting 1,000 iterations in each condition, this dissertation satisfied all of the aspects of a high quality IRT Monte Carlo study identified by Harwell et al. (2002). Self-contained, reproducible code for this simulation is available in Appendices A, B, C, D, E, and F (note that the PARSCALE program is required for this code to execute successfully).

Data Analysis Strategy

Identifying Meaningful Type I Error Rate Inflations

The central focus of this dissertation is to determine what psychometric conditions lead to meaningfully inflated Type I error rates for different scoring techniques for a latent construct. Therefore, it will be necessary to specify a criterion to determine whether the empirical Type I error rates are a reasonable approximation of the nominal Type I error rate, or alpha, of .05.

Two criteria were used to determine whether the empirical Type I error rates were meaningfully inflated above the nominal alpha level of .05. First, using the criteria specified by Bradley (1978) and Robey and Barcikowski (1992), values of π within the interval of $.025 \leq \alpha \leq .075$ were identified as reasonable approximations of α. Values of π larger than .075 were identified as significant departures from α and therefore spurious interaction effects. The aforementioned interval was chosen based on Bradley's (1978) liberal interval of $\alpha \pm \frac{1}{2}\alpha$ for a Type I error rate of .05, and Robey and Barcikowski's (1992) values for achieving a power of .90 with 1,000 iterations in a Monte Carlo study with α and π both set to .05.

Second, a binomial test based on a z-approximation distribution was conducted for each condition. The binomial test is a non-parametric procedure for determining whether the observed proportions of some event are equal to a known binomial distribution. An expected proportion of .05 was specified, and significant departures from this proportion were noted.

Identifying Effects of the Independent Variables

To determine which independent variables had the greatest impact on the occurrence of a spurious interaction effect, a direct logistic regression analysis was conducted for the raw and estimated theta scores. The independent variables were entered simultaneously into the model so that the impact of each variable could be evaluated in the presence the other variables. This analysis was conducted at the iteration level, and the dependent variable was a dichotomous 1,0 variable indicating the presence or absence of a significant interaction effect respectively.

Testing Hypothesized Interactions

Hypotheses 3 and 4 specified interactions between assessment appropriateness and fidelity and item discrimination and beta weights respectively. These hypotheses were derived from observed results of the Kang and Waller (2005) study. Keeping in context with Kang and Waller, hypotheses 3 and 4 were tested at the aggregate level using 2x2 factorial ANOVA analyses. The dependent variable in these analyses was the empirical Type I error rate for the raw scores.

As an additional examination, these hypotheses were also tested at the iteration level using a stepwise logistic regression model. The independent variables were entered simultaneously at the first step, followed by the interaction term at the second step. The dependent variable in these analyses was a dichotomous 1,0 variable indicating the presence or absence of a significant interaction effect respectively.

Assessing Linearity and Interval-Level Scaling

Finally, the linearity and interval-level scaling of the dependent variables was assessed using a Pearson product-moment correlation and an examination of the scatter plots for the relationship between actual theta scores, raw scores, and estimated theta scores. Conditions 80, 88, and 96 were chosen from the simulation for the purposes of this analysis. Due to the large number of data points (750,000) generated within each of these conditions, a random sample of 1,875, or .25%, were selected from each to generate the scatter plots. The correlations were calculated based on the full sample.

CHAPTER 4: RESULTS

Tables 4, 5, 6, and 7 contain the main results of the four simulations that were conducted. Each table contains 24 conditions (for a total of 96 conditions) represented in rows that vary based on sample size, scale length, item discrimination, item category difficulty, and regression (beta) weights. The simulated theta distribution, which represents the individuals' actual latent construct scores, for all of the conditions was standard normal N(0, 1).

The tables are organized into pairs based on scale fidelity. Tables 4 and 5 represent the normal fidelity conditions, in which the distributions of the difficulty parameters had a standard deviation of 1.0. Tables 6 and 7 represent the high fidelity conditions in which the distributions for the difficulty parameters had a standard deviation of 0.5. This distinction represents the fidelity manipulation, which was hypothesized to create conditions of extreme assessment inappropriateness. Within each fidelity pair, the tables are distinguished by sample size. Specifically, tables 4 and 6 represent simulated samples of 250 individuals, and tables 5 and 7 represent simulated samples of 750 individuals. Each row in the four tables represents a single condition in the simulation, and the individual row entries represent averaged results across the 1,000 iterations in each condition.

In each table, the independent variables that were manipulated in the simulation are represented in the columns for sample size (n), item category difficulty ($b_{i,j-1}$), item discrimination (a_i), beta weights (β), and the number of items (k). Additionally, it is helpful to note that each table can be subdivided into three primary sections based on

Table 4

Results of simulation 1 (normal fidelity, distribution of latent construct scores = standard normal N(0,1))

c	n	$b_{i,j-1}$	a_i	β	k	π_θ	π_x	$\pi_{\hat{\theta}}$	α	*rmse*	SW_θ	SW_x	$SW_{\hat{\theta}}$	sk_x	kt_x	$sk_{\hat{\theta}}$	$kt_{\hat{\theta}}$
1	250	Difficult	Low	.3	15	0.062	0.055	0.047	0.66	0.86	0.96	0.06	0.78	0.54	0.03	0.23	0.12
2	250	Difficult	Low	.3	30	0.062	0.068†	0.052	0.80	0.84	0.96	0.06	0.77	0.56	0.02	0.22	0.14
3	250	Difficult	Low	.5	15	0.062	0.088*,†	0.058	0.66	0.86	0.96	0.23	0.84	0.54	0.03	0.24	0.08
4	250	Difficult	Low	.5	30	0.062	0.113*,†	0.061	0.80	0.83	0.96	0.30	0.82	0.57	0.02	0.24	0.10
5	250	Difficult	High	.3	15	0.062	0.055	0.047	0.66	1.44	0.96	0.06	0.78	0.54	0.03	0.23	0.12
6	250	Difficult	High	.3	30	0.061	0.105*,†	0.082*,†	0.92	1.44	0.96	0.00	0.41	1.01	0.68	0.34	0.13
7	250	Difficult	High	.5	15	0.061	0.296*,†	0.089*,†	0.86	1.44	0.96	0.01	0.57	0.99	0.64	0.34	0.11
8	250	Difficult	High	.5	30	0.062	0.372*,†	0.098*,†	0.92	1.44	0.96	0.02	0.61	1.01	0.68	0.34	0.15
9	250	Moderate	Low	.3	15	0.062	0.046	0.056	0.69	0.70	0.96	0.83	0.86	0.02	0.51	0.01	0.30
10	250	Moderate	Low	.3	30	0.062	0.042	0.048	0.82	0.70	0.96	0.86	0.86	0.02	0.50	0.02	0.28
11	250	Moderate	Low	.5	15	0.062	0.041	0.052	0.69	0.70	0.96	0.91	0.86	0.02	0.51	0.02	0.28
12	250	Moderate	Low	.5	30	0.062	0.037	0.037	0.82	0.70	0.96	0.93	0.86	0.02	0.50	0.03	0.26
13	250	Moderate	High	.3	15	0.062	0.051	0.047	0.88	0.53	0.96	0.43	0.88	0.04	0.78	0.01	0.22
14	250	Moderate	High	.3	30	0.062	0.048	0.059	0.94	0.52	0.96	0.50	0.88	0.03	0.79	0.00	0.25
15	250	Moderate	High	.5	15	0.062	0.041	0.050	0.88	0.53	0.96	0.86	0.91	0.04	0.78	0.00	0.25
16	250	Moderate	High	.5	30	0.061	0.039	0.055	0.94	0.52	0.96	0.89	0.91	0.03	0.79	0.00	0.24
17	250	Easy	Low	.3	15	0.061	0.064†	0.053	0.65	0.89	0.96	0.03	0.72	0.58	0.02	0.21	0.19
18	250	Easy	Low	.3	30	0.061	0.067†	0.052	0.79	0.88	0.96	0.04	0.70	0.60	0.08	0.22	0.20
19	250	Easy	Low	.5	15	0.061	0.106*,†	0.075†	0.65	0.89	0.96	0.15	0.79	0.58	0.03	0.21	0.21
20	250	Easy	Low	.5	30	0.061	0.138*,†	0.055	0.79	0.88	0.96	0.23	0.78	0.60	0.08	0.20	0.21
21	250	Easy	High	.3	15	0.061	0.112*,†	0.098*,†	0.85	1.58	0.96	0.00	0.30	1.06	0.86	0.37	0.10
22	250	Easy	High	.3	30	0.061	0.123*,†	0.070†	0.92	1.59	0.96	0.00	0.26	1.08	0.88	0.38	0.08
23	250	Easy	High	.5	15	0.061	0.317*,†	0.120*,†	0.85	1.59	0.96	0.00	0.49	1.06	0.88	0.38	0.09
24	250	Easy	High	.5	30	0.061	0.386*,†	0.102*,†	0.92	1.59	0.96	0.01	0.47	1.08	0.90	0.38	0.08

* Significant Type I Error rate based on $\alpha \pm \frac{1}{2}\alpha$; †Significant Type I Error rate based on the results of a binomial test; Iterations per condition = 1,000

Table Key: c = condition; n = number of individuals; $b_{i,j-1}$ = item category difficulty distribution, Difficult = N(-1.5,1), Moderate = N(0,1), Easy = N(1.5,1); a_i = item discrimination distribution, Low = U(.31,.58), High = U(.58,1.13); β = regression weight; k = number of items; π_θ = empirical Type I error rate for actual theta scores; π_x = empirical Type I error rate for raw scores; $\pi_{\hat{\theta}}$ = empirical Type I error rate for estimated theta scores; α = average internal consistency for the raw scores; *rmse* = root mean square error for the estimated theta scores; SW_θ = proportion of n.s. Shapiro-Wilk tests for the actual theta scores; SW_x = proportion of n.s. Shapiro-Wilk tests for the raw scores; $SW_{\hat{\theta}}$ = proportion of n.s. Shapiro-Wilk tests for the estimated theta scores; sk_x = |skewness| for the raw scores (abs. value); kt_x = |kurtosis| for the raw scores; $sk_{\hat{\theta}}$ = |skewness| for the estimated theta scores; $kt_{\hat{\theta}}$ = |kurtosis| for the estimated theta scores

Table 5

Results of simulation 2 (normal fidelity, distribution of latent construct scores = standard normal N(0,1))

c	n	$b_{i,j-1}$	a_i	β	k	π_θ	π_x	$\pi_{\hat{\theta}}$	α	$rmse$	SW_θ	SW_x	$SW_{\hat{\theta}}$	sk_x	kt_x	$sk_{\hat{\theta}}$	$kt_{\hat{\theta}}$
25	750	Difficult	Low	.3	15	0.049	0.074†	0.056	0.66	0.82	0.95	0.00	0.44	0.55	0.01	0.23	0.12
26	750	Difficult	Low	.3	30	0.049	0.069†	0.057	0.80	0.84	0.95	0.00	0.45	0.57	0.02	0.24	0.13
27	750	Difficult	Low	.5	15	0.049	0.167*,†	0.089*,†	0.66	0.82	0.95	0.00	0.65	0.55	0.01	0.22	0.16
28	750	Difficult	Low	.5	30	0.049	0.222*,†	0.079*,†	0.80	0.84	0.95	0.01	0.66	0.56	0.01	0.24	0.12
29	750	Difficult	High	.3	15	0.049	0.162*,†	0.084*,†	0.86	1.31	0.95	0.00	0.11	1.00	0.66	0.31	0.20
30	750	Difficult	High	.3	30	0.049	0.142*,†	0.066†	0.92	1.31	0.95	0.00	0.11	1.01	0.67	0.32	0.19
31	750	Difficult	High	.5	15	0.049	0.627*,†	0.173*,†	0.86	1.31	0.95	0.00	0.52	1.00	0.66	0.31	0.19
32	750	Difficult	High	.5	30	0.049	0.710*,†	0.158*,†	0.92	1.31	0.95	0.00	0.52	1.00	0.67	0.32	0.19
33	750	Moderate	Low	.3	15	0.049	0.056	0.052	0.69	0.68	0.95	0.32	0.76	0.02	0.50	0.01	0.30
34	750	Moderate	Low	.3	30	0.049	0.046	0.043	0.82	0.66	0.95	0.44	0.77	0.02	0.50	0.02	0.30
35	750	Moderate	Low	.5	15	0.049	0.046	0.055	0.69	0.68	0.95	0.69	0.82	0.02	0.50	0.02	0.27
36	750	Moderate	Low	.5	30	0.049	0.043	0.038	0.82	0.66	0.95	0.81	0.82	0.02	0.50	0.02	0.29
37	750	Moderate	High	.3	15	0.049	0.050	0.053	0.88	0.51	0.95	0.01	0.69	0.03	0.78	0.00	0.24
38	750	Moderate	High	.3	30	0.049	0.044	0.044	0.94	0.53	0.95	0.01	0.71	0.03	0.79	0.01	0.25
39	750	Moderate	High	.5	15	0.049	0.065†	0.068†	0.88	0.50	0.95	0.51	0.79	0.03	0.78	0.00	0.23
40	750	Moderate	High	.5	30	0.049	0.056	0.055	0.94	0.53	0.95	0.67	0.80	0.03	0.79	0.02	0.24
41	750	Easy	Low	.3	15	0.048	0.075†	0.058	0.66	0.86	0.95	0.00	0.34	0.58	0.06	0.21	0.19
42	750	Easy	Low	.3	30	0.049	0.081*,†	0.043	0.79	0.87	0.95	0.00	0.36	0.60	0.10	0.20	0.21
43	750	Easy	Low	.5	15	0.049	0.164*,†	0.075†	0.66	0.86	0.95	0.00	0.56	0.58	0.06	0.18	0.21
44	750	Easy	Low	.5	30	0.049	0.269*,†	0.059	0.79	0.87	0.95	0.00	0.60	0.60	0.10	0.19	0.19
45	750	Easy	High	.3	15	0.049	0.159*,†	0.066†	0.85	1.40	0.95	0.00	0.07	1.07	0.91	0.33	0.19
46	750	Easy	High	.3	30	0.049	0.180*,†	0.065†	0.92	1.41	0.95	0.00	0.07	1.09	0.94	0.33	0.20
47	750	Easy	High	.5	15	0.049	0.635*,†	0.193*,†	0.85	1.40	0.95	0.00	0.49	1.07	0.90	0.33	0.19
48	750	Easy	High	.5	30	0.049	0.765*,†	0.192*,†	0.92	1.41	0.95	0.00	0.48	1.09	0.94	0.33	0.20

* Significant Type I Error rate based on $\alpha \pm \frac{1}{2}\alpha$; †Significant Type I Error rate based on the results of a binomial test

Iterations per condition = 1,000

See Table 4 for key

Table 6

Results of simulation 3 (high fidelity, distribution of latent construct scores = standard normal N(0,1))

c	n	$b_{i,j-1}$	a_i	β	k	π_θ	π_x	$\pi_{\hat{\theta}}$	α	*rmse*	SW_θ	SW_x	$SW_{\hat{\theta}}$	sk_x	kt_x	$sk_{\hat{\theta}}$	$kt_{\hat{\theta}}$
49	250	Difficult	Low	.3	15	0.062	0.067†	0.055	0.64	0.7	0.96	0.01	0.78	0.64	0.08	0.23	0.21
50	250	Difficult	Low	.3	30	0.062	0.078*,†	0.054	0.78	0.69	0.96	0.01	0.75	0.67	0.15	0.24	0.22
51	250	Difficult	Low	.5	15	0.062	0.099*,†	0.064†	0.64	0.70	0.96	0.10	0.89	0.64	0.09	0.23	0.21
52	250	Difficult	Low	.5	30	0.062	0.132*,†	0.057	0.78	0.69	0.96	0.15	0.83	0.67	0.15	0.24	0.22
53	250	Difficult	High	.3	15	0.062	0.128*,†	0.079*,†	0.84	1.57	0.96	0.00	0.03	1.34	1.52	0.76	0.72
54	250	Difficult	High	.3	30	0.061	0.152*,†	0.085*,†	0.91	1.56	0.96	0.00	0.04	1.38	1.63	0.76	0.70
55	250	Difficult	High	.5	15	0.061	0.390*,†	0.215*,†	0.84	1.57	0.96	0.00	0.26	1.34	1.53	0.77	0.73
56	250	Difficult	High	.5	30	0.062	0.467*,†	0.224*,†	0.91	1.56	0.96	0.00	0.27	1.37	1.62	0.76	0.67
57	250	Moderate	Low	.3	15	0.062	0.047	0.044	0.68	0.56	0.96	0.77	0.96	0.01	0.59	0.00	0.24
58	250	Moderate	Low	.3	30	0.062	0.044	0.058	0.81	0.56	0.96	0.80	0.97	0.01	0.59	0.00	0.24
59	250	Moderate	Low	.5	15	0.062	0.041	0.050	0.68	0.56	0.96	0.89	0.97	0.01	0.59	0.00	0.25
60	250	Moderate	Low	.5	30	0.062	0.040	0.050	0.81	0.56	0.96	0.90	0.96	0.01	0.59	0.00	0.24
61	250	Moderate	High	.3	15	0.062	0.047	0.056	0.88	0.38	0.96	0.11	0.93	0.02	1.01	0.01	0.44
62	250	Moderate	High	.3	30	0.062	0.042	0.059	0.93	0.39	0.96	0.14	0.93	0.02	1.02	0.01	0.43
63	250	Moderate	High	.5	15	0.062	0.031	0.042	0.88	0.38	0.96	0.79	0.95	0.02	1.02	0.01	0.43
64	250	Moderate	High	.5	30	0.061	0.032	0.054	0.93	0.39	0.96	0.86	0.95	0.02	1.03	0.01	0.43
65	250	Easy	Low	.3	15	0.061	0.075†	0.050	0.63	0.72	0.96	0.01	0.74	0.66	0.11	0.24	0.23
66	250	Easy	Low	.3	30	0.061	0.066†	0.049	0.78	0.71	0.96	0.01	0.73	0.69	0.19	0.24	0.22
67	250	Easy	Low	.5	15	0.061	0.115*,†	0.071†	0.63	0.72	0.96	0.07	0.83	0.66	0.13	0.24	0.21
68	250	Easy	Low	.5	30	0.061	0.150*,†	0.067†	0.78	0.70	0.96	0.11	0.83	0.69	0.20	0.24	0.22
69	250	Easy	High	.3	15	0.061	0.141*,†	0.098*,†	0.83	1.65	0.96	0.00	0.03	1.39	1.71	0.80	0.79
70	250	Easy	High	.3	30	0.061	0.155*,†	0.093*,†	0.91	1.65	0.96	0.00	0.02	1.42	1.81	0.80	0.79
71	250	Easy	High	.5	15	0.061	0.411*,†	0.246*,†	0.83	1.65	0.96	0.00	0.22	1.40	1.74	0.81	0.84
72	250	Easy	High	.5	30	0.061	0.488*,†	0.235*,†	0.91	1.65	0.96	0.00	0.24	1.43	1.82	0.81	0.82

* Significant Type I Error rate based on $\alpha \pm \frac{1}{2}\alpha$; †Significant Type I Error rate based on the results of a binomial test

Iterations per condition = 1,000

See Table 4 for key

Table 7

Results of simulation 4 (high fidelity, distribution of latent construct scores = standard normal N(0,1))

c	n	$b_{i,j-1}$	a_i	β	k	π_θ	π_x	$\pi_{\hat{\theta}}$	α	*rmse*	SW_θ	SW_x	$SW_{\hat{\theta}}$	sk_x	kt_x	$sk_{\hat{\theta}}$	$kt_{\hat{\theta}}$
73	750	Difficult	Low	.3	15	0.049	0.086*,†	0.066†	0.64	0.65	0.95	0.00	0.16	0.65	0.10	0.26	0.24
74	750	Difficult	Low	.3	30	0.049	0.080*,†	0.056	0.78	0.65	0.95	0.00	0.16	0.67	0.16	0.26	0.24
75	750	Difficult	Low	.5	15	0.049	0.183*,†	0.100*,†	0.64	0.65	0.95	0.00	0.54	0.65	0.10	0.26	0.24
76	750	Difficult	Low	.5	30	0.049	0.268*,†	0.086*,†	0.78	0.65	0.95	0.00	0.55	0.67	0.15	0.26	0.26
77	750	Difficult	High	.3	15	0.049	0.216*,†	0.109*,†	0.84	1.36	0.95	0.00	0.00	1.35	1.55	0.62	0.25
78	750	Difficult	High	.3	30	0.049	0.199*,†	0.095*,†	0.91	1.36	0.95	0.00	0.00	1.38	1.63	0.62	0.24
79	750	Difficult	High	.5	15	0.049	0.740*,†	0.407*,†	0.84	1.36	0.95	0.00	0.07	1.35	1.56	0.62	0.24
80	750	Difficult	High	.5	30	0.049	0.842*,†	0.388*,†	0.91	1.35	0.95	0.00	0.08	1.38	1.62	0.62	0.23
81	750	Moderate	Low	.3	15	0.049	0.047	0.049	0.68	0.56	0.95	0.15	0.90	0.01	0.58	0.00	0.27
82	750	Moderate	Low	.3	30	0.049	0.045	0.046	0.81	0.56	0.95	0.23	0.89	0.01	0.59	0.00	0.28
83	750	Moderate	Low	.5	15	0.049	0.043	0.047	0.68	0.56	0.95	0.61	0.93	0.01	0.58	0.00	0.27
84	750	Moderate	Low	.5	30	0.049	0.046	0.042	0.81	0.56	0.95	0.74	0.94	0.01	0.59	0.00	0.28
85	750	Moderate	High	.3	15	0.049	0.043	0.049	0.88	0.39	0.95	0.00	0.70	0.02	1.02	0.00	0.44
86	750	Moderate	High	.3	30	0.049	0.041	0.042	0.93	0.39	0.95	0.00	0.72	0.02	1.03	0.00	0.44
87	750	Moderate	High	.5	15	0.049	0.045	0.054	0.88	0.39	0.95	0.32	0.92	0.02	1.02	0.00	0.44
88	750	Moderate	High	.5	30	0.049	0.040	0.047	0.93	0.39	0.95	0.46	0.92	0.02	1.03	0.00	0.42
89	750	Easy	Low	.3	15	0.048	0.080*,†	0.059	0.64	0.68	0.95	0.00	0.12	0.67	0.15	0.26	0.23
90	750	Easy	Low	.3	30	0.049	0.094*,†	0.041	0.78	0.67	0.95	0.00	0.12	0.69	0.22	0.26	0.25
91	750	Easy	Low	.5	15	0.049	0.180*,†	0.094*,†	0.64	0.68	0.95	0.00	0.50	0.67	0.15	0.26	0.23
92	750	Easy	Low	.5	30	0.049	0.315*,†	0.076*,†	0.78	0.67	0.95	0.00	0.51	0.69	0.22	0.27	0.26
93	750	Easy	High	.3	15	0.049	0.199*,†	0.106*,†	0.84	1.46	0.95	0.00	0.00	1.40	1.77	0.66	0.38
94	750	Easy	High	.3	30	0.049	0.236*,†	0.107*,†	0.91	1.46	0.95	0.00	0.00	1.43	1.87	0.65	0.33
95	750	Easy	High	.5	15	0.049	0.734*,†	0.436*,†	0.84	1.46	0.95	0.00	0.06	1.40	1.77	0.66	0.38
96	750	Easy	High	.5	30	0.049	0.849*,†	0.404*,†	0.91	1.46	0.95	0.00	0.05	1.43	1.88	0.65	0.36

* Significant Type I Error rate based on $\alpha \pm \frac{1}{2}\alpha$; †Significant Type I Error rate based on the results of a binomial test

Iterations per condition = 1,000

See Table 4 for key

item category difficulty. The first eight rows in each table are the "difficult" scales, the middle eight rows are the "moderate" scales, and the last eight rows are the "easy" scales. The moderate scales represent assessment appropriateness because the distribution of the item category difficulty parameters is congruent with the theta distribution. The difficult and easy scales represent assessment inappropriateness because the distribution of the item category difficulty parameters is either more positive or more negative than the theta distribution respectively. The primary dependent variables, empirical Type I error rates (π) for the interaction term, are represented in the columns for actual latent construct scores (π_θ), raw scores (π_X), and estimated theta scores ($\pi_{\hat{\theta}}$). The values in these columns represent the proportion of times that a significant ΔR^2 was identified between the additive and multiplicative regression models across 1,000 iterations for the respective dependent variable. The theoretical alpha level (α) was set at $p < .05$ for the tests of the regression model within each iteration. This proportion represents the empirical, or observed, Type I error rate for each score type in each condition.

To assess the occurrence of spurious interaction effects, one would want to determine whether the empirical Type I error rate, π, reasonably approximates the conventional theoretical Type I error rate, α, of 5%. However, rather than treating π and α as point estimates, an interval approach has been suggested to account for sampling error that is inherent in simulation studies. Using the Bradley (1978) and Robey and Barcikowski (1992) criteria for a liberal interval with a power of .90 at 1,000 iterations, the interval of $\alpha \pm \frac{1}{2}\alpha$ or $.025 \leq \alpha \leq .075$ was used to determine whether π was a reasonable approximation of α. Therefore, columns π_θ, π_X, and $\pi_{\hat{\theta}}$ in Tables 4, 5, 6,

and 7 are marked with an asterisk (*) to indicate significant departures of π from α for the actual theta scores, raw scores, and estimated theta scores respectively. Additionally, a binomial test was conducted for each iteration to determine whether the observed proportion of Type I errors was significantly different from .05. Columns π_{θ}, π_{X}, and $\pi_{\hat{\theta}}$ in Tables 4, 5, 6, and 7 are marked with a dagger (†) to indicate significant departures of π from α for the actual theta scores, raw scores, and estimated theta scores respectively.

The average internal consistency of the raw score matrices for the two simulated independent variables, X_1 and X_2, and the one simulated dependent variable, X_3, is represented in the column for Cronbach's alpha (α). The average root mean squared error (*rmse*) is reported to represent the estimated theta parameter recovery. Root mean squared error is regarded as an appropriate statistic representing the congruence of the estimated parameters with the actual parameters in an IRT Monte Carlo study with multiple iterations (Harwell et al., 1996). The result of the Shapiro-Wilk tests are represented in the columns labeled SW_{θ}, SW_{X}, and $SW_{\hat{\theta}}$ for the actual latent construct scores, raw scores, and estimated theta scores respectively. The Shapiro-Wilk test evaluates the assumption that the residuals in the regression model are normally distributed (Shapiro & Wilk, 1965). The values in these columns represent the proportion of times that the Shapiro-Wilk test was non-significant across 1,000 iterations for the respective dependent variables, indicating violations of the residual normality assumption. Finally, the average skewness and kurtosis for the raw scores are represented in the columns denoted as sk_x and kt_x respectively. The average skewness and kurtosis for the estimated theta scores are represented in the columns denoted as $sk_{\hat{\theta}}$ and

$kt_{\hat{\theta}}$ respectively. Note that the values represented in these four columns are all absolute values.

Simulation Checks

A series of simulation checks were conducted by observing the values reported in Tables 4, 5, 6, and 7. First, the results of the average internal consistency, Cronbach's alpha, for the raw score matrices are an indication that the simulated scores are indeed following a pattern indicative of a general, underlying latent structure. As indicated by Cortina (1993), acceptable levels of internal consistency will be observed when a general, latent factor better accounts for the variance in responses than individual item effects. Although there is debate regarding an "appropriate" criterion for alpha, it is generally accepted that alpha levels of .60 to .70 and above represent adequate to good internal consistency[3]. Across Tables 4, 5, 6, and 7, the internal consistency values range between .63 and .94. In general, these scores represent appropriate levels of internal consistency, and are an indication that the variance in the simulated raw scores are indeed reflecting a general, latent construct rather than individual item effects.

Differences in the values of alpha can be easily understood by examining the respective levels of the independent variables. Specifically, the lowest alpha levels are observed in conditions with shorter scale lengths and smaller item discrimination values. Each of these factors plays an important role for internal consistency. Specifically, alpha

[3] Cortina (1993) addresses this debate by arguing that the .70 criterion for alpha is often invoked as a measure of unidimensionality. Although the .70 level does represent an adequate degree of internal consistency, it is not a measure of unidimensionality. Cortina cogently demonstrates that assessments with very low inter-item correlations can still result in alpha levels at or above the .70 criterion, and researchers interested in assessing unidimensionality must also evaluate the inter-item correlation matrix to verify unidimensional consistency.

tends to increase as assessment length increases (Crocker & Algina, 1986), as well as in assessments with highly discriminating items (Ostini & Nering, 2006). The influence of item discrimination on internal consistency with polytomous items is such that item discrimination can be thought of as a type of factor loading representing the ability of an item to differentiate individuals with varying construct scores (Ostini & Nering, 2006). Therefore, all other factors held constant, alpha should be larger in the highly discriminating scales with 30 items. Indeed, this pattern is realized across all simulation conditions.

Additionally, the distributional characteristics and the result of the Shapiro-Wilk test have important implications for the results of the simulations. First, higher degrees of nonnormality were observed in the raw scores than in the estimated theta scores. A common assumption stipulates that latent constructs are normally distributed (Embretson & Reise, 2000), and the actual theta scores in these simulations were drawn from a standard normal distribution. However, Maxwell and Delaney (1985) demonstrated that, when measuring a latent construct, the shape of the raw score distribution is often skewed. This condition was observed in these simulations, such that the skewness of the raw scores was always greater than the skewness of the estimated theta scores. Further, Maxwell and Delaney (1985) found that skewness is exacerbated in assessments with highly discriminating items and under conditions of assessment inappropriateness. This effect was also observed in the simulations, such that the skewness of the raw scores increased with higher discrimination values and under conditions of assessment inappropriateness. This suggests that the distributions of the estimated theta scores were

more representative of the actual latent construct than the distributions of the raw scores, and certain psychometric conditions only exacerbated these differences.

The second implication related to the score distributions pertains to the relationship between nonnormality and the Type I error rate. Across all simulated conditions, significant positive correlations between the empirical Type I error rate and the average absolute value of the skewness of the raw score distributions, $r = .665$, $p < .001$, and the estimated theta score distributions, $r = .637$, $p < .001$ were observed. It should be noted, however, that although these relationships are very similar in magnitude and direction, the actual values for the raw scores exceeded the values of the estimated theta scores in every condition (see Tables 4, 5, 6, and 7). These results indicate that there may be an effect of nonnormality on the robustness of the MMR models that were evaluated in these simulations. Additionally, significant negative correlations between the results of the Shapiro-Wilk tests and the empirical Type I error rates were observed for the raw scores, $r = -.512$, $p < .001$, and the estimated theta scores, $r = -.405$, $p < .001$. These results indicate that as the empirical Type I error rate increased, the number of non-significant Shapiro-Wilk tests decreased. These violations of the normality of the errors in the MMR models may have also contributed to the respective empirical Type I error rates. However, in these simulations, no conclusions about causality can be drawn for these factors. Kang and Waller (2005) also reported similar findings in their investigation of dichotomous data.

Omnibus Impact on Empirical Type I Error Rates

As a test of the influence of all of the manipulated factors on the occurrence of Type I errors in this simulation, a direct logistic regression analysis was conducted in which all of the independent variables were entered into the model simultaneously as categorical predictors. This method is preferable for assessing multiple independent variables in an exploratory and comparative manner (Tabachnick & Fidell, 2007). This analysis was conducted at the individual iteration level, not the aggregated condition level. This strategy allows an examination of the likelihood of a Type I error with the same criteria that was used to establish the empirical Type I error rates reported in Tables 4, 5, 6, and 7. The dependent variable in this analysis was the occurrence of a Type I error, and was coded as a 1 if the ΔR^2 between the additive and multiplicative model was significant at the theoretical alpha level (α) of $p < .05$ or a 0 if it was not significant. The result of this analysis is presented in Table 8 for the raw scores and Table 9 for the estimated theta scores. For both dependent variables, the full models were significant when compared to the constant-only models, and each independent variable reliably predicted the respective dependent variable.

Several important findings can be identified from these results. First, the psychometric characteristics that were manipulated in this simulation had a stronger overall effect on Type I errors when the variables were operationalized as raw scores when compared to estimated theta scores. These results suggest that raw scores are more sensitive to measurement effects in parametric analyses than are IRT-derived theta estimates. For both dependent variables, assessment appropriateness was the most

Table 8

Direct logistic regression for raw score Type I errors

Omnibus full model, χ^2 (1, N = 96,000) = 17,157.51, $p < .001$, $R^2 = .27$

	Wald χ^2	df	*B*	*OR*◊
Appropriateness (difficulty)	5,008.55***	1	2.096	8.13
Beta Weights	4,876.13***	1	1.417	4.12
Discrimination	4,437.19***	1	1.339	3.82
Fidelity	165.42***	1	0.242	1.27
Sample size	154.09***	1	0.234	1.26
Items	154.09***	1	0.234	1.26

***$p < .001$

◊ In each case, the odds ratio (OR) reported corresponds to increases in the predictor variable (e.g., increased assessment inappropriateness results in higher likelihoods of Type I errors, increases in discrimination results in higher likelihoods of Type I errors, etc.).

Table 9

Direct logistic regression for estimated theta score Type I errors

Omnibus full model, χ^2 (1, N = 96,000) = 3,571.47, $p < .001$, $R^2 = .08$

	Wald χ^2	df	*B*	*OR*◊
Appropriateness (difficulty)	918.15***	1	0.875	2.40
Discrimination	1,185.01***	1	0.836	2.30
Beta Weights	881.98***	1	0.710	2.03
Fidelity	364.77***	1	0.446	1.56
Sample size	9.93***	1	0.073	1.08
Items	9.93***	1	0.073	1.08

***$p < .001$

◊ In each case, the odds ratio (OR) reported corresponds to increases in the predictor variable (e.g., increased assessment inappropriateness results in higher likelihoods of Type I errors, increases in discrimination results in higher likelihoods of Type I errors, etc.).

impactful predictor of Type I errors, followed by item discrimination and regression weights. This result replicates the effects of assessment appropriateness identified by Kang and Waller (2005), as well as arguments raised by Busemeyer (1980) on the role of

assessment difficulty in parametric statistics. Finally, the finding that stronger regression weights resulted in a higher likelihood of Type I errors in both raw scores (OR = 4.12) and estimated theta scores (OR = 2.03) corroborates arguments raised by Rogers (2002). Specifically, Rogers (2002) found that the strength of the main effects in MMR is directly related to the likelihood of identifying an interaction effect and that an effect-size ceiling is placed on the interaction term as the main effects decrease in strength. This effect appears to be at least partially supported by these results. Further implications of the individual variables included in these analyses will be discussed as they relate to specific hypotheses.

General Findings and Hypothesis Tests

There were several findings in the simulation results that are important to emphasize in light of the hypothesis tests. First, descriptive statistics were calculated for the empirical Type I error rates in each scoring condition collapsing across each independent variable (see Table 10). A general pattern can be identified in these results such that higher empirical Type I error rates were observed for the stronger level of each independent variable. This pattern would indicate that each psychometric characteristic that was varied in the simulations had an overall effect on the empirical Type I error rates for the interaction terms. However, it should be noted that this effect was limited to the raw score and estimated theta scoring techniques.

Hypothesis 1

Hypothesis 1 stated that, under conditions in which no significant interaction is present, the use of raw scores to operationalize a latent construct will result in higher

Table 10

Mean empirical Type I error rates across the independent variables

		Mean (s.d.)		
		π_θ	π_X	$\pi_{\hat{\theta}}$
Appropriateness (difficulty)	Appropriate	0.06 (.229)	0.04 (.306)	0.05 (.218)
	Inappropriate	0.06 (.228)	0.24 (.417)	0.11 (.313)
Discrimination	Low	0.06 (.229)	0.10 (.294)	0.06 (.234)
	High	0.06 (.228)	0.25 (.435)	0.12 (.327)
Sample Size	250	0.06 (.228)	0.16 (.368)	0.09 (.290)
	750	0.06 (.228)	0.19 (.391)	0.09 (.282)
Fidelity	Normal	0.06 (.228)	0.16 (.367)	0.07 (.259)
	High	0.06 (.228)	0.19 (.392)	0.11 (.310)
Items	15	0.06 (.228)	0.16 (.368)	0.09 (.290)
	30	0.06 (.228)	0.19 (.391)	0.09 (.282)
Beta Weights	.3	0.06 (.228)	0.09 (.288)	0.06 (.242)
	.5	0.06 (.228)	0.26 (.438)	0.12 (.322)

Type I error rates than the use of actual or estimated theta scores derived using an IRT approach. Partial support was found for hypothesis 1. The empirical Type I error rate for the raw scores was higher than the empirical Type I error rate of the actual theta scores in 66 of 96 conditions (69%), as well as the estimated theta scores in 71 of 96 conditions (74%). However, in some conditions the empirical Type I error rate for the raw scores was actually lower than the empirical Type I error rate for the actual and/or estimated theta scores. Although unexpected, it is important to note that the difference between the empirical Type I error rates in these conditions was never greater than 3.2%. This

suggests that although the raw scores did perform better than the actual and/or estimated theta scores in some conditions, the performance differences in these conditions were not substantial.

The preceding results are an indication of the relative performance of the three scoring techniques, although they do not make any distinction as to meaningfully inflated Type I error rates. Recall that the interval of $\alpha \pm \frac{1}{2}\alpha$, or $.025 \leq \alpha \leq .075$, was one criterion that was used to determine whether π was a reasonable approximation of α for each scoring technique, and that values of π greater than .075 represent meaningfully inflated Type I error rates. Using this criterion, the results indicated meaningfully inflated Type I error rates in 53 of 96 conditions (55%) when raw scores were used to operationalize the latent constructs. These conditions are marked with an asterisk under the column labeled π_X in Tables 4, 5, 6, and 7. The results also indicated meaningfully inflated Type I error rates in 33 of 96 conditions (34%) when estimated theta scores were used to operationalize the latent constructs. These conditions are marked with an asterisk under the column labeled $\pi_{\hat{\theta}}$ in Tables 4, 5, 6, and 7.

The second criterion for identifying meaningfully inflated Type I error rates was the result of a binomial test for the proportion of significant interactions in each condition. This test was slightly more conservative than the approximation interval. Using this criterion, the results indicated meaningfully inflated Type I error rates in 63 of 96 conditions (66%) when raw scores were used to operationalize the latent constructs. These conditions are marked with a dagger under the column labeled π_X in Tables 4, 5, 6, and 7. The results also indicated meaningfully inflated Type I error rates in 44 of 96

conditions (46%) when estimated theta scores were used to operationalize the latent constructs. These conditions are marked with a dagger under the column labeled $\pi_{\hat{\theta}}$ in Tables 4, 5, 6, and 7.

An important finding to highlight here is that, of the conditions with meaningfully inflated Type I error rates for the estimated theta scores, none were unique with regard to the raw scores using either criteria. In other words, no meaningful inflations existed for the estimated theta scores that did not also exist for the raw scores. This finding was true regardless of the criteria used to determine a meaningful inflation of the Type I error rate. This suggests that there were no unique conditions in which the raw scores performed practically better than the estimated theta scores. This finding gives better insight into the previously reported result that the empirical Type I error rate for raw scores was actually lower than the empirical Type I error rate of the actual and/or estimated theta scores in some conditions. Specifically, these differences were always observed in conditions in which the theoretical alpha level was well approximated by all three scoring techniques and therefore do not suggest any noteworthy advantages for the raw scores.

Hypothesis 1a

Hypothesis 1a stated that, under conditions of assessment appropriateness, the Type I error rates for raw scores, actual, and estimated theta scores will not exceed the nominal criterion of $\alpha = .05$. Full support was found for hypothesis 1a. Under conditions of assessment appropriateness there were no significant departures from the acceptable interval of $\alpha \pm \frac{1}{2}\alpha$ for any of the scoring methods or other simulated factors (see rows 9-16 in table 4, 33-40 in table 5, 57-64 in table 6, and 81-88 in table 7). This suggests that

when assessment difficulty and the distribution of construct scores are reasonably matched, there is little concern for an increased risk of Type I errors for any scoring technique. Indeed, this pattern was also realized by Kang and Waller (2005) in their investigation of dichotomous data. These results are also meaningful in the interpretation of hypothesis 2.

Hypothesis 2

Hypothesis 2 stated that, assessment inappropriateness will influence the prevalence of Type I error rates for the interaction term in moderated multiple regression. Support was identified for Hypothesis 2 such that all of the meaningfully inflated Type I error rates were observed when assessment inappropriateness was present. These results imply that the risk of spurious interactions is only a concern when the difficulty of the assessment is poorly matched to the construct levels of the individuals. A further examination of this trend reveals support for Hypothesis 2a.

Hypothesis 2a

Hypothesis 2a stated that, under conditions of assessment inappropriateness, the use of raw scores to operationalize a latent construct will result in Type I error rates that exceed the nominal criterion of $\alpha = .05$. Specifically, raw scores resulted in empirical Type I error rates that were above the acceptable interval in 53 of 64 (83%) conditions in which assessment inappropriateness was present. A direct logistic regression analysis that included all of the simulated independent variables was conducted to test the impact of assessment inappropriateness on Type I errors for the raw scores. According to the Wald criterion, assessment appropriateness reliably predicted a Type I error for the raw scores,

$\chi^2(1, N = 96{,}000) = 5{,}008.55$, $p < .001$, $OR = 8.13$. This result suggests that, in the presence of the other predictors, the likelihood of committing a Type I error is 8.13 times higher under conditions of assessment inappropriateness than under conditions of assessment appropriateness when raw scores are used to operationalize the variables.

Additionally, Figure 7 represents the descriptive statistics for the empirical Type I error rates for the raw scores and estimated theta scores. For the raw scores, these data indicate a positively skewed distribution (skew = 2.04), with a mean empirical Type I error rate of 17.5%, median of 8.7%, and a standard deviation of 20%. Finally, the values range from 3.1% to 84.9%. If we were to only examine those values outside of the acceptable interval for alpha (greater than .075), the mean empirical Type I error rate was 27.6%, with values ranging from 7.8% to 84.9%.

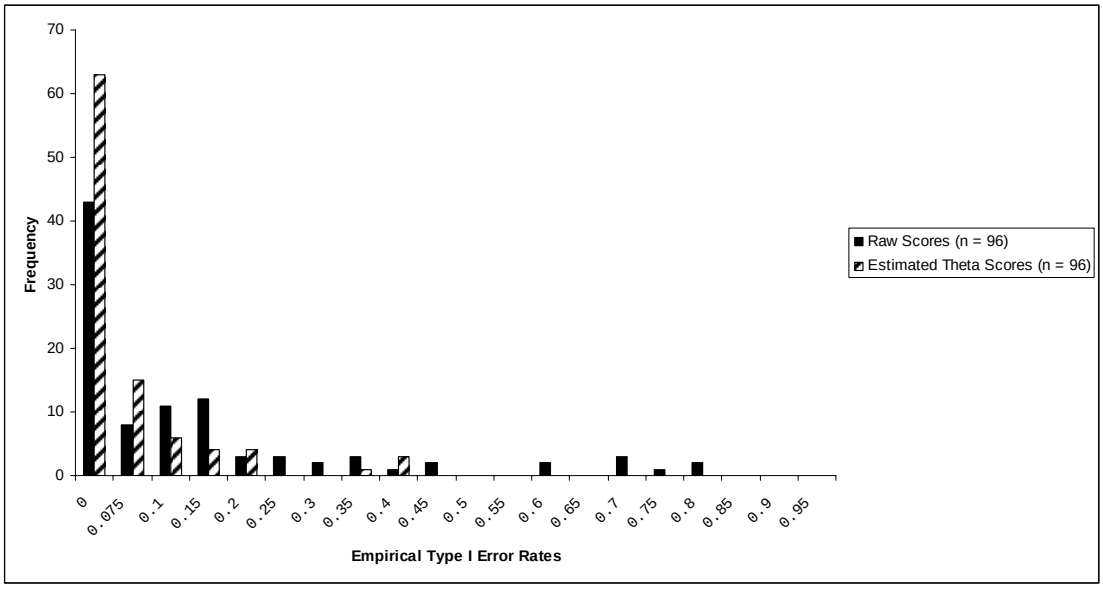

Figure 7. Distribution of spurious interactions for raw scores and estimated theta scores

Hypothesis 2b

Hypothesis 2b stated that, under conditions of assessment inappropriateness, the use of estimated or actual latent trait scores to operationalize a latent construct will not result in Type I error rates that exceed the nominal criterion of $\alpha = .05$. Hypothesis 2b was supported for the actual theta scores, but not for the estimated theta scores. For the estimated theta scores, meaningfully inflated Type I error rates were identified in 33 of 64 (51%) inappropriate assessment conditions. A direct logistic regression analysis that included all of the simulated independent variables was conducted to test the impact of assessment inappropriateness on Type I errors for the estimated theta scores. According to the Wald criterion, assessment appropriateness reliably predicted a Type I error for the estimated theta scores, $\chi^2(1, N = 96{,}000) = 918.15$, $p < .001$, $OR = 2.40$. This result suggests that, in the presence of the other predictors, the likelihood of committing a Type I error is 2.4 times higher under conditions of assessment inappropriateness than under conditions of assessment appropriateness when estimated theta scores are used to operationalize the variables. The descriptive statistics for the estimated theta score distribution also indicate a positively skewed distribution (skew = 2.97), with a mean empirical Type I error rate of 9.0%, median of 5.9%, and a standard deviation of 8.0% (see Figure 7). Finally, the values range from 3.7% to 43.6%. If we were to only examine those values outside of the acceptable interval for alpha (greater than .075), the mean empirical Type I error rate was 15.9%, with values ranging from 7.6% to 43.6%.

Although Hypothesis 2b was not directly supported, the estimated theta scores did fare substantially better with regards to spurious interaction effects when compared to the raw scores. An examination of Figure 7 clearly illustrates a more troubling picture for raw scores than for estimated theta scores. Specifically, the empirical Type I error rates for raw scores that were beyond the acceptable interval ranged from slightly above the interval at 7.8% to an extremely high rate of 84.9% observed in condition 96. However, the empirical Type I error rates for estimated theta scores that were beyond the acceptable interval ranged from slightly above the interval at 7.6% to a moderately high rate of 43.6%. In most cases of empirical Type I error rate divergence, the rate for the raw scores was substantially higher than the corresponding rate for the estimated theta scores.

One reason that Hypothesis 2b was not fully supported may be that the estimated theta scores were less accurate in conditions of assessment inappropriateness. This reduction in parameter accuracy would be expected given the score attenuation that is present in exceedingly easy or difficult assessments (Busemeyer, 1980; Embretson & Reise, 2000; Hambleton et al., 1991). An indication that this trend is present is observable in the root mean squared error (*rmse*) values in Tables 4, 5, 6, and 7. In these simulations, the root mean squared error is an estimate of parameter recovery, or the congruence between the estimated and actual theta scores. Larger values of *rmse* suggest greater differences between the actual and estimated theta scores. To test the differences in the *rmse* values between the appropriate and inappropriate assessment conditions, an independent samples t-test was conducted. The results indicated a significant difference, $t(94) = 8.74$, $p < .001$, such that the *rmse* in the inappropriate assessment conditions was

significantly larger ($M = 1.12$, $sd = 0.37$) than the *rmse* in the appropriate assessment conditions ($M = 0.54$, $sd = 0.11$). This finding could provide some indication as to the meaningfully inflated Type I error rates for the estimated theta scores under conditions of assessment inappropriateness.

Hypothesis 3

Hypothesis 3 stated that, under conditions of extreme assessment inappropriateness, the use of raw scores to operationalize a latent construct will result in the highest prevalence of Type I error rates beyond the nominal criterion of $\alpha = .05$ for the interaction term in moderated multiple regression. The results did not support this Hypothesis. Hypothesis 3 was tested using a 2x2 between subjects analysis of variance. The full model included the main effects of fidelity and assessment appropriateness and the interaction between fidelity and appropriateness. The dependent variable in this analysis was the empirical Type I error rate for the raw scores. As a follow-up to Kang and Waller (2005), this analysis served to assess the prediction that the interaction of appropriateness and fidelity would create conditions of extreme assessment inappropriateness, and result in a high rate of Type I errors when raw scores are used to operationalize a latent construct. The interaction between fidelity and assessment appropriateness was not significant, $F(1, 92) = .417$, $p = .520$, $\eta^2 = .005$.

To further examine this result, a test of the model that included only the main effects of fidelity and assessment appropriateness was conducted. These results indicated that assessment appropriateness was the only significant predictor of differences in the empirical Type I error rate for raw scores, $F(1, 93) = 25.65$, $p < .001$, $\eta^2 = .216$, such that

inappropriate assessments were associated with significantly higher empirical Type I error rates than appropriate assessments. The means plots presented in Figure 8 confirms that there is a strong main effect of assessment appropriateness, but no effects of fidelity or the interaction between the two variables. These results suggest that assessments with

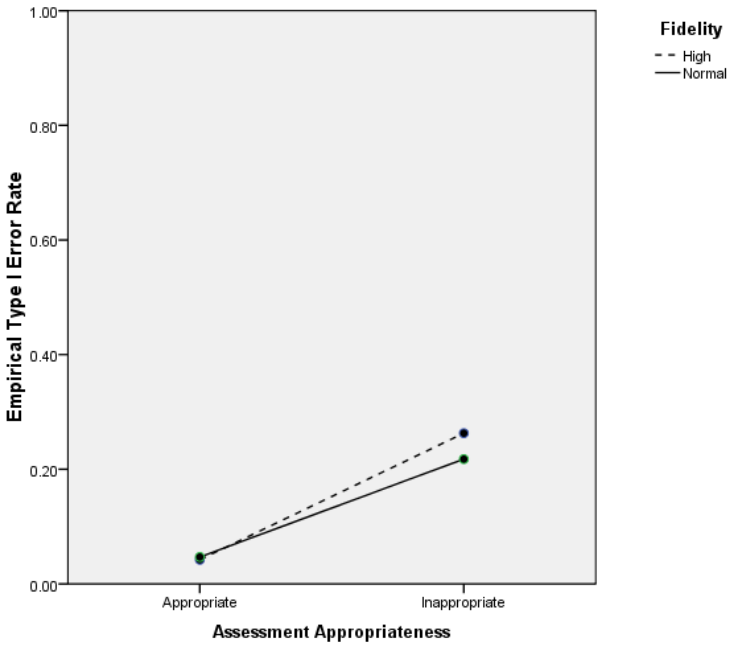

Figure 8. Interaction of fidelity and assessment appropriateness on the empirical Type I error rate for raw scores

high fidelity, or highly "peaked" tests, do not necessarily contribute to spurious interaction effects above and beyond the effect of assessment appropriateness. In other words, the effect of assessment appropriateness was the dominant factor in the omnibus model.

As an additional test of this Hypothesis, a stepwise logistic regression model was created to examine the interaction of assessment appropriateness and fidelity at the iteration level. The first step of this model included the main effects of appropriateness and fidelity, and the second step introduced the interaction between appropriateness and fidelity. The results of this analysis indicated that the main effects model was significant when compared to the constant-only model, χ^2 (2, N = 96,000) = 6,955.77, $p < .001$, R^2 = .1156, and that the model including the interaction between appropriateness and fidelity was also significant when compared to the main effects only model, χ^2 (3, N = 96,000) = 6,995.58, $p < .001$, R^2 = .1163. According to the Wald criterion, the interaction between assessment appropriateness and fidelity reliably predicted a Type I error for the estimated theta scores, χ^2(1, N = 96,000) = 39.75, $p < .001$, *OR* = 1.44. This result indicates that the likelihood of a spurious interaction effect is 1.44 times greater for high fidelity assessments when compared to low fidelity assessments under conditions of assessment inappropriateness. Although this result contradicts the results of the ANOVA conducted at the aggregate level, it should be noted that the significance observed here is likely due to the large sample size. An examination of the change in the amount of variance that is accounted for by the interaction term reveals that the interaction between assessment appropriateness and fidelity only accounts for an additional .07% of the variance in spurious interaction effects.

Hypothesis 4

Hypothesis 4 stated that, simulated assessments with higher item discrimination scores and stronger regression coefficients will result in the highest occurrence of Type I

errors for the interaction term in moderated multiple regression when raw scores are used to operationalize a latent construct. The results supported this Hypothesis. Hypothesis 4 was tested using a 2x2 between subjects analysis of variance. The full model included the main effects of item discrimination and regression weights and the interaction between discrimination and regression weights. The dependent variable in this analysis was the empirical Type I error rate for the raw scores. As a follow-up to Kang and Waller (2005), this analysis served to assess the prediction that the interaction of item discrimination and regression weights would create conditions with a high rate of Type I errors when raw scores are used to operationalize a latent construct. The interaction between discrimination and regression weights was significant, $F(1, 92) = 10.83$, $p = .01$, $\eta^2 = .105$. An examination of the means plots presented in Figure 9 suggests that there are significant differences in the empirical Type I error rate for raw scores due to discrimination and regression weights, and the effect of regression weights is greatest at high levels of item discrimination.

As an additional test of this Hypothesis, a stepwise logistic regression model was created to examine the interaction of item discrimination and regression weights at the iteration level. The first step of this model included the main effects of item discrimination and regression weights, and the second step introduced the interaction between item discrimination and regression weights. The results of this analysis indicated that the main effects model was significant when compared to the constant-only model, $\chi^2 (2, N = 96{,}000) = 9{,}354.15$, $p < .001$, $R^2 = .154$, and that the model including the interaction between item discrimination and regression weights was also significant when

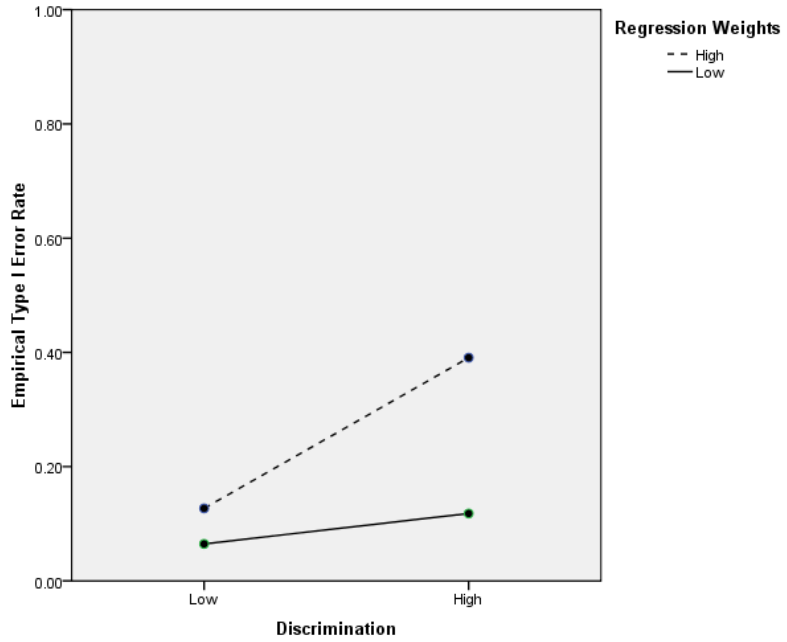

Figure 9. Interaction of item discrimination and regression weights on the empirical Type I error rate for raw scores

compared to the main effects only model, χ^2 (3, N = 96,000) = 9,759.19, $p < .001$, R^2 = .160. According to the Wald criterion, the interaction between item discrimination and regression weights reliably predicted a Type I error for the estimated theta scores, χ^2(1, N = 96,000) = 416.59, $p < .001$, *OR* = 2.29. This result indicates that the likelihood of a spurious interaction effect is 2.29 times greater with strong regression weights when compared to weaker regression weights when assessments have highly discriminating items. An examination of the change in the amount of variance that is accounted for by the interaction term reveals that the interaction between item discrimination and regression weights accounted for an additional .6% of the variance in spurious interaction effects.

Scale of Measurement and Linearity of Raw Scores and Estimated Theta Scores

Previous research has advanced the idea that a primary mechanism underlying spurious interaction effects is the scale of measurement on which various scores can be classified. Specifically, data that fails to achieve interval level scaling can be predisposed to an increased risk for Type I errors in moderated statistical models (Davison & Sharma, 1990; Embretson, 1996; Kang & Waller, 2005). It is often considered that the raw scores that are generated from latent construct assessments under the classical test theory framework achieve ordinal scales of measurement at best, but that theta estimates derived from IRT models can achieve interval or nearly interval level scaling. (Borsboom, 2008; Embretson, 1996; Embretson & DeBoeck, 1994; Harwell & Gatti, 2001; Kang & Waller, 2005; Perline et al., 1979; Reise & Haviland, 2005; Reise et al., 2005; Rupp & Zumbo, 2006; Wainer, 1982). Measurement researchers have appeared to reach a general consensus that the Rasch model and the scores that are derived from it can fully achieve interval level scaling, and other IRT models can achieve a scale closer to interval level than simple raw scores (Embretson, 2006). Reise and Haviland (2005) reflect this principle by referring to IRT-derived theta estimates as being *optimally* scaled. So far, only one study has been located that directly assessed the degree to which multicategory response data and a polytomous IRT model achieve interval-level scaling. Harwell and Gatti (2001) provided some evidence that scores generated using the graded response model (GRM) can achieve interval, or nearly interval-level scaling. The authors tested this assumption by generating distributions of the residuals between the estimated theta

scores and the actual theta scores, and concluded that the residuals were within a range attributable to sampling error. However, Harwell and Gatti cautioned that their results were preliminary and that more research on these relationships is needed.

Similar to Harwell and Gatti (2001), the root mean squared error (rmse) was averaged over all of the iterations in each condition of this dissertation. These data are presented in the column labeled *rmse* in Tables 4, 5, 6, and 7. A trend is observable in the rmse values in Tables 4, 5, 6, and 7 such that larger rmse's are typically associated with larger empirical Type I error rates. This would provide initial evidence that the estimated theta scores are more poorly approximating the actual theta scores as the empirical Type I error rate increases. To the extent that the actual theta scores achieve an interval scale, these results would lend support to Harwell and Gatti's (2001) findings.

A more compelling argument can possibly be made for the interval-level scaling of the data that was generated in this dissertation by assessing the linearity of the data. One of the qualities of scale of measurement classifications is that they provide us with admissible transformations for the data (Stevens, 1946; Stine, 1989). Linear transformations are needed for interval level data, but not for ordinal level data. Therefore, tests of linear associations, such as the Pearson product-moment correlation, should reasonably differentiate between interval and ordinal data. As an example, the actual theta scores, estimated theta scores, and raw scores that were generated in conditions 80, 88, and 96 were correlated within each iteration. In condition 80, an average correlation over all of the iterations in this condition revealed that actual theta scores were correlated with raw scores, $r = .882$, $p < .001$, and with estimated theta

scores, $r = .925$, $p < .001$. In condition 88, an average correlation over all of the iterations in this condition revealed that actual theta scores were correlated with raw scores, $r = .953$, $p < .001$, and with estimated theta scores, $r = .962$, $p < .001$. In condition 96, an average correlation over all of the iterations in this condition revealed that actual theta scores were correlated with raw scores, $r = .877$, $p < .001$, and with estimated theta scores, $r = .941$, $p < .001$.

When data meet the linearity assumption, differences in correlations are due to the magnitude of the covariance between the variables. However, nonlinearity can attenuate correlations as well. An examination of the scatter plots from 1,875 randomly selected data points in conditions 80, 88, and 96 reveals that the nonlinearity in the raw score – theta score relationship is likely to be the attenuating factor in the differences between the aforementioned correlation coefficients (see Figure 10). Additionally, a ceiling and floor effect can be observed in conditions 80 and 88 respectively due to the relative item category difficulty parameters in each condition. These conditions also reflect the effects of assessment inappropriateness on the relationships between the scores and the actual latent construct.

These results would appear to indicate that there is a stronger linear relationship between estimated theta scores and actual theta scores than there is between raw scores and actual theta scores. Although this may provide some indication of the interval-level scaling of the estimated theta scores, it is always tempered by the extent to which the actual theta scores are at an interval scale themselves. Recall that earlier work has indicated that the only IRT model that is thought to be theoretically capable of achieving

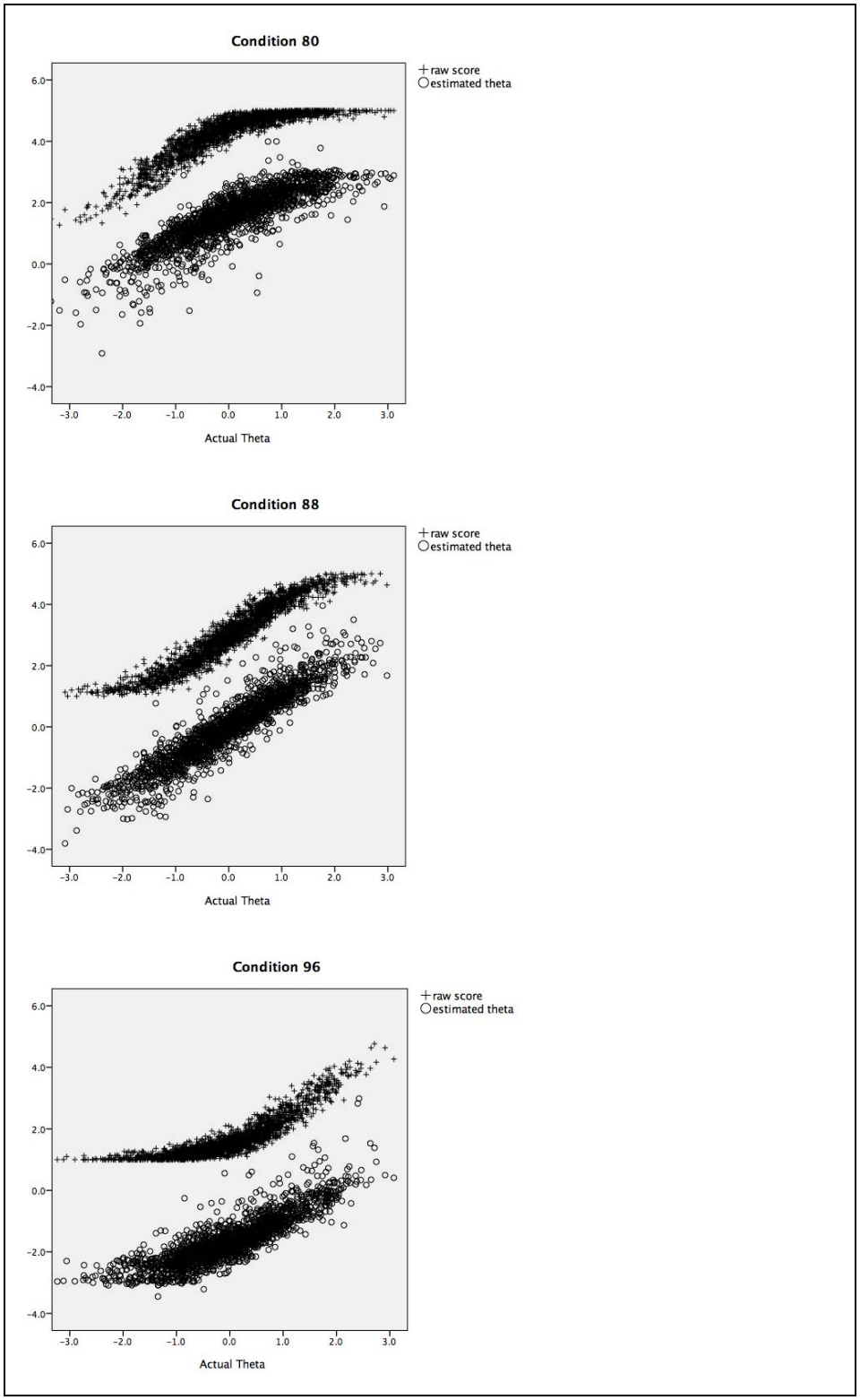

Figure 10. Linearity of the relationships between actual theta scores, raw scores, and estimated theta scores from three simulation conditions

interval scales is the Rasch model (Embretson, 2006; Embretson & DeBoeck, 1994; Perline et al., 1979; Reise & Haviland, 2005; Reise et al., 2005; Rupp & Zumbo, 2006; Wainer, 1982). Thus, data that is generated using alternative IRT models may not perfectly achieve interval scales of measurement. This evidence provides some support for the interval scale argument for polytomous models, but further research into these relationships is certainly needed.

CHAPTER 5: DISCUSSION

General Discussion

The Monte Carlo simulation that was conducted in this dissertation was designed to investigate the effects of the psychometric characteristics of a latent construct assessment and scoring techniques on the empirical Type I error rate for the interaction term of a moderated multiple regression model. This simulation was a direct extension of the simulation conducted by Kang and Waller (2005). The Kang and Waller study investigated the relative performance of raw scores and estimated theta scores with regards to Type I error rates in MMR using dichotomous data and the two-parameter logistic IRT model. The simulation conducted in this dissertation modeled multi-category, Likert-type data, and applied a polytomous IRT model, the graded response model. Additionally, this simulation was designed to reflect common scale characteristics in applied psychology to enhance the generalizability of the findings to this field. This simulation was also designed to achieve a higher degree of experimental rigor such that the number of iterations per condition was increased from 500 in Kang and Waller's study to 1,000. These characteristics are important for the generalizability of the results when sampling from a pseudo-random population in a simulation study (Harwell et al., 1996). Finally, an approximation interval for the empirical Type I error rate was established using Robey and Barcikowski's (1992) recommendations and a binomial test was conducted for the purposes of identifying meaningfully inflated Type I errors.

This dissertation was conducted to address two primary research goals. Theoretical and empirical evidence have emerged to suggest that using IRT to

operationalize an individual's standing on a latent construct has important measurement implications over the use of raw scores (Borsboom, 2008; Embretson, 1996; Embretson, 2006; Embretson & DeBoeck, 1994; Harwell & Gatti, 2001; Kang & Waller, 2005; Perline et al., 1979; Reise & Haviland, 2005; Reise et al., 2005; Rupp & Zumbo, 2006; Wainer, 1982). Specifically, IRT-derived theta scores have been demonstrated to be resistant to inflated Type I error rates in moderated statistical models due to achieving an interval, or nearly interval, scale of measurement (Embretson, 1996; Kang & Waller, 2005). This previous work, although highly illuminating, has been limited to applications of dichotomous data and restrictive IRT models. Therefore, my first goal in this dissertation was to extend our understanding of these potentially beneficial measurement properties by modeling multicategory data and implementing a polytomous IRT model.

The second goal was to generalize these findings to the applied psychological literature with the purpose of promoting IRT as a useful statistical tool. Although many researchers in this area have called attention to the importance of measurement (Austin et al., 2002; Cortina, 1993; Phillips & Lord, 1986; Podsakoff & Dalton, 1987; Scandura & Williams, 2000; Smith & Stanton, 1998; Stone-Romero, 1994), and the usefulness of IRT (Hulin & Ilgen, 1990; Zickar, 1998), summative evidence suggests that few are embracing their calls. However, due to the prevalence of polytomous data (Aguinis et al., 2009; Austin, et al., 2002; Fields, 2002), and the popularity of moderated statistical models (Aguinis, 2004; Aguinis, et al., 2005; Aguinis & Stone-Romero, 1997; Stone-Romero et al., 1994; Bartlett et al., 1978; Lubinski & Humphreys, 1990), a strong case can be made for adding IRT to the general statistical repertoire of applied psychological

researchers. The contribution of IRT being represented here is a scoring technique that, in certain contexts, can result in better accuracy for parametric analyses.

Among several important findings to address, it is imperative to point out that under certain conditions, the meaningfully inflated Type I error rates were observed for both the estimated theta scores and the raw scores. This result for the raw scores was expected, however, this result for the estimated theta scores was somewhat unexpected. This finding should caution researchers that a perfectly functioning metric was not identified in this dissertation for the graded response model. However, it was often the case that the Type I error rate of the raw scores far exceeded that of the estimated theta scores. This finding provided some support for Hypothesis 1, which stated, "Under conditions in which no significant interaction is present, the use of raw scores to operationalize a latent construct will result in higher Type I error rates than the use of actual or estimated latent trait scores derived using an IRT approach". For example, in conditions 7, 8, 23, and 24 in Table 4, the Type I error rates for the raw scores ranged from .296 to .386 whereas the respective Type I error rates for the estimated theta scores ranged from .089 to .120. Clearly, given the alternative, the estimated theta scores would be more attractive to researchers in these conditions. Additionally, cases in which the Type I error rates were grossly inflated such as conditions 31, 32, 47, and 48 in Table 5 and 79, 80, 95, and 96 in Table 7, the Type I error rate for the raw scores was approximately 200% to 450% higher than the Type I error rate for the estimated theta scores. An examination of Figure 7 clearly reveals these disparities between the raw scores and estimated theta scores. Again, although the estimated theta scores did not

perform perfectly within the acceptable limits, these results demonstrate a clear preference for their use in applied research when certain psychometric conditions exist.

Another prominent result was that of the role of assessment appropriateness on Type I error rates for both raw scores and estimated theta scores. Assessment appropriateness is defined as the congruence between the reliability of an assessment and the latent construct distribution of the individuals responding to an assessment. In IRT, an assessment's reliability is variable and linked to item difficulty. Peak item reliability occurs at the point along the construct continuum at which the individuals have a fifty percent chance of responding correctly, or positively, to an item. Peak assessment reliability is a cumulative function of item reliability, and hence, item difficulty (Embretson & Reise, 2000; Hambleton et al., 1991). Inappropriate assessments are defined as those that violate this congruence, such that they are either too "easy" or too "difficult" for the individuals. The results of this simulation demonstrated that, under conditions of assessment appropriateness, there is no concern as to unacceptable Type I error rates for any psychometric condition or scoring technique. This finding provided full support for Hypothesis 1a, which stated, "Under conditions of assessment appropriateness, the Type I error rates for raw scores, actual, and estimated latent construct scores will not exceed the nominal criterion of $\alpha = .05$".

The complimentary result, that spurious interactions were only observed under some conditions of assessment inappropriateness, was also true. These results were not unexpected and supported Hypothesis 2, which stated, "Assessment inappropriateness will influence the prevalence of Type I error rates for the interaction term in moderated

multiple regression". Embretson (1996) and Kang and Waller (2005) also identified assessment appropriateness and inappropriateness as the primary factor in their simulations of spurious interaction effects. Embretson (1996) determined that the degree and direction of the inappropriateness fully accounted for the nature of the interaction with regard to treatment groups in a simulated factorial ANOVA. Prior to these studies, Maxwell and Delaney (1985) cogently demonstrated how various distributional shapes of latent constructs can interact with assessment difficulty (appropriateness) to result in artificial group mean differences when the observed scores and latent scores were related through a non-linear, monotonic relationship. These relationships were also reflected in the findings in this dissertation. Full support was identified for Hypothesis 2a, which stated, "Under conditions of assessment inappropriateness, the use of raw scores to operationalize a latent construct will result in Type I error rates that exceed the nominal criterion of $\alpha = .05$". However, the prediction that this effect would not occur when using estimated theta scores was not fully supported. Indeed, Hypothesis 2b, which stated, "Under conditions of assessment inappropriateness, the use of estimated or actual latent trait scores to operationalize a latent construct will not result in Type I error rates that exceed the nominal criterion of $\alpha = .05$" was not fully supported. This finding suggests that certain conditions may reduce the level of linearity in the theta – estimated theta relationship. Specifically, an examination of the rmse values in these conditions indicates less precise parameter recovery. This would suggest that the degree of congruence between the actual and estimated theta scores was being eroded in exceedingly easy or difficult assessments.

Other findings were also very similar to previous investigations. Both Embretson (1996) and Kang and Waller (2005) found little effect of assessment length on Type I error rates for any scoring condition. This result was also replicated in this dissertation. Other factors being held constant, the Type I error rates for the 15 and 30 item assessments were reasonably similar. Under classical test theory, increasing scale length is one method of increasing the reliability of an assessment. However, these results suggest that this approach would not help increase the accuracy of statistical analyses. Kang and Waller (2005) also found that item discrimination and regression coefficients to be influential factors for spurious interaction effects. Indeed, this result was replicated here by finding support for Hypothesis 4, which stated, "Simulated assessments with higher item discrimination scores and stronger regression coefficients will result in the highest occurrence of Type I errors for the interaction term in moderated multiple regression when raw scores are used to operationalize a latent construct".

One expected result from previous research was not supported. Specifically, Hypothesis 3, which stated, "Under conditions of extreme assessment inappropriateness, the use of raw scores to operationalize a latent construct will result in the highest prevalence of Type I error rates beyond the nominal criterion of $\alpha = .05$ for the interaction term in moderated multiple regression" was not supported. Extreme assessment inappropriateness was defined as assessment inappropriateness combined with a high fidelity assessment. This would create peaked tests that are either too easy or too difficult for the individuals answering the items. In the context of attitude-based assessments such as those modeled here, this can be thought of as creating assessments to

which most individuals fully agreed or fully disagreed with the majority of the items. The findings in this dissertation did not support the hypothesized interaction between appropriateness and fidelity. Instead, the results appeared to suggest that fidelity had a very small effect on the empirical Type I error rate, and assessment appropriateness dominated the relationship.

Finally, it is important to highlight the distributional characteristics of the dependent variables and the tests of a key assumption in MMR analyses that are included in Tables 4, 5, 6, and 7. Specifically, the average skewness and kurtosis of the distribution of the variables was significantly positively correlated with the empirical Type I error rate for both the raw scores and the estimated theta scores. These characteristics can have implications for the robustness of parametric analyses (Bradley, 1982; 1984). The same pattern was observed by Kang and Waller (2005) in their simulation involving dichotomous data. The researchers speculated that correcting this non-normality in the raw scores with the Box-Cox transformation (Box & Cox, 1964) may help to attenuate the empirical Type I error rate for the raw scores to a level similar to that of the estimated theta scores. If successful, this may provide researchers with another method of operationalizing their data that does not involve fitting an IRT model. In a brief empirical treatment, Kang and Waller (2005) concluded that this process does work in some cases, but not in others. Specifically, moderate empirical Type I error rates, such as $\pi = .15$, responded well to this correction and were reduced to approximately .05, but high rates of $\pi = .40$ and $\pi = .53$ were only reduced to .19 and .31 respectively. Given that the most extreme empirical Type I error rates that were observed in this dissertation

exceeded values of .60 and .80, this correction may be of lesser value. A more thorough examination of alternative score transformations such as the Box-Cox transformation would certainly be of interest.

Implications for Measurement

A primary impetus for conducting this simulation was to extend the results of Embretson (1996) and Kang and Waller (2005). Building upon theoretical arguments posed during the 1980's (Busemeyer, 1980; Davison & Sharma, 1990; Maxwell & Delaney, 1985), Embretson (1996) found that actual theta scores were resistant to both Type I and Type II errors in factorial ANOVA whereas raw scores were not. Kang and Waller (2005) extended this work to find that estimated theta scores from the two-parameter logistic IRT model were also resistant to Type I errors in MMR analyses. Finally, the results of the simulation conducted in this dissertation suggest that estimated theta scores from the GRM were also more resistant to Type I errors in MMR than were raw scores. These studies represent a generalizability trend such that each successive study branched further away from the measurement ideal and into the realities of psychological data. This dissertation represents a major step in this area as it was the first study to investigate these effects using polytomous data.

The investigation of the performance of polytomous IRT models in a variety of contexts is still an important avenue of measurement research. Much attention has been paid to the mathematical benefits of the Rasch model for dichotomous data with regard to measurement scale (Embretson & Reise, 2000; Fischer, 1995; Perline et al., 1979) and parametric analyses (Fraley et al., 2000; Reise & Haviland, 2005). However, much less

attention has been paid to polytomous IRT models. Therefore, research that directly evaluates the performance of polytomous IRT models in a variety of measurement and statistical contexts is of significant importance. In their examination of the scale of measurement of the data generated with the GRM, Harwell and Gatti (2001) indicated that more research is needed to understand the properties of estimated theta scores from more complex IRT models such as the GRM. Specifically, the authors call for careful examinations of a variety of item and scale properties that can influence the data that is generated with these models. The simulations conducted in this dissertation represent an important step forward in filling these requests.

Implications for Applied Psychology

In addressing the implications of this dissertation in applied psychology, we must return to a discussion of the current state of measurement and psychometrics in the field. Although measurement is recognized as a key feature of research quality in applied psychology, it is often relegated to a brief mention of reliability (Podsakoff & Dalton, 1987; Scandura & Williams, 2000), an afterthought (Austin et al., 2002; Stone-Romero, 1994), or worse yet, a blatant misinterpretation of psychometric indicators (Phillips & Lord, 1986; Cortina, 1993). However, as Schriesheim, Powers, Scandura, Gardiner, and Lankau (1993) as well as Schoenfeldt (1984) indicated, the validity of applied psychological research is wholly dependent on the quality of our measurement. Indeed, this sentiment holds true for any research area that relies heavily on the assessment of latent constructs as a primary data collection methodology.

Compounding the aforementioned factors is the overwhelming reliance on classical test theory approaches to psychometric evaluation in applied psychology. The availability and utility of modern test theory approaches, such as IRT, can improve the quality of our measurement practices if they are adopted. The results of this dissertation provide two arguments for integrating modern measurement theory into applied psychological research. The first argument is related to a general shift towards modern measurement practices in applied psychology for scale evaluation. The second argument is related to the use of IRT as a scoring method to increase the quality of parametric analyses. These two arguments are presented in detail below.

IRT for Scale Evaluation

In their seminal chapter in the *Handbook of Industrial and Organizational Psychology*, Drasgow and Hulin (1990) made a general call to applied psychological researchers for the use of IRT for scale development, maintenance, and evaluation. Since this time, IRT has had trouble gaining ground with psychometric research in applied psychology (Austin et al., 2002). However, a clear demonstration of the unique benefits of IRT over classical test theory methods may provide a stronger case for IRT as a psychometric methodology.

The identification of assessment appropriateness using variable reliability estimates is an example of a unique quality of IRT that was found to be the most impactful factor on differences in empirical Type I error rates. However, due to the invariance problem, assessment appropriateness is more difficult to determine using classical test theory approaches. Assessment appropriateness is a function of the variable

reliability estimates that are possible with IRT (Embretson & Reise, 2000; Hambleton et al., 1991), and assessment information curves that can be derived in all IRT models provide a very useful way to observe these properties. Therefore, the most serious concern identified in this dissertation and in two prior simulations (Embretson, 1996; Kang & Waller, 2005) for spurious interaction effects is more accessible using IRT for the evaluation of our scales. To date, only one study has specifically examined this property for a commonly used assessment in applied psychology. Morse and Griffeth (2009) found peaked information curves for the MSQ short form at the lower end of the construct continuum suggesting possible assessment inappropriateness if the respondents are of moderate or high job satisfaction. Due to the influence of assessment appropriateness on spurious interaction effects, there is clearly a need to examine other assessments for these characteristics.

Very few commonly used scales in applied psychology have been assessed using IRT models (c.f. Hulin & Mayer, 1986; Reeve & Smith, 2001; Zagorsek et al., 2006), but the tools and expertise are available to expand the application of model-based measurement in applied psychological research. Researchers and practitioners in this area wishing to utilize IRT models to evaluate existing assessments for these characteristics could follow these five steps. First, sufficient data must be collected for the purposes of fitting the IRT model. Recent research has indicated that sample sizes of at least 250 individuals are preferable for fitting polytomous IRT models (Ostini & Nering, 2006). Second, the dimensionality of the construct assessment must be determined. This information can often be assessed using previous research related to the development and

validation of the assessment. For most IRT models, unidimensionality must be obtained either by verifying that the assessment has a single, dominant factor or by separating the assessment into its unidimensional sub-scales (Hulin & Ilgen, 1990). Third, an IRT model must be chosen to apply to the data. Currently, there are several polytomous IRT models that are readily available for use and the interested reader can reference Ostini and Nering (2006) for a full treatment of the similarities and differences between them. However, given its conceptual appeal as a difference family model, the widespread availability of software that includes it, and the existing simulation information related to Type I errors, Samejima's graded response model (GRM) is a reasonable choice for many applications in applied psychology. Fourth, the scale response data must be fit to the model using an IRT software package. Currently, there are several free packages available in the R environment such as the latent trait model package (Rizopoulos, 2005), as well as commercially available software such as *PARSCALE* (Muraki & Bock, 2003) that can appropriately handle polytomous IRT models including the GRM. Finally, assessment information curves can be generated from the software that will indicate where the reliability of the assessment is highest and lowest. These curves can give the researcher an indication of whether assessment appropriateness is likely to be fulfilled based on whether and where the reliability peaks in relation to the latent construct continuum. Additional information such as item discrimination and item difficulty can also be generated for the purposes of referencing the tables presented in this dissertation as an approximate estimate of the Type I error risk. Interested readers can also reference

Harwell and Gatti (2001) for a similar step-wise treatment for rescaling ordinal data with IRT for use in parametric analyses.

As an example of this process, a recent study was conducted to investigate the psychometric properties of a popular job satisfaction measure, the MSQ short form, using a polytomous IRT model. Morse and Griffeth (2009) subjected the MSQ to an IRT analysis based on a recent debate regarding the item structure of its two dominant factors, intrinsic and extrinsic satisfaction. The results of their analysis suggested that the MSQ had peak reliability at the lower end of the construct continuum. This would indicate that this measure of job satisfaction may be predisposed to assessment inappropriateness if the individuals responding to it were not primarily low-satisfaction employees. Given other characteristics that were identified in the Morse and Griffeth study, such as the average item discrimination (.638 and .682 for the two sub-scales) and the number of items in the scale, we could reference conditions 5, 7, 29, or 31 in Tables 4 and 5 as possible Type I error rates (note that the possible conditions are reflecting unknown sample sizes and regression coefficients in a hypothetical MMR analysis). For raw scores, all of these conditions except for condition 5 would result in higher than acceptable Type I error rates for the MSQ.

Rescaling Data with IRT

The second argument for implementing IRT in applied psychology deals with our treatment of raw data from polytomous latent construct assessments. The results of this dissertation overwhelmingly indicate that estimated theta scores from the GRM resulted in better overall accuracy than raw scores in MMR models. Although there were some

cases in which the empirical Type I error rate for the estimated theta scores exceeded the acceptable interval, these rates were always a magnitude of order smaller than those for the raw scores.

These results also suggest that there is a more complex relationship underlying data structures and the assessment of moderators in MMR analyses than perhaps previously thought. Paunonen and Jackson (1988) conducted a simulation in which Type I error rates were compared between ordinary least squares regression (OLS) and principle components regression (PCR) in relation to the multicollinearity of the predictors. Their results indicated that OLS performed much better than PCR with regard to accurate moderator detection, and that linear transformations of the data had little effect on the Type I error rates for either procedure. In their study, the researchers simulated random effects data from normal distributions just as in this dissertation, but did not investigate any influences of psychometric characteristics on the data (i.e., difficulty, discrimination, assessment appropriateness, etc.). Conceptually, Paunonen and Jackson (1988) generated data as if they were able to collect actual theta scores. It is not surprising, therefore, that their results were well within the normal Type I error rate for MMR. An examination of the empirical Type I error rates for the actual theta scores in Tables 4, 5, 6, and 7 replicate these findings. The generalizability of the results from Paunonen and Jackson's simulation are severely restricted, and the results of this dissertation indicate that there is a more complex story to tell. Specifically, psychometric characteristics such as assessment appropriateness appear to have a significant influence on the performance of MMR analyses based on how the data is operationalized.

Given these results, researchers in applied psychology can use IRT models to more appropriately operationalize latent construct assessments and to increase the accuracy of parametric analyses. IRT models, therefore, can be used as a data management tool as well as a psychometric evaluation technique. Past research that has incorporated IRT into applied psychological research has often focused on narrow measurement applications (c.f., Facteau & Craig, 2001; Zickar et al., 2004). However, the results reported in this dissertation provide some support for a more generalized application of IRT. Specifically, under certain conditions such as assessment inappropriateness, strong regression weights, and high item discrimination, estimated theta scores were found to be better approximations of a latent construct than raw scores. The use of IRT as a "rescaling technique" in these conditions would be an advantage for researchers. However, based on the results identified in this dissertation, using an IRT model to rescale raw data is not necessary under conditions of assessment appropriateness. This finding provides researchers with the option, in some cases, to use raw scores with little or no consequences to the results of a moderated statistical analysis. Given the relative ease of generating raw scores compared to estimated theta scores, this result is noteworthy in terms of the balance between research quality and practicality.

Limitations and Future Research

Several limitations were present in this dissertation. First, the simulation that was conducted was designed to investigate the prevalence of Type I error rates for the interaction term in a moderated multiple regression analysis. A Type I error occurs when a researcher erroneously rejects a null Hypothesis, and is possible due to probability-

based statistical decision-making. A related problem is a Type II error, or failing to reject a null Hypothesis when a difference actually exists. There is ongoing debate as to the relative importance of these errors. Some argue that scientists in general have been more concerned with Type I errors for fear of appearing too "loose" with the conclusions that are drawn and presented to the general public (Rosenthal & Rosnow, 2008). Thus, for the purposes of protecting the integrity of our work, we may want to be overly cautious. However, especially in applied psychology, many researchers focus predominantly on Type II errors because of statistical artifacts that can preclude the detection of significant results (Aguinis & Stone-Romero, 1997; McClelland & Judd, 1993; Rogers, 2002; Stone, 1988; Zedeck, 1971). In investigations of moderators, this position has been especially salient due to the historical difficulty of identifying interaction effects that were strongly theorized to exist. Due to the design of this simulation, it was not possible to investigate the relationships between response score scaling and psychometric characteristics on Type II errors.

However, one study has provided some evidence related to the question of Type II errors. In her investigation of measurement effects on interactions in factorial ANOVA, Embretson (1996) found that the same conditions that give rise to Type I errors contribute similarly to Type II errors. Therefore, psychometric characteristics can give rise to a sensitivity effect (Type I errors) as well as a dampening effect (Type II errors) with regard to statistical decisions about interactions. Limitations in Embretson's study such as the application of the most restrictive IRT model, the Rasch model, to dichotomously scored data reduces the generalizability of these results. A possible avenue for future

research may include attempting to replicate these results in alternative analyses, such as MMR, with more complex IRT models, such as the two-parameter logistic model and the GRM. These models have been found, in many cases, to be more widely applicable to real data due to the inclusion of discrimination paramters and polytomous response categories respectively (Embretson & Reise, 2000). Together, the popularity of MMR, the heavy reliance on latent construct assessment data, and the strong concerns for Type II errors for interaction effects is sufficient cause to warrant this investigation.

The influence of distributional non-normality is also an issue that raises interesting follow-up possibilities. A common assumption of parametric statistical tests is that data is normally distributed around the mean of a particular variable (Rosenthal & Rosnow, 2008; Tabachnick & Fidell, 2007). Most statistics textbooks will provide a proof related to the central limit theorem assuring researchers that, regardless of the true shape of the parent population distribution, the sampling distribution of some statistic (usually the mean) will approach normality as sample size increases. Violations of normality can, in some cases, cause problems for the conclusions that are drawn from parametric tests because the critical distribution regions associated with Hypothesis evaluation are modeled from normally distributed data. In this dissertation, as well as in the Kang and Waller (2005) study, non-normality was observed in the simulated variables, and departures from normality were observed in a consistent, positive manner with the empirical Type I error rate. Additionally, the results of the Shapiro-Wilk test for the normality of the residuals, an assumption of MMR, indicated that cases of residual non-normality also increased with the empirical Type I error rate. These findings create a

potential confound such that the Type I error rate may be partially related to these violations of normality, which were observed as a result of the simulation rather than a controlled factor. As a future study, it may be helpful to manipulate these conditions to determine their causal impact.

Kang and Waller (2005) briefly delved into possible transformations for non-normality (e.g., the Box-Cox transformation) that could help attenuate the empirical Type I error rate, but their treatment of this issue was very preliminary. Given that similar patterns in the data emerged in this study, a more rigorous simulation could be conducted in this area. Specifically, it would be helpful to compare various methods of data transformations in situations where inflated Type I error rates are known to exist. The information from these studies could arm researchers with varying techniques, including IRT scaling that have known efficacies for preventing unreasonably high Type I error rates in defined contexts.

Another methodological limitation that could be raised for the simulations that were conducted in this dissertation relates to the use of the GRM to generate raw scores and to derive estimated theta scores. Recall that the simulated raw score matrices were derived using the GRM to impose particular psychometric factors and reflect responses to a Likert-type scale. The GRM was then invoked using *PARSCALE* to derive the estimated theta values from the raw score matrices. This approach is often taken in IRT Monte Carlo studies to increase the likelihood of model fit (Harwell et al., 1996). However, interesting follow-up studies may include using different models to generate and fit the data. Indeed, this approach may provide more insight into the degree to which

various IRT models reach interval scales of measurement. If the data are generated with the Rasch model (which is thought to best achieve interval-level scaling) and fitted with a more complex model, the degree of model fit could be another metric of interval-level scaling. This seems to be a particularly interesting avenue of future research.

Finally, two limitations were present in relation to the simulated variables in this dissertation. Recall that the two main effects in the simulated regression models were randomly selected from a standard normal distribution. Conceptually, this creates a regression model using continuous variables as the predictors. These are also referred to as random-effects, because the values are assumed to be randomly sampled from some larger distribution (Cohen et al., 2003). Examples of random-effects variables can include latent constructs such as intelligence or job-satisfaction in which each individual has a construct score and is assumed to be a member of a larger population. Another type of variable that can be included in an MMR analysis is a fixed-effect, or a discreet condition that is set in the model. Fixed-effects can also be thought of as categorical predictors such as manipulated treatment conditions. A primary distinction between the two types of variables is that random-effects make assumptions about parent distributions for both the variable and its associated error, whereas fixed-effects do not (Fisicaro & Tisak, 1994). The decision to simulate random-effects in this dissertation was made to retain as much comparability with the Kang and Waller (2005) study as possible. However, to increase the generalizability of these findings, it would be useful to also investigate fixed-effects due to the popularity of categorical variables in MMR in applied psychological research (Aguinis, 2004; Aguinis et al., 2005).

A more obscure but perhaps noteworthy concern related to the random/fixed-effects variable structure arises in relation to testing interaction terms in MMR. In two mathematical elaborations, Fisicaro and Tisak (1994) and Sockloff (1976) argued that random effects predictors are mathematically constrained such that an interaction effect cannot be computed. Therefore, they concluded that all significant interactions arising from random effects are Type I errors. Based on this logic, the results of this dissertation would not be meaningful due to this constraint. However, little follow-up evidence has been offered to further justify these claims, and seminal references such as the Cohen et al. (2003) regression text devote entire chapters to random effects and interactions between random effects predictors. Although important to note as a possible point of future study, it appears that the popular applied consensus has not placed much weight on this issue.

Conclusion

Overall, the results of this dissertation provide some support for the use of model-based measurement techniques for both general scale evaluation and the operationalization of latent constructs. However, this support is not ubiquitous and, in deference to parsimony, the generation of estimated theta scores should not be taken as the default data-handling method in all situations. Specifically, although there was some evidence for the influence of all of the manipulated psychometric characteristics in this dissertation, it is clear that assessment appropriateness is the most important factor. If a researcher can justify that the reliability of the assessment and the distribution of construct scores for the individuals are reasonably matched, there is no evidence here that

the Type I error rate will be inflated to an unacceptable level for any scoring technique. Although software packages developed to generate IRT-derived theta scores have greatly reduced the computational complexity that has long restricted their use, the use of simple raw scores is still more straightforward and sufficient in these cases.

The finding that assessment appropriateness was the most impactful psychometric factor for spurious interaction effects again raises the need for more widespread use of model-based measurement practices in applied psychology. Specifically, the generation of scale information curves that can clearly depict the reliability profile of an assessment is a key factor in fulfilling this need. However, only one study has examined a commonly used latent construct assessment in applied psychology for these properties (c.f., Morse & Griffeth, 2009). Based on the results of this dissertation, the need for further inquiry into this area is undeniable.

Finally, it was clear from the results of these simulations that the GRM was not fully resistant to inflated Type I error rates. In some cases, the empirical Type I error rates that were observed in this dissertation were inflated above the acceptable criterion for the estimated theta scores. This also suggests that the GRM may not have fully achieved interval-level scaling. However, it is important to remember that the estimated theta scores performed consistently better than the raw scores in every condition with an unacceptably high Type I error rate. In many cases, this performance difference was substantial. Therefore, although not perfect, the GRM did produce scores that were more robust than the raw scores for use in moderated multiple regression models.

Perhaps the most general message being promoted through the research in this dissertation is that we cannot lose sight of the importance and concrete implications of good measurement practices in psychological research. One would be hard-pressed to find a psychological researcher who disagrees with this sentiment in principle, but the published track record tells a different tale. It is therefore important to maintain the visibility of the importance and applicability of modern measurement theory and research in psychology and psychology-related fields. Embretson (2006) illustrated this position very effectively by stating,

> Applications of model-based measurement are rapidly increasing in the testing industry, but applications to psychological research are lagging. Although a wide variety of models with explanatory potential are now available and accessible through popular software, they will not be applied effectively unless psychologists are better prepared in measurement and statistics. Meeting this challenge will require a refocusing of efforts on several levels in the training of psychological researchers (p. 54).

The results of this dissertation provide a very specific indicator of the costs of measurement ambivalence on popularly used parametric analyses. Given that psychometric quality has been found to extend beyond the question of construct definitions and into the accuracy of parametric analyses, the need for increased attention to measurement seems especially salient. Hopefully, further research and exposure in fields like applied psychology will help elucidate these issues and improve the overall quality of behavioral research.

REFERENCES

Aguinis, H. (2004). *Regression Analysis for Categorical Moderators*. New York, NY: Guilford.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, 90(1), 94 – 107.

Aguinis, H. & Pierce, C. A. (2008). Enhancing the relevance of organizational behavior by embracing performance management research. *Journal of Organizational Behavior*, 29, 139 – 145.

Aguinis, H., Pierce, C. A., Bosco, F. A., & Muslin, I. S. (2009). First decade of Organizational Research Methods: Trends in design, measurement, and data-analysis topics. *Organizational Research Methods*, 12(1), 69 – 112.

Aguinis, H. & Stone-Romero, F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology*, *82*(1), 192-206.

Aiken, L. S. & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Thousand Oaks, CA: Sage Publications, Inc.

Anderson, N. H. (1961). Scales and statistics: Parametric and nonparametric. *Psychological Bulletin*, 58, 305 – 316.

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

Austin, J. T., Scherbaum, C. A., & Mahlman, R. A. (2002). History of research methods in industrial and organizational psychology: Measurement, design, analysis. In S. G. Rogelberg (Ed) *Handbook of research methods in industrial and organizational psychology*, (p. 3-33). Malden, MA, Blackwell Publishing.

Baker, F. B. & Kim, S. H. (2004). *Item response theory parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker Inc.

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology*, *31*(2), 233-241.

Borsboom, D. (2008). Latent variable theory. *Measurement*, 6, 25 – 53.

Bollen, K. & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin,* 110(2), 305 – 314.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, 26, 211 – 243.

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

Bradley, J. V. (1982). The insidious L-shaped distribution. *Bulletin of the Psychonomic Society*, 20(2), 85 – 88.

Bradley, J. V. (1984). The complexity of nonrobustness effects. *Bulletin of the Psychonomic Society*, 22(3), 250 – 253.

Brayfield, A. H. & Crockett, W. H. (1955). Employee attitudes and performance. *Psychological Bulletin*, 52(5), 396 – 424.

Brief, A. P. (1998). *Attitudes in and around organizations*. Thousand Oaks, CA: Sage Publications.

Burke, C. J. (1953). Additive scales and statistics. *Psychological Review*, 60, 73 – 75.

Busemeyer, J. R. (1980). Importance of measurement theory, error theory, and experimental design for testing the significance of interactions. *Psychological Bulletin*, *88*(1), 237-244.

Campbell, N. R. (1928). *An account of the principles of measurement and calculation*. London: Longmans & Green.

Cascio, W. F. & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Childs, R. A. & Chen, W. (1999). Obtaining comparable item parameter estimates in MULTILOG and PARSCALE for two polytomous IRT models. *Applied Psychological Measurement*, 23(4), 371 – 379.

Clogg, C. C. & Shihadeh, E. S. (1994). *Statistical models for ordinal variables*. Thousand Oaks, CA: Sage Publications.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: LEA.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98 – 104.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: HBJ Publishing.

Davison, M. L. & Sharma, R. (1988). Parametric statistics and levels of measurement. *Psychological Bulletin*, *104*(1), 137-144.

Davison, M. L & Sharma, R. (1990). Parametric statistics and levels of measurement: Factorial designs and multiple regression. *Psychological Bulletin*, *107*(3), 394-400.

Dawis, R. V. (1992). Person-environment fit and job satisfaction. In C. J. Cranny, P. C. Smith, & E. F. Stone (Eds.), *Job Satisfaction* (pp. 69 – 88). New York, NY: Lexington.

DeMars, C. E. (2002). Recovery of graded response and partial credit parameters in MULTILOG and PARSCALE. *Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.*

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, 13, 77 – 90.

Drasgow, F. & Hulin, C. L. (1991). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.) *Handbook of Industrial and Organizational Psychology*: Vol. 1 (2nd ed., p. 577-636). Palo Alto, CA: Consulting Psychologists Press.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*(3), 201-212.

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist*, 61(1), 50 – 55.

Embretson, S. E. & DeBoeck, P. (1994). Latent trait theory. In R. J. Sternberg (Ed.), *Encyclopedia of Intelligence* (p. 644 – 647). New York, NY: MacMillan.

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

Facteau, J. D. & Craig, B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology*, 86(2), 215-227.

Fields, D. L. (2002). *Taking the measure of work: A guide to validated scales for organizational research and diagnosis*. Thousand Oaks, CA: Sage Publications.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. Fischer & I. Molenaar (Eds.), *Rasch Models: Foundations, recent developments, and applications*. New York, NY: Springer-Verlag.

Fisicaro, S. A. & Tisak, J. (1994). A theoretical note on the stochastics of moderated multiple regression. *Educational and Psychological Measurement*, 54(1), 32 – 41

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, 78(2), 350 – 365.

Gagne, P., Furlow, C., & Ross, T. (2009). Increasing the number of replications in item response theory simulations. *Educational and Psychological Measurement*, 69(1), 79 – 84.

Gaito, J. (1959). Non-parametric methods in psychological research. *Psychological Reports*, 5, 115 – 120.

Gaito, J. (1960). Scale classification and statistics. *Psychological Review*, 67, 277 – 278.

Gardner, P. L. (1975). Scales and statistics. *Review of Educational Research*, 45(1), 43 – 57.

Gibbons, J. D. (1993). *Nonparametric statistics: An introduction*. Newbury Park, CA: Sage Publications.

Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of Management*, 26, 463 – 488.

Hackman, J. R. & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance*, 16, 250 – 179.

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

Harwell, M. & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, 71(1), 105 – 131.

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

Hattie, J. A. (1985). Methodological review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9, 139 – 164.

Hulin, C. L. (1966). Job satisfaction and turnover in a female clerical population. *Journal of Applied Psychology*, 50, 280 – 285.

Hulin, C. L. (1968). Effects of changes in job-satisfaction levels on employee turnover. *Journal of Applied Psychology*, 52, 122 – 126.

Hulin, C. L. & Ilgen D.R. (2000). Introduction to computational modeling in organizations: The good that modeling does. In C. L. Hulin and D. R. Ilgen (Eds.)

Computational Modeling of Behavior in Organizations (p. 3 – 18). Washington, D.C.: American Psychological Association.

Hulin, C. L. & Judge, T. A. (2003). Job attitudes. In I. B. Weiner (Series Ed.) & W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Vol. Eds.), *Handbook of Psychology: Vol. 12. Industrial and Organizational Psychology* (pp. 255 – 276). Hoboken, NJ: Wiley.

Hulin, C. L., & Mayer, J. (1986). Psychometric equivalence of a translation of the Job Descriptive Index into Hebrew. *Journal of Applied Psychology*, 71(1), 83-94.

Ihaka, R. & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5, 299 – 314.

Jacoby, J. (1978). How valid and useful are all our consumer behavior research findings. *Consumer Research: A State of the Art Review*, 42(2), 87 – 96.

Jaccard, J. & Turrisi, R. (2003). *Interaction effects in multiple regression*. Thousand Oaks, CA: Sage Publications.

Jenson, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Kang, T. & Chen, T. T. (2007, June). *An investigation of the performance of the generalized S-X^2 item-fit index for polytomous IRT models* (ACT Research Report Series 2007-1).

Kang, S. & Waller, G. (2005). Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement*, *29*(2), 87-105.

Lee, T. W. & Mitchell, T. R. (1994). An alternative approach: The unfolding model of voluntary turnover. *Academy of Management Review*, 19, 51 – 58.

Lee, T. W., Mitchell, T. R., Holtom, B. C., McDaniel, L. S., & Hill, J. W. (1999). The unfolding model of voluntary turnover: A replication and extension. *Academy of Management Journal*, 42(4), 450 – 462.

Lee, T.W., Mitchell, T.R., Wise, L., & Fireman, S. (1996). An unfolding model of voluntary employee turnover. *Academy of Management Journal, 39,* 5–36.

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

Lord, F. M. (1953a). On the statistical treatment of football numbers. *American Psychologist*, 8, 750 – 751.

Lord, F. M. (1953b). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13(4), 517 – 549.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 28, 989 – 1020.

Lubinski, D. & Humphreys, L. G. (1990). Assessing spurious "moderator effects": Illustrated substantively with the hypothesized ("synergistic") relation between spatial and mathematical ability. *Psychological Bulletin*, 107(3), 385 – 393.

Mair, P. & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, 20(9), 1 – 20.

March, J. G. & Simon, H. A. (1958). *Organizations*. New York: Wiley.

Maxwell, S. & Delaney, H. (1985). Measurement and statistics: An examination of construct validity. *Psychological Bulletin*, 97, 85 – 93.

McClelland, G. H. & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114, 376 – 390.

McNemar, Q. (1969). *Psychological statistics* (4th ed.). New York, NY: Wiley.

Meade, A. W., Lautenschlager, G. J., & Johnson, E. C. (2007). A Monte Carlo examination of the sensitivity of the differential functioning of items and tests framework for tests of measurement invariance with Likert data. *Applied Psychological Measurement*, 31(5), 430 – 455.

Mellenbergh, G. J. (1999). A note on simple gain score precision. *Applied Psychological Measurement*, 23, 87 – 89.

Mislevy, R. J. & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, 13, 57 – 75.

Mobley, W. H., Griffeth, R. W., Hand, H. H., & Meglino, B. M. (1979). Review and conceptual analysis of the employee turnover process. *Psychological Bulletin*, 86(3), 493 – 522.

Mobley, W. H., Horner, S., & Hollingsworth, A. (1978). An evaluation of precursors of hospital employee turnover. *Journal of Applied Psychology*, 63, 408 – 414.

Mooney, C. Z. (1997). *Monte Carlo Simulation*. Thousand Oaks, CA: Sage Publications.

Morris, J. H., Sherman, J. D., & Mansfield, E. R. (1986). Failures to detect moderating effects with ordinary least squares-moderated multiple regression: Some reasons and a remedy. *Psychological Bulletin*, 99, 282 – 288.

Morse, B. & Griffeth, R. W. (2009). An item response theory assessment of three intrinsic/extrinsic item-factor appropriations of the Minnesota Satisfaction

Questionnaire. (*manuscript presented at the 2009 Academy of Management Meeting, Chicago, IL).*

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

Muraki, E. & Bock, R. D. (2003). *PARSCALE: IRT item analysis and test scoring for rating-scale data* (Version 4.1) [computer software]. Lincolnwood, IL: Scientific Software International, Inc.

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

Partchev, I. (2007, May). *The irtoys package.* (R Technical Report).

Paunonen, S. V. & Jackson, D. N. (1988). Type I error rates for moderated multiple regression. *Journal of Applied Psychology*, 73(3), 569 – 573.

Phillips, J. S. & Lord, R. G. (1986). Notes on the practical and theoretical consequences of implicit leadership theories for the future of leadership measurement. *Journal of Management*, 12, 21 – 42.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, 3(2), 237 – 255.

Podsakoff, P. M. & Dalton, D. R. (1987). Research methodology in organizational studies. *Journal of Management*, 13, 419 – 441.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In M. Glegvad (Ed.), *The Danish Yearbook of Philosophy* (p. 58 – 94). Copenhagen: Munksgaard.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21(1), 25 – 36.

Reeve, C. L. & Smith, S. (2000). Refining Lodahl and Kejner's Job Involvement Scale with a convergent evidence approach: Applying multiple methods to multiple samples. *Organizational Research Methods*, 4(2), 91-111.

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science*, 14(2), 95 – 101.

Reise, S. P. & Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment*, 84(3), 228 – 238.

Reise, S. P. & Waller, G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, *8*(2), 164-184.

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.

Rosenthal, R. & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). Boston, MA: McGraw Hill.

Rupp, A. A. & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, 66(1), 63 – 84.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1-25.

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

Samejima, F. (1996). The graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*. New York, NY: Springer.

Scandura, T. A. & Williams, E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management Journal*, 43(6), 1248-1264.

Schoenfeldt, L. F. (1984). Psychometric properties of organizational research instruments. In T. S. Bateman and G. R. Ferris (Eds.) *Method and analysis in organizational research*. Reston, VA: Reston Publishing.

Schriesheim, C. A., Powers, K. J., Scandura, T. A., Gardiner, C. C., & Lankau, M. J. (1993). Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management*, 19(2), 385 – 417.

Schwab, D. P. (1980). Construct validity in organizational behavior. In B. M. Shaw & L. L. Cummings (Eds.) *Research in organizational behavior*, (p. 3 – 44), Greenwich, CT: JAI Press.

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591-611.

Smith, P. C. & Stanton, J. M. (1998). Perspectives on the measurement of job attitudes: The long view. *Human Resource Management Review*, 8(4), 367 – 386.

Society for Industrial and Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.).

Sockloff, A. L. (1976). The analysis of nonlinearity via linear regression with polynomial and product variables: An examination. *Review of Educational Research*, 46, 267 – 291.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677 – 680.

Stine, W. W. (1989). Meaningful inference: The role of measurement in statistics. *Psychological Bulletin*, 105(1), 147 – 155.

Stocking, M. L. (1987). Two simulated feasibility studies in computerised adaptive testing. *Applied Psychology: An International Review*, *36*(3-4), 263-277.

Stone, E. F. (1988). Moderator variables in research: A review and analysis of conceptual and methodological issues. In G. R. Ferris & K. M. Rowland (Eds.), *Research in personnel and human resources management*, *Vol. 6* (p. 191 – 229). Greenwich, CT: JAI.

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 – 6.

Stone-Romero, E. F. (1994). Construct validity issues in organizational behavior research. In J. Greenberg (Ed.) *Organizational behavior: The state of the science.* (1st Ed., p. 155-179). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Stone-Romero, E. F., Alliger, G. M., & Aguinis, H. (1994). Type II error problems in the use of moderated multiple regression for the detection of moderating effects of dichotomous variables. *Journal of Management*, 20(1), 167 – 178.

Swaminathan, H. & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, 50, 349 – 364.

Takane, Y. & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52, 392 – 408.

Thissen, D. & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin*, *104*(3), 385-395.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273 – 286.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33(4), 529 – 554.

Wainer, H. (1982). Robust statistics: A survey and some prescriptions. In G. Keren (Ed) *Statistical and methodological issues in psychology and social sciences research* (pp. 187 – 214). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Weiss, D. J., Dawis, R. V., England, G. W., & Lofquist, L. H. (1967). *Manual for the Minnesota Satisfaction Questionnaire*. Minneapolis, MN: Industrial Relations Center – University of Minnesota.

Witt, L. A. (1998). Enhancing organizational goal congruence: A solution to organizational politics. *Journal of Applied Psychology*, 83(4), 666-674.

Zagorsek, H., Stough, S. J., & Jaklic, M. (2006). Analysis of the reliability of the Leadership Practices Inventory in the item response theory framework. *International Journal of Selection and Assessment*, 14(2), 180-191.

Zedeck, S. (1971). Problems with the use of "moderator" variables. *Psychological Bulletin*, 76, 295 – 310.

Zickar, M. J. (1998). Modeling item-level data with item response theory. *Current Directions in Psychological Science*, 7(4), 104 – 109.

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods*, 7(2), 168-190.

Zickar, M. J. & Slaughter, J. E. (2002). Computational modeling. In S. Rogelberg (Ed) *Handbook of Research Methods in Industrial and Organizational Psychology* (p. 184 – 197). Malden, MA: Blackwell Publishing.

APPENDIX A: R CODE FOR SIMULATION 1

```
#Morse Dissertation Table 1 (n=250, normal)

#Load latent trait model library
library("ltm")

#Set number of iterations per condition
n.it<-1000

#Individual Monte Carlo loop structure
study1<-function(seednum, numSubj=numSubj, Numiter=n.it,
b.mean, b.sd, a.low, a.high, w1, w2, numItem, results.file)

#=================================================================#
#Simulation loops for spurious interactions (n=250, k=15, normal) #
#=================================================================#

{

setwd("C:/Program Files/PARSCALE4")

#Generate raw response matrix for IV1, IV2, and DV
score.item.prg<-function(numItem,numSubj,Ptheta,a,b,score,theta)
{
b1<-b
b2<-b1+.70
b3<-b2+.70
b4<-b3+.70

for(i in 1:numItem){

Ptheta1[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b1[i]))) #CBRF 1
Ptheta2[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b2[i]))) #CBRF 2
Ptheta3[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b3[i]))) #CBRF 3
Ptheta4[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b4[i]))) #CBRF 4

Ptheta1a[, i]<-1.0 - Ptheta1[, i] #CRF for option 1
Ptheta2b[, i]<-Ptheta1[, i] - Ptheta2[, i] #CRF for option 2
Ptheta3c[, i]<-Ptheta2[, i] - Ptheta3[, i] #CRF for option 3
Ptheta4d[, i]<-Ptheta3[, i] - Ptheta4[, i] #CRF for option 4
Ptheta5e[, i]<-Ptheta4[, i] #CRF for option 5

#Generating a response matrix by comparing a random value from a
#uniform distribution U(0,1) to the relative score categories
r<-runif(numSubj)
response1[,i]<-ifelse(r < Ptheta1a[,i],1,0)
response2[,i]<-ifelse(r < Ptheta1a[,i] + Ptheta2b[,i] & r >=
Ptheta1a[,i],2,0)
response3[,i]<-ifelse(r<Ptheta1a[,i]+Ptheta2b[,i]+Ptheta3c[,i]&r >=
Ptheta1a[,i]+Ptheta2b[,i],3,0)
response4[,i]<-ifelse(r<Ptheta1a[,i]+Ptheta2b[,i] + Ptheta3c[,i] +
Ptheta4d[,i] & r >= Ptheta1a[,i] + Ptheta2b[,i] + Ptheta3c[,i],4,0)
response5[,i]<-ifelse(r >=Ptheta1a[,i]+Ptheta2b[,i]+ Ptheta3c[,i] +
```

```
53: Ptheta4d[,i],5,0)
54:
55: #Compiling the response matrix to object 'score'
56: score<-response1+response2+response3+response4+response5
57: }
58: return(score)
59: }
60:
61: #Function to calculate skewness
62: skew <- function (x)
63: {
64: sk <- function(xx) {
65: n <- length(xx)
66: mn <- mean(xx)
67: dif.x <- xx - mn
68: m2 <- sum(dif.x^2)/n
69: m3 <- sum(dif.x^3)/n
70: m3/(m2^(3/2))
71: }
72: if (ncol(x) == 1 || is.null(dim(x)))
73: return(sk(x))
74: else return(apply(x, 2, sk))
75: }
76:
77: #Function to calculate kurtosis
78: kurtosis <-function (x)
79: {
80: kt <- function(xx) {
81: n <- length(xx)
82: mn <- mean(xx)
83: dif.x <- xx - mn
84: m2 <- sum(dif.x^2)/n
85: m4 <- sum(dif.x^4)/n
86: (m4/m2^2) - 3
87: }
88: if (ncol(x) == 1 || is.null(dim(x)))
89: return(kt(x))
90: else return(apply(x, 2, kt))
91: }
92:
93: #------------------- fixed conditions ----------------------------#
94:
95: result <- matrix(0, nrow = Numiter, ncol = 9)
96:
97: #------------------- starting for-loop ---------------------------#
98:
99: iter<-0
100: good.iter<-1
101: while(good.iter <= Numiter) {
102:
103: iter<-iter+1
104: set.seed(seednum+iter)
105:
106: #----------------- initializing values -------------------------#
```

```
107:
108: #Create a person by item matrix for the scores of CBRF 1 through 4
109: Ptheta1 <- matrix(0, nrow = numSubj, ncol = numItem)
110: Ptheta2 <- matrix(0, nrow = numSubj, ncol = numItem)
111: Ptheta3 <- matrix(0, nrow = numSubj, ncol = numItem)
112: Ptheta4 <- matrix(0, nrow = numSubj, ncol = numItem)
113:
114: #Create a person x item matrix for the scores of CRF 1 through 5
115: Ptheta1a <- matrix(0, nrow = numSubj, ncol = numItem)
116: Ptheta2b <- matrix(0, nrow = numSubj, ncol = numItem)
117: Ptheta3c <- matrix(0, nrow = numSubj, ncol = numItem)
118: Ptheta4d <- matrix(0, nrow = numSubj, ncol = numItem)
119: Ptheta5e <- matrix(0, nrow = numSubj, ncol = numItem)
120:
121: #Create a person x item matrix for raw scores of cat 1 through 5
122: response1 <- matrix(0, nrow = numSubj, ncol = numItem)
123: response2 <- matrix(0, nrow = numSubj, ncol = numItem)
124: response3 <- matrix(0, nrow = numSubj, ncol = numItem)
125: response4 <- matrix(0, nrow = numSubj, ncol = numItem)
126: response5 <- matrix(0, nrow = numSubj, ncol = numItem)
127:
128: #Create the final person by item matrix of raw responses
129: score <- matrix(0, nrow = numSubj, ncol = numItem)
130:
131: #---------------- Generating IV1, IV2, and DV -------------------#
132:
133: theta1<-scale(rnorm(numSubj))
134: theta2<-scale(rnorm(numSubj))
135: theta3<-scale(w1*theta1+w2*theta2+sqrt(1-(w1^2+w2^2))*scale(rnorm(numSubj)))
136:
137: #--------------- Specifying item parameters ---------------------#
138:
139: a1 <- runif(numItem, a.low, a.high)
140: a2 <- runif(numItem, a.low, a.high)
141: a3 <- runif(numItem, a.low, a.high)
142: b1a <- rnorm(numItem, b.mean, b.sd)
143: b2a <- rnorm(numItem, b.mean, b.sd)
144: b3a <- rnorm(numItem, b.mean, b.sd)
145:
146: #--------------- Generating reponse patterns --------------------#
147:
148: score1<-score.item.prg(numItem,numSubj,Ptheta1,a1,b1a,score, theta1)
149: score2<-score.item.prg(numItem,numSubj,Ptheta1,a2,b2a,score, theta2)
150: score3<-score.item.prg(numItem,numSubj,Ptheta1,a3,b3a,score, theta3)
151:
152: #-------- Compute Cronbach's Alpha for reliability -------------#
153:
154: alpha1<-cronbach.alpha(score1)
155: alpha2<-cronbach.alpha(score2)
156: alpha3<-cronbach.alpha(score3)
157: alpha.score1<-alpha1$alpha
158: alpha.score2<-alpha2$alpha
159: alpha.score3<-alpha3$alpha
160:
```

```
#--------- Estimating parameters using PARSCALE4.1 -------------#

#Command to invoke PARSCALE to generate theta estimates
#Note that all files must be located in the PARSCALE directory

#-------------------------n=250, k=15----------------------------#
#-------------------------score 1-------------------------------#
score1psl<-data.frame(1001:1250,score1)
write.table(score1psl,"C:\\Program Files\\PARSCALE4\\testscore_15-250.dat",
sep="",row.names=FALSE,col.names=FALSE)
system("score15-250.bat",show.output.on.console = FALSE)
theta.ab1<-read.table("C:\\Program Files\\PARSCALE4\\testscore15-250.SCO",
head=F,fill=T)[(1:250)*2,7]
theta.ab1<-as.matrix(theta.ab1)
#-------------------------score 2-------------------------------#
score2psl<-data.frame(1001:1250,score2)
write.table(score2psl,"C:\\Program Files\\PARSCALE4\\testscore_15-250.dat",
sep="",row.names=FALSE,col.names=FALSE)
system("score15-250.bat",show.output.on.console = FALSE)
theta.ab2<-read.table("C:\\Program Files\\PARSCALE4\\testscore15-250.SCO",
head=F,fill=T)[(1:250)*2,7]
theta.ab2<-as.matrix(theta.ab2)
#-------------------------score 3-------------------------------#
score3psl<-data.frame(1001:1250,score3)
write.table(score3psl,"C:\\Program Files\\PARSCALE4\\testscore_15-
250.dat",
sep="",row.names=FALSE,col.names=FALSE)
system("score15-250.bat",show.output.on.console = FALSE)
theta.ab3<-read.table("C:\\Program Files\\PARSCALE4\\testscore15-250.SCO",
head=F,fill=T)[(1:250)*2,7]
theta.ab3<-as.matrix(theta.ab3)
#---------------------------------------------------------------#

#-- Computing the rmsq, total scores, skew and kurtosis----------#

rmsq<- sqrt( mean( (theta1 - theta.ab1)^2 ) )

score1 <- apply(score1, 1, mean)
score2 <- apply(score2, 1, mean)
score3 <- apply(score3, 1, mean)

score1.skew<-skew(score1)
score1.kurtosis<-kurtosis(score1)
score2.skew<-skew(score2)
score2.kurtosis<-kurtosis(score2)
score3.skew<-skew(score3)
score3.kurtosis<-kurtosis(score3)

theta.ab1skew<-skew(theta.ab1)
theta.ab1kurtosis<-kurtosis(theta.ab1)
theta.ab2skew<-skew(theta.ab2)
theta.ab2kurtosis<-kurtosis(theta.ab2)
theta.ab3skew<-skew(theta.ab3)
theta.ab3kurtosis<-kurtosis(theta.ab3)
```

```
214:
215: #--- Applying additive and multiplicative regression models -----#
216:
217: #Actual theta scores
218: theta.add<-lm(theta3~theta1+theta2)
219: theta.mul<-lm(theta3~theta1+theta2+I(theta1*theta2))
220:
221: #Raw scores
222: sum.add<-lm(score3~score1+score2)
223: sum.mul<-lm(score3~score1+score2+I(score1*score2))
224:
225: #Estimated theta scores
226: theta.ab.add<-lm(theta.ab3~theta.ab1+theta.ab2)
227: theta.ab.mul<-lm(theta.ab3~theta.ab1+theta.ab2+I(theta.ab1*theta.ab2))
228:
229: #-------- Shapiro-Wilk test for checking normality --------------#
230:
231: theta.orderres <- summary(theta.add)$res
232: sum.orderres <- summary(sum.add)$res
233: thetahat.orderres <- summary(theta.ab.add)$res
234:
235: swtheta.p <- round(shapiro.test(theta.orderres)$p,5)
236: swsum.p <- round(shapiro.test(sum.orderres)$p,5)
237: swthetahat.p <- round(shapiro.test(thetahat.orderres)$p,5)
238:
239: #report r-square, r-square change sig., and rm squared deviations#
240:
241: cat("Working on sample",seednum,"iteration",iter,good.iter, "\n")
242: theta.add.rsq <- summary(theta.add)$r.squared
243: theta.mul.rsq <- summary(theta.mul)$r.squared
244: theta.p <- anova(theta.add, theta.mul)$"Pr(>F)"[2]
245: theta.p[is.na(theta.p)] <- 1.00
246:
247: sum.add.rsq <- summary(sum.add)$r.squared
248: sum.mul.rsq <- summary(sum.mul)$r.squared
249: sum.p <- round(anova(sum.add, sum.mul)$"Pr(>F)"[2], 4)
250: sum.p[is.na(sum.p)] <- 1.00
251:
252: theta.ab.add.rsq <- summary(theta.ab.add)$r.squared
253: theta.ab.mul.rsq <- summary(theta.ab.mul)$r.squared
254: theta.ab.p<-round(anova(theta.ab.add, theta.ab.mul)$"Pr(>F)"[2],
4)
255: theta.ab.p[is.na(theta.ab.p)] <- 1.00
256:
257: #--------- Summarize results of each loop ----------------------#
258:
259: iter.results<-as.vector(c(iter,
260: seednum,
261: numItem,
262: a.low,
263: a.high,
264: b.mean,
265: b.sd,
266: w1,
```

```
267: w2,
268: theta.add.rsq,
269: theta.mul.rsq,
270: theta.p,
271: sum.add.rsq,
272: sum.mul.rsq,
273: sum.p,
274: theta.ab.add.rsq,
275: theta.ab.mul.rsq,
276: theta.ab.p,
277: rmsq,
278: alpha.score1,
279: alpha.score2,
280: alpha.score3,
281: swtheta.p,
282: swsum.p,
283: swthetahat.p,
284: score1.skew,
285: score2.skew,
286: score3.skew,
287: score1.kurtosis,
288: score2.kurtosis,
289: score3.kurtosis,
290: theta.ab1skew,
291: theta.ab2skew,
292: theta.ab3skew,
293: theta.ab1kurtosis,
294: theta.ab2kurtosis,
295: theta.ab3kurtosis))
296:
297: names(iter.results)<-NULL
298: sink(results.file,append=TRUE)
299: print(iter.results,digits=4,quote=FALSE)
300: sink()
301: good.iter<-good.iter+1
302: }
303:
304: }
305: #-------------------- End loop structure ----------------------#
306:
307: #================================================================#
308: # Begin looping individual conditions #
309: #================================================================#
310:
311: options(width=2000)
312:
313: {
314:
315: results.file1<-"C:/Documents and Settings/Admin/Desktop/DissModel/C1.txt"
316: study1(seednum = 1,
317: numSubj = 250,
318: Numiter = n.it,
319: b.mean = -2.5,
320: b.sd = 0.7,
```

```
a.low = .31,
a.high = .58,
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file1)

results.file3<-"C:/Documents and Settings/Admin/Desktop/DissModel/C3.txt"
study1(seednum = 3,
numSubj = 250,
Numiter = n.it,
b.mean = -2.5,
b.sd = 0.7,
a.low = .31,
a.high = .58,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file3)

results.file5<-"C:/Documents and Settings/Admin/Desktop/DissModel/C5.txt"
study1(seednum = 5,
numSubj = 250,
Numiter = n.it,
b.mean = -2.5,
b.sd = 0.7,
a.low = .58,
a.high = 1.13,
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file5)

results.file7 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C7.txt"
study1(seednum = 7,
numSubj = 250,
Numiter = n.it,
b.mean = -2.5,
b.sd = 0.7,
a.low = .58,
a.high = 1.13,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file7)

results.file9 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C9.txt"
study1(seednum = 9,
numSubj = 250,
Numiter = n.it,
b.mean = -1.0,
b.sd = 0.7,
a.low = .31,
a.high = .58,
```

```
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file9)

results.file11<-"C:/Documents and Settings/Admin/Desktop/DissModel/C11.txt"
study1(seednum = 11,
numSubj = 250,
Numiter = n.it,
b.mean = -1.0,
b.sd = 0.7,
a.low = .31,
a.high = .58,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file11)

results.file13<-"C:/Documents and Settings/Admin/Desktop/DissModel/C13.txt"
study1(seednum = 13,
numSubj = 250,
Numiter = n.it,
b.mean = -1.0,
b.sd = 0.7,
a.low = .58,
a.high = 1.13,
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file13)

results.file15 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C15.txt"
study1(seednum = 15,
numSubj = 250,
Numiter = n.it,
b.mean = -1.0,
b.sd = 0.7,
a.low = .58,
a.high = 1.13,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file15)

results.file17 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C17.txt"
study1(seednum = 17,
numSubj = 250,
Numiter = n.it,
b.mean = 0.5,
b.sd = 0.7,
a.low = .31,
a.high = .58,
w1 = .3,
w2 = .3,
```

```
numItem = 15,
results.file = results.file17)

results.file19 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C19.txt"
study1(seednum = 19,
numSubj = 250,
Numiter = n.it,
b.mean = 0.5,
b.sd = 0.7,
a.low = .31,
a.high = .58,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file19)

results.file21 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C21.txt"
study1(seednum = 21,
numSubj = 250,
Numiter = n.it,
b.mean = 0.5,
b.sd = 0.7,
a.low = .58,
a.high = 1.13,
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file21)

results.file23 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C23.txt"
study1(seednum = 23,
numSubj = 250,
Numiter = n.it,
b.mean = 0.5,
b.sd = 0.7,
a.low = .58,
a.high = 1.13,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file23)

}


#=====================================================================#
#Simulation loops for spurious interactions (n=250,k=30,normal)#
#=====================================================================#

{

setwd("C:/Program Files/PARSCALE4")

```

```
#Generate raw response matrix for IV1, IV2, and DV
score.item.prg<-function(numItem,numSubj,Ptheta,a,b,score,theta)
{
b1<-b
b2<-b1+.70
b3<-b2+.70
b4<-b3+.70

for(i in 1:numItem){

Ptheta1[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b1[i]))) #CBRF 1
Ptheta2[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b2[i]))) #CBRF 2
Ptheta3[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b3[i]))) #CBRF 3
Ptheta4[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b4[i]))) #CBRF 4

Ptheta1a[, i]<-1.0 - Ptheta1[, i] #CRF for option 1
Ptheta2b[, i]<-Ptheta1[, i] - Ptheta2[, i] #CRF for option 2
Ptheta3c[, i]<-Ptheta2[, i] - Ptheta3[, i] #CRF for option 3
Ptheta4d[, i]<-Ptheta3[, i] - Ptheta4[, i] #CRF for option 4
Ptheta5e[, i]<-Ptheta4[, i] #CRF for option 5

#Generating a response matrix by comparing a random value from a
#uniform distribution U(0,1) to the relative score categories
r<-runif(numSubj)
response1[,i]<-ifelse(r < Ptheta1a[,i],1,0)
response2[,i]<-ifelse(r < Ptheta1a[,i]+Ptheta2b[,i] & r >=
Ptheta1a[,i],2,0)
response3[,i]<-ifelse(r<Ptheta1a[,i]+Ptheta2b[,i]+Ptheta3c[,i]&r >=
Ptheta1a[,i] + Ptheta2b[,i],3,0)
response4[,i]<-ifelse(r < Ptheta1a[,i] + Ptheta2b[,i] + Ptheta3c[,i]+
Ptheta4d[,i] & r >= Ptheta1a[,i] + Ptheta2b[,i] + Ptheta3c[,i],4,0)
response5[,i]<-ifelse(r >= Ptheta1a[,i] + Ptheta2b[,i] + Ptheta3c[,i] +
Ptheta4d[,i],5,0)

#Compiling the response matrix to object 'score'
score<-response1+response2+response3+response4+response5
}
return(score)
}

#Function to calculate skewness
skew <- function (x)
{
sk <- function(xx) {
n <- length(xx)
mn <- mean(xx)
dif.x <- xx - mn
m2 <- sum(dif.x^2)/n
m3 <- sum(dif.x^3)/n
m3/(m2^(3/2))
}
if (ncol(x) == 1 || is.null(dim(x)))
return(sk(x))
else return(apply(x, 2, sk))
```

```
535: }
536:
537: #Function to calculate kurtosis
538: kurtosis <-function (x)
539: {
540: kt <- function(xx) {
541: n <- length(xx)
542: mn <- mean(xx)
543: dif.x <- xx - mn
544: m2 <- sum(dif.x^2)/n
545: m4 <- sum(dif.x^4)/n
546: (m4/m2^2) - 3
547: }
548: if (ncol(x) == 1 || is.null(dim(x)))
549: return(kt(x))
550: else return(apply(x, 2, kt))
551: }
552:
553: #---------------------- fixed conditions -------------------------#
554:
555: result <- matrix(0, nrow = Numiter, ncol = 9)
556:
557: #--------------------- starting for-loop -------------------------#
558:
559: iter<-0
560: good.iter<-1
561: while(good.iter <= Numiter) {
562:
563: iter<-iter+1
564: set.seed(seednum+iter)
565:
566: #------------------- initializing values ------------------------#
567:
568: #Create a person by item matrix for the scores of CBRF 1 through 4
569: Ptheta1 <- matrix(0, nrow = numSubj, ncol = numItem)
570: Ptheta2 <- matrix(0, nrow = numSubj, ncol = numItem)
571: Ptheta3 <- matrix(0, nrow = numSubj, ncol = numItem)
572: Ptheta4 <- matrix(0, nrow = numSubj, ncol = numItem)
573:
574: #Create a person x item matrix for the scores of CRF 1 through 5
575: Ptheta1a <- matrix(0, nrow = numSubj, ncol = numItem)
576: Ptheta2b <- matrix(0, nrow = numSubj, ncol = numItem)
577: Ptheta3c <- matrix(0, nrow = numSubj, ncol = numItem)
578: Ptheta4d <- matrix(0, nrow = numSubj, ncol = numItem)
579: Ptheta5e <- matrix(0, nrow = numSubj, ncol = numItem)
580:
581: #Create a person x item matrix for raw scores of cat 1 through 5
582: response1 <- matrix(0, nrow = numSubj, ncol = numItem)
583: response2 <- matrix(0, nrow = numSubj, ncol = numItem)
584: response3 <- matrix(0, nrow = numSubj, ncol = numItem)
585: response4 <- matrix(0, nrow = numSubj, ncol = numItem)
586: response5 <- matrix(0, nrow = numSubj, ncol = numItem)
587:
588: #Create the final person by item matrix of raw responses
```

```
589: score <- matrix(0, nrow = numSubj, ncol = numItem)
590:
591: #------------- Generating IV1, IV2, and DV ----------------------#
592:
593: theta1<-scale(rnorm(numSubj))
594: theta2<-scale(rnorm(numSubj))
595: theta3<-scale(w1*theta1+w2*theta2+sqrt(1-(w1^2+w2^2))*scale(rnorm(numSubj)))
596:
597: #----------- Specifying item parameters ------------------------#
598:
599: a1 <- runif(numItem, a.low, a.high)
600: a2 <- runif(numItem, a.low, a.high)
601: a3 <- runif(numItem, a.low, a.high)
602: b1a <- rnorm(numItem, b.mean, b.sd)
603: b2a <- rnorm(numItem, b.mean, b.sd)
604: b3a <- rnorm(numItem, b.mean, b.sd)
605:
606: #----------- Generating reponse patterns -----------------------#
607:
608: score1<-score.item.prg(numItem,numSubj,Ptheta1,a1,b1a,score,theta1)
609: score2<-score.item.prg(numItem,numSubj, Ptheta1, a2, b2a, score, theta2)
610: score3<-score.item.prg(numItem,numSubj, Ptheta1, a3, b3a, score, theta3)
611:
612: #--------- Compute Cronbach's Alpha for reliability -------------#
613:
614: alpha1<-cronbach.alpha(score1)
615: alpha2<-cronbach.alpha(score2)
616: alpha3<-cronbach.alpha(score3)
617: alpha.score1<-alpha1$alpha
618: alpha.score2<-alpha2$alpha
619: alpha.score3<-alpha3$alpha
620:
621: #---------- Estimating parameters using PARSCALE4.1 -------------#
622:
623: #Command to invoke PARSCALE to generate theta estimates
624: #Note that all files must be located in the PARSCALE directory
625:
626: #-------------------------n=250, k=30----------------------------#
627: #----------------------------score 1-----------------------------#
628: score1psl<-data.frame(1001:1250,score1)
629: write.table(score1psl,"C:\\Program Files\\PARSCALE4\\testscore_30-250.dat",
630: sep="",row.names=FALSE,col.names=FALSE)
631: system("score30-250.bat",show.output.on.console = FALSE)
632: theta.ab1<-read.table("C:\\Program Files\\PARSCALE4\\testscore30-250.SCO",
633: head=F,fill=T)[(1:250)*2,7]
634: theta.ab1<-as.matrix(theta.ab1)
635: #----------------------------score 2-----------------------------#
636: score2psl<-data.frame(1001:1250,score2)
637: write.table(score2psl,"C:\\Program Files\\PARSCALE4\\testscore_30-250.dat",
638: sep="",row.names=FALSE,col.names=FALSE)
639: system("score30-250.bat",show.output.on.console = FALSE)
640: theta.ab2<-read.table("C:\\Program Files\\PARSCALE4\\testscore30-250.SCO",
641: head=F,fill=T)[(1:250)*2,7]
642: theta.ab2<-as.matrix(theta.ab2)
```

```
643: #---------------------------score 3-----------------------------#
644: score3psl<-data.frame(1001:1250,score3)
645: write.table(score3psl,"C:\\Program Files\\PARSCALE4\\testscore_30-250.dat",
646: sep="",row.names=FALSE,col.names=FALSE)
647: system("score30-250.bat",show.output.on.console = FALSE)
648: theta.ab3<-read.table("C:\\Program Files\\PARSCALE4\\testscore30-250.SCO",
649: head=F,fill=T)[(1:250)*2,7]
650: theta.ab3<-as.matrix(theta.ab3)
651: #-----------------------------------------------------------------#
652:
653: #--- Computing the rmsq, total scores, skew and kurtosis --------#
654:
655: rmsq<- sqrt( mean( (theta1 - theta.ab1)^2 ) )
656:
657: score1 <- apply(score1, 1, mean)
658: score2 <- apply(score2, 1, mean)
659: score3 <- apply(score3, 1, mean)
660:
661: score1.skew<-skew(score1)
662: score1.kurtosis<-kurtosis(score1)
663: score2.skew<-skew(score2)
664: score2.kurtosis<-kurtosis(score2)
665: score3.skew<-skew(score3)
666: score3.kurtosis<-kurtosis(score3)
667:
668: theta.ab1skew<-skew(theta.ab1)
669: theta.ab1kurtosis<-kurtosis(theta.ab1)
670: theta.ab2skew<-skew(theta.ab2)
671: theta.ab2kurtosis<-kurtosis(theta.ab2)
672: theta.ab3skew<-skew(theta.ab3)
673: theta.ab3kurtosis<-kurtosis(theta.ab3)
674:
675: #-- Applying additive and multiplicative regression models ------#
676:
677: #Actual theta scores
678: theta.add<-lm(theta3~theta1+theta2)
679: theta.mul<-lm(theta3~theta1+theta2+I(theta1*theta2))
680:
681: #Raw scores
682: sum.add<-lm(score3~score1+score2)
683: sum.mul<-lm(score3~score1+score2+I(score1*score2))
684:
685: #Estimated theta scores
686: theta.ab.add<-lm(theta.ab3~theta.ab1+theta.ab2)
687: theta.ab.mul<-lm(theta.ab3~theta.ab1+theta.ab2+I(theta.ab1*theta.ab2))
688:
689: #-------- Shapiro-Wilk test for checking normality --------------#
690:
691: theta.orderres <- summary(theta.add)$res
692: sum.orderres <- summary(sum.add)$res
693: thetahat.orderres <- summary(theta.ab.add)$res
694:
695: swtheta.p <- round(shapiro.test(theta.orderres)$p,5)
696: swsum.p <- round(shapiro.test(sum.orderres)$p,5)
```

```
swthetahat.p <- round(shapiro.test(thetahat.orderres)$p,5)

#report r-square, r-square change sig., and rm squared deviations#

cat("Working on sample", seednum,"iteration",iter,good.iter, "\n")
theta.add.rsq <- summary(theta.add)$r.squared
theta.mul.rsq <- summary(theta.mul)$r.squared
theta.p <- anova(theta.add, theta.mul)$"Pr(>F)"[2]
theta.p[is.na(theta.p)] <- 1.00

sum.add.rsq <- summary(sum.add)$r.squared
sum.mul.rsq <- summary(sum.mul)$r.squared
sum.p <- round(anova(sum.add, sum.mul)$"Pr(>F)"[2], 4)
sum.p[is.na(sum.p)] <- 1.00

theta.ab.add.rsq <- summary(theta.ab.add)$r.squared
theta.ab.mul.rsq <- summary(theta.ab.mul)$r.squared
theta.ab.p<-round(anova(theta.ab.add, theta.ab.mul)$"Pr(>F)"[2],
4)
theta.ab.p[is.na(theta.ab.p)] <- 1.00

#------------ Summarize results of each loop --------------------#

iter.results<-as.vector(c(iter,
seednum,
numItem,
a.low,
a.high,
b.mean,
b.sd,
w1,
w2,
theta.add.rsq,
theta.mul.rsq,
theta.p,
sum.add.rsq,
sum.mul.rsq,
sum.p,
theta.ab.add.rsq,
theta.ab.mul.rsq,
theta.ab.p,
rmsq,
alpha.score1,
alpha.score2,
alpha.score3,
swtheta.p,
swsum.p,
swthetahat.p,
score1.skew,
score2.skew,
score3.skew,
score1.kurtosis,
score2.kurtosis,
score3.kurtosis,
```

```
750: theta.ab1skew,
751: theta.ab2skew,
752: theta.ab3skew,
753: theta.ab1kurtosis,
754: theta.ab2kurtosis,
755: theta.ab3kurtosis))
756:
757: names(iter.results)<-NULL
758: sink(results.file,append=TRUE)
759: print(iter.results,digits=4,quote=FALSE)
760: sink()
761: good.iter<-good.iter+1
762: }
763:
764: }
765: #-------------------- End loop structure ----------------------#
766:
767: #=====================================================================#
768: # Begin looping individual conditions #
769: #=====================================================================#
770:
771: options(width=2000)
772:
773: {
774:
775: results.file2<-"C:/Documents and Settings/Admin/Desktop/DissModel/C2.txt"
776: study1(seednum = 2,
777: numSubj = 250,
778: Numiter = n.it,
779: b.mean = -2.5,
780: b.sd = 0.7,
781: a.low = .31,
782: a.high = .58,
783: w1 = .3,
784: w2 = .3,
785: numItem = 30,
786: results.file = results.file2)
787:
788: results.file4<-"C:/Documents and Settings/Admin/Desktop/DissModel/C4.txt"
789: study1(seednum = 4,
790: numSubj = 250,
791: Numiter = n.it,
792: b.mean = -2.5,
793: b.sd = 0.7,
794: a.low = .31,
795: a.high = .58,
796: w1 = .5,
797: w2 = .5,
798: numItem = 30,
799: results.file = results.file4)
800:
801: results.file6 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C6.txt"
802: study1(seednum = 6,
803: numSubj = 250,
```

```
804: Numiter = n.it,
805: b.mean = -2.5,
806: b.sd = 0.7,
807: a.low = .58,
808: a.high = 1.13,
809: w1 = .3,
810: w2 = .3,
811: numItem = 30,
812: results.file = results.file6)
813:
814: results.file8 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C8.txt"
815: study1(seednum = 8,
816: numSubj = 250,
817: Numiter = n.it,
818: b.mean = -2.5,
819: b.sd = 0.7,
820: a.low = .58,
821: a.high = 1.13,
822: w1 = .5,
823: w2 = .5,
824: numItem = 30,
825: results.file = results.file8)
826:
827: results.file10<-"C:/Documents and Settings/Admin/Desktop/DissModel/C10.txt"
828: study1(seednum = 10,
829: numSubj = 250,
830: Numiter = n.it,
831: b.mean = -1.0,
832: b.sd = 0.7,
833: a.low = .31,
834: a.high = .58,
835: w1 = .3,
836: w2 = .3,
837: numItem = 30,
838: results.file = results.file10)
839:
840: results.file12 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C12.txt"
841: study1(seednum = 12,
842: numSubj = 250,
843: Numiter = n.it,
844: b.mean = -1.0,
845: b.sd = 0.7,
846: a.low = .31,
847: a.high = .58,
848: w1 = .5,
849: w2 = .5,
850: numItem = 30,
851: results.file = results.file12)
852:
853: results.file14<-"C:/Documents and Settings/Admin/Desktop/DissModel/C14.txt"
854: study1(seednum = 14,
855: numSubj = 250,
856: Numiter = n.it,
857: b.mean = -1.0,
```

```
858: b.sd = 0.7,
859: a.low = .58,
860: a.high = 1.13,
861: w1 = .3,
862: w2 = .3,
863: numItem = 30,
864: results.file = results.file14)
865:
866: results.file16 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C16.txt"
867: study1(seednum = 16,
868: numSubj = 250,
869: Numiter = n.it,
870: b.mean = -1.0,
871: b.sd = 0.7,
872: a.low = .58,
873: a.high = 1.13,
874: w1 = .5,
875: w2 = .5,
876: numItem = 30,
877: results.file = results.file16)
878:
879: results.file18 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C18.txt"
880: study1(seednum = 18,
881: numSubj = 250,
882: Numiter = n.it,
883: b.mean = 0.5,
884: b.sd = 0.7,
885: a.low = .31,
886: a.high = .58,
887: w1 = .3,
888: w2 = .3,
889: numItem = 30,
890: results.file = results.file18)
891:
892: results.file20<-"C:/Documents and Settings/Admin/Desktop/DissModel/C20.txt"
893: study1(seednum = 20,
894: numSubj = 250,
895: Numiter = n.it,
896: b.mean = 0.5,
897: b.sd = 0.7,
898: a.low = .31,
899: a.high = .58,
900: w1 = .5,
901: w2 = .5,
902: numItem = 30,
903: results.file = results.file20)
904:
905: results.file22 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C22.txt"
906: study1(seednum = 22,
907: numSubj = 250,
908: Numiter = n.it,
909: b.mean = 0.5,
910: b.sd = 0.7,
911: a.low = .58,
```

```
912: a.high = 1.13,
913: w1 = .3,
914: w2 = .3,
915: numItem = 30,
916: results.file = results.file22)
917:
918: results.file24 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C24.txt"
919: study1(seednum = 24,
920: numSubj = 250,
921: Numiter = n.it,
922: b.mean = 0.5,
923: b.sd = 0.7,
924: a.low = .58,
925: a.high = 1.13,
926: w1 = .5,
927: w2 = .5,
928: numItem = 30,
929: results.file = results.file24)
930:
931: }
932:
933: #=================================================================#
934: # Summarize Results for Table 1 (n=250, normal) #
935: #=================================================================#
936:
937: n.it<-1000
938:
939: results.list<-paste("C:/Documents and Settings/Admin/Desktop/DissModel/C",
940: 1:24,sep="")
941: results.list<-paste(results.list,".txt",sep="")
942:
943: type1.theta<-rep(0,24)
944: type1.sum <- rep(0,24)
945: type1.thetahat <- rep(0,24)
946: rdiff.theta <- rep(0,24)
947: rdiff.sum <- rep(0,24)
948: rdiff.thetahat <- rep(0,24)
949: mn.rmsq<-rep(0,24)
950: pvalue.score1.mn<-rep(0,24)
951: pvalue.score2.mn<-rep(0,24)
952: pvalue.score3.mn<-rep(0,24)
953: pvalue.score1.sd<-rep(0,24)
954: pvalue.score2.sd<-rep(0,24)
955: pvalue.score3.sd<-rep(0,24)
956: alpha.score1<-rep(0,24)
957: alpha.score2<-rep(0,24)
958: alpha.score3<-rep(0,24)
959: sw.theta.p<-rep(0,24)
960: sw.sum.p<-rep(0,24)
961: sw.thetahat.p<-rep(0,24)
962:
963: skew.score1<-rep(0,24)
964: skew.score2<-rep(0,24)
965: skew.score3<-rep(0,24)
```

```
966:
967: kurtosis.score1<-rep(0,24)
968: kurtosis.score2<-rep(0,24)
969: kurtosis.score3<-rep(0,24)
970:
971: skew.theta.ab1<-rep(0,24)
972: skew.theta.ab2<-rep(0,24)
973: skew.theta.ab3<-rep(0,24)
974:
975: kurtosis.theta.ab1<-rep(0,24)
976: kurtosis.theta.ab2<-rep(0,24)
977: kurtosis.theta.ab3<-rep(0,24)
978:
979: for(i in 1:24) {
980: infile <- read.table(results.list[[i]], header=FALSE)
981: infile<- infile[,2:ncol(infile)]
982: names(infile)<-list( "iter",
983: "seednum",
984: "numItem",
985: "a.low",
986: "a.high",
987: "b.mean",
988: "b.sd",
989: "w1",
990: "w2",
991: "theta.add.rsq",
992: "theta.mul.rsq",
993: "theta.p",
994: "sum.add.rsq",
995: "sum.mul.rsq",
996: "sum.p",
997: "theta.ab.add.rsq",
998: "theta.ab.mul.rsq",
999: "theta.ab.p",
1000: "rmsq",
1001: "alpha.score1",
1002: "alpha.score2",
1003: "alpha.score3",
1004: "swtheta.p",
1005: "swsum.p",
1006: "swthetahat.p",
1007: "skew.score1",
1008: "skew.score2",
1009: "skew.score3",
1010: "kurtosis.score1",
1011: "kurtosis.score2",
1012: "kurtosis.score3",
1013: "skew.theta.ab1",
1014: "skew.theta.ab2",
1015: "skew.theta.ab3",
1016: "kurtosis.theta.ab1",
1017: "kurtosis.theta.ab2",
1018: "kurtosis.theta.ab3")
1019:
```

```
write.table(infile,
"C:/Documents and Settings/Admin/Desktop/DissModel/norm 250 full.txt")

type1.theta[i]<-sum(infile["theta.p"] <= .05)/n.it
type1.sum[i]<-sum(infile["sum.p"] <=.05)/n.it
type1.thetahat[i]<-sum(infile["theta.ab.p"] <=.05)/n.it

rdiff.theta[i]<-round(sum(infile["theta.mul.rsq"]-infile["theta.add.rsq"])/
n.it,2)
rdiff.sum[i]<-round(sum(infile["sum.mul.rsq"]-infile["sum.add.rsq"])/n.it,2)
rdiff.thetahat[i]<-round(sum(infile["theta.ab.mul.rsq"] -
infile["theta.ab.add.rsq"])/n.it,2)
mn.rmsq[i]<-round(mean(infile["rmsq"]),2)
alpha.score1[i]<-round(mean(infile["alpha.score1"]),2)
alpha.score2[i]<-round(mean(infile["alpha.score2"]),2)
alpha.score3[i]<-round(mean(infile["alpha.score3"]),2)

sw.theta.p[i]<-round(sum(infile["swtheta.p"] > .05)/n.it,5)
sw.sum.p[i]<-round(sum(infile["swsum.p"] > .05)/n.it,5)
sw.thetahat.p[i]<-round(sum(infile["swthetahat.p"] > .05)/n.it,5)

skew.score1[i]<-round(mean(infile["skew.score1"]),5)
skew.score2[i]<-round(mean(infile["skew.score2"]),5)
skew.score3[i]<-round(mean(infile["skew.score3"]),5)
kurtosis.score1[i]<-round(mean(infile["kurtosis.score1"]),5)
kurtosis.score2[i]<-round(mean(infile["kurtosis.score2"]),5)
kurtosis.score3[i]<-round(mean(infile["kurtosis.score3"]),5)
skew.theta.ab1[i]<-round(mean(infile["skew.theta.ab1"]),5)
skew.theta.ab2[i]<-round(mean(infile["skew.theta.ab2"]),5)
skew.theta.ab3[i]<-round(mean(infile["skew.theta.ab3"],na.rm=TRUE),5)
kurtosis.theta.ab1[i]<-round(mean(infile["kurtosis.theta.ab1"]),5)
kurtosis.theta.ab2[i]<-round(mean(infile["kurtosis.theta.ab2"]),5)
kurtosis.theta.ab3[i]<-round(mean(infile["kurtosis.theta.ab3"],na.rm=TRUE),5)

}

n <- c(rep(250,24))
b <- c(rep("N(-1.5,1.0)",8),rep("N(0,1)",8),rep("N(1.5,1.0)",8))
a <- c(rep("U(0.31, 0.58)",4),rep("U(0.58, 1.13)",4))
a <- rep(a,3)
B1B2 <- rep(c(.3,.3,.5,.5),6)
Items<-rep(c(15,30),12)

mean.alpha<-round(apply(cbind(alpha.score1,alpha.score2,alpha.score3),
1,mean),2)

type1.theta<-round(type1.theta,2)
type1.sum<-round(type1.sum,2)
type1.thetahat<-round(type1.thetahat,2)

sw.theta.p<-round(sw.theta.p,2)
sw.sum.p<-round(sw.sum.p,2)
sw.thetahat<-round(sw.thetahat.p,2)

```

```
1074: sktab1<-round(data.frame(skew.score3, kurtosis.score3, skew.theta.ab3,
1075: kurtosis.theta.ab3),2)
1076:
1077: table1<-data.frame(n,b,a,B1B2,Items,type1.theta,type1.sum,type1.thetahat,
1078: mean.alpha,sw.theta.p,sw.sum.p,sw.thetahat,sktab1)
1079:
1080: print(table1)
1081: write.table(table1,
1082: "C:/Documents and Settings/Admin/Desktop/DissModel/Table1 norm 250.txt")
1083: #=======End simulation for Table 1 (n=250, normal)=============#
```

APPENDIX B: R CODE FOR SIMULATION 2

```
#Morse Dissertation Table 2 (n=750, normal)

#Load latent trait model library
library("ltm")

#Set number of iterations per condition
n.it<-1000

#Individual Monte Carlo loop structure
study1<-function(seednum, numSubj=numSubj, Numiter=n.it,
b.mean, b.sd, a.low, a.high, w1, w2, numItem, results.file)

#=================================================================#
#Simulation loops for spurious interactions (n=750,k=15, normal)#
#=================================================================#

{

setwd("C:/Program Files/PARSCALE4")

#Generate raw response matrix for IV1, IV2, and DV
score.item.prg<-function(numItem, numSubj,Ptheta,a,b,score, theta)
{
b1<-b
b2<-b1+.70
b3<-b2+.70
b4<-b3+.70

for(i in 1:numItem){

Ptheta1[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b1[i]))) #CBRF 1
Ptheta2[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b2[i]))) #CBRF 2
Ptheta3[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b3[i]))) #CBRF 3
Ptheta4[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b4[i]))) #CBRF 4

Ptheta1a[, i]<-1.0 - Ptheta1[, i] #CRF for option 1
Ptheta2b[, i]<-Ptheta1[, i] - Ptheta2[, i] #CRF for option 2
Ptheta3c[, i]<-Ptheta2[, i] - Ptheta3[, i] #CRF for option 3
Ptheta4d[, i]<-Ptheta3[, i] - Ptheta4[, i] #CRF for option 4
Ptheta5e[, i]<-Ptheta4[, i] #CRF for option 5

#Generating a response matrix by comparing a random value from a
#uniform distribution U(0,1) to the relative score categories
r<-runif(numSubj)
response1[,i]<-ifelse(r < Ptheta1a[,i],1,0)
response2[,i]<-ifelse(r < Ptheta1a[,i] + Ptheta2b[,i] & r >=
Ptheta1a[,i],2,0)
response3[,i]<-ifelse(r<Ptheta1a[,i]+Ptheta2b[,i]+Ptheta3c[,i]&r >=
Ptheta1a[,i]+Ptheta2b[,i],3,0)
response4[,i]<-ifelse(r<Ptheta1a[,i]+Ptheta2b[,i] + Ptheta3c[,i] +
Ptheta4d[,i] & r >= Ptheta1a[,i] + Ptheta2b[,i] + Ptheta3c[,i],4,0)
response5[,i]<-ifelse(r>=Ptheta1a[,i]+Ptheta2b[,i] + Ptheta3c[,i] +
```

```
53: Ptheta4d[,i],5,0)
54:
55: #Compiling the response matrix to object 'score'
56: score<-response1+response2+response3+response4+response5
57: }
58: return(score)
59: }
60:
61: #Function to calculate skewness
62: skew <- function (x)
63: {
64: sk <- function(xx) {
65: n <- length(xx)
66: mn <- mean(xx)
67: dif.x <- xx - mn
68: m2 <- sum(dif.x^2)/n
69: m3 <- sum(dif.x^3)/n
70: m3/(m2^(3/2))
71: }
72: if (ncol(x) == 1 || is.null(dim(x)))
73: return(sk(x))
74: else return(apply(x, 2, sk))
75: }
76:
77: #Function to calculate kurtosis
78: kurtosis <-function (x)
79: {
80: kt <- function(xx) {
81: n <- length(xx)
82: mn <- mean(xx)
83: dif.x <- xx - mn
84: m2 <- sum(dif.x^2)/n
85: m4 <- sum(dif.x^4)/n
86: (m4/m2^2) - 3
87: }
88: if (ncol(x) == 1 || is.null(dim(x)))
89: return(kt(x))
90: else return(apply(x, 2, kt))
91: }
92:
93: #---------------------- fixed conditions -------------------------#
94:
95: result <- matrix(0, nrow = Numiter, ncol = 9)
96:
97: #---------------------- starting for-loop ------------------------#
98:
99: iter<-0
100: good.iter<-1
101: while(good.iter <= Numiter) {
102:
103: iter<-iter+1
104: set.seed(seednum+iter)
105:
106: #-------------------- initializing values ----------------------#
```

```
107:
108: #Create a person by item matrix for the scores of CBRF 1 through 4
109: Ptheta1 <- matrix(0, nrow = numSubj, ncol = numItem)
110: Ptheta2 <- matrix(0, nrow = numSubj, ncol = numItem)
111: Ptheta3 <- matrix(0, nrow = numSubj, ncol = numItem)
112: Ptheta4 <- matrix(0, nrow = numSubj, ncol = numItem)
113:
114: #Create a person x item matrix for the scores of CRF 1 through 5
115: Ptheta1a <- matrix(0, nrow = numSubj, ncol = numItem)
116: Ptheta2b <- matrix(0, nrow = numSubj, ncol = numItem)
117: Ptheta3c <- matrix(0, nrow = numSubj, ncol = numItem)
118: Ptheta4d <- matrix(0, nrow = numSubj, ncol = numItem)
119: Ptheta5e <- matrix(0, nrow = numSubj, ncol = numItem)
120:
121: #Create a person by item matrix for raw scores of cat 1 through 5
122: response1 <- matrix(0, nrow = numSubj, ncol = numItem)
123: response2 <- matrix(0, nrow = numSubj, ncol = numItem)
124: response3 <- matrix(0, nrow = numSubj, ncol = numItem)
125: response4 <- matrix(0, nrow = numSubj, ncol = numItem)
126: response5 <- matrix(0, nrow = numSubj, ncol = numItem)
127:
128: #Create the final person by item matrix of raw responses
129: score <- matrix(0, nrow = numSubj, ncol = numItem)
130:
131: #------------- Generating IV1, IV2, and DV ----------------------#
132:
133: theta1<-scale(rnorm(numSubj))
134: theta2<-scale(rnorm(numSubj))
135: theta3<-scale(w1*theta1+w2*theta2+sqrt(1-(w1^2+w2^2))*scale(rnorm(numSubj)))
136:
137: #------------ Specifying item parameters ----------------------#
138:
139: a1 <- runif(numItem, a.low, a.high)
140: a2 <- runif(numItem, a.low, a.high)
141: a3 <- runif(numItem, a.low, a.high)
142: b1a <- rnorm(numItem, b.mean, b.sd)
143: b2a <- rnorm(numItem, b.mean, b.sd)
144: b3a <- rnorm(numItem, b.mean, b.sd)
145:
146: #------------ Generating response patterns ----------------------#
147:
148: score1<-score.item.prg(numItem,numSubj,Ptheta1,a1,b1a,score, theta1)
149: score2<-score.item.prg(numItem,numSubj, Ptheta1, a2, b2a, score, theta2)
150: score3<-score.item.prg(numItem,numSubj, Ptheta1, a3, b3a, score, theta3)
151:
152: #--------- Compute Cronbach's Alpha for reliability ------------#
153:
154: alpha1<-cronbach.alpha(score1)
155: alpha2<-cronbach.alpha(score2)
156: alpha3<-cronbach.alpha(score3)
157: alpha.score1<-alpha1$alpha
158: alpha.score2<-alpha2$alpha
159: alpha.score3<-alpha3$alpha
160:
```

```
#---------- Estimating parameters using PARSCALE4.1 ------------#

#Command to invoke PARSCALE to generate theta estimates
#Note that all files must be located in the PARSCALE directory

#-------------------------n=750, k=15----------------------------#
#---------------------------score 1------------------------------#
score1psl<-data.frame(1001:1750,score1)
write.table(score1psl,"C:\\Program Files\\PARSCALE4\\testscore_15-750.dat",
sep="",row.names=FALSE,col.names=FALSE)
system("score15-750.bat",show.output.on.console = FALSE)
theta.ab1<-read.table("C:\\Program Files\\PARSCALE4\\testscore15-750.SCO",
head=F,fill=T)[(1:750)*2,7]
theta.ab1<-as.matrix(theta.ab1)
#---------------------------score 2------------------------------#
score2psl<-data.frame(1001:1750,score2)
write.table(score2psl,"C:\\Program Files\\PARSCALE4\\testscore_15-750.dat",
sep="",row.names=FALSE,col.names=FALSE)
system("score15-750.bat",show.output.on.console = FALSE)
theta.ab2<-read.table("C:\\Program Files\\PARSCALE4\\testscore15-750.SCO",
head=F,fill=T)[(1:750)*2,7]
theta.ab2<-as.matrix(theta.ab2)
#---------------------------score 3------------------------------#
score3psl<-data.frame(1001:1750,score3)
write.table(score3psl,"C:\\Program Files\\PARSCALE4\\testscore_15-750.dat",
sep="",row.names=FALSE,col.names=FALSE)
system("score15-750.bat",show.output.on.console = FALSE)
theta.ab3<-read.table("C:\\Program Files\\PARSCALE4\\testscore15-750.SCO",
head=F,fill=T)[(1:750)*2,7]
theta.ab3<-as.matrix(theta.ab3)
#----------------------------------------------------------------#

#-- Computing the rmsq, total scores, skew and kurtosis----------#

rmsq<- sqrt( mean( (theta1 - theta.ab1)^2 ) )

score1 <- apply(score1, 1, mean)
score2 <- apply(score2, 1, mean)
score3 <- apply(score3, 1, mean)

score1.skew<-skew(score1)
score1.kurtosis<-kurtosis(score1)
score2.skew<-skew(score2)
score2.kurtosis<-kurtosis(score2)
score3.skew<-skew(score3)
score3.kurtosis<-kurtosis(score3)

theta.ab1skew<-skew(theta.ab1)
theta.ab1kurtosis<-kurtosis(theta.ab1)
theta.ab2skew<-skew(theta.ab2)
theta.ab2kurtosis<-kurtosis(theta.ab2)
theta.ab3skew<-skew(theta.ab3)
theta.ab3kurtosis<-kurtosis(theta.ab3)

```

```
#-- Applying additive and multiplicative regression models ------#

#Actual theta scores
theta.add<-lm(theta3~theta1+theta2)
theta.mul<-lm(theta3~theta1+theta2+I(theta1*theta2))

#Raw scores
sum.add<-lm(score3~score1+score2)
sum.mul<-lm(score3~score1+score2+I(score1*score2))

#Estimated theta scores
theta.ab.add<-lm(theta.ab3~theta.ab1+theta.ab2)
theta.ab.mul<-lm(theta.ab3~theta.ab1+theta.ab2+I(theta.ab1*theta.ab2))

#-------- Shapiro-Wilk test for checking normality --------------#

theta.orderres <- summary(theta.add)$res
sum.orderres <- summary(sum.add)$res
thetahat.orderres <- summary(theta.ab.add)$res

swtheta.p <- round(shapiro.test(theta.orderres)$p,5)
swsum.p <- round(shapiro.test(sum.orderres)$p,5)
swthetahat.p <- round(shapiro.test(thetahat.orderres)$p,5)

#report r-square, r-square change sig., and rm squared deviations#

cat("Working on sample",seednum,"iteration",iter,good.iter, "\n")
theta.add.rsq <- summary(theta.add)$r.squared
theta.mul.rsq <- summary(theta.mul)$r.squared
theta.p <- anova(theta.add, theta.mul)$"Pr(>F)"[2]
theta.p[is.na(theta.p)] <- 1.00

sum.add.rsq <- summary(sum.add)$r.squared
sum.mul.rsq <- summary(sum.mul)$r.squared
sum.p <- round(anova(sum.add, sum.mul)$"Pr(>F)"[2], 4)
sum.p[is.na(sum.p)] <- 1.00

theta.ab.add.rsq <- summary(theta.ab.add)$r.squared
theta.ab.mul.rsq <- summary(theta.ab.mul)$r.squared
theta.ab.p<-round(anova(theta.ab.add,theta.ab.mul)$"Pr(>F)"[2], 4)
theta.ab.p[is.na(theta.ab.p)] <- 1.00

#----------- Summarize results of each loop --------------------#

iter.results<-as.vector(c(iter,
seednum,
numItem,
a.low,
a.high,
b.mean,
b.sd,
w1,
w2,
theta.add.rsq,
```

```
theta.mul.rsq,
theta.p,
sum.add.rsq,
sum.mul.rsq,
sum.p,
theta.ab.add.rsq,
theta.ab.mul.rsq,
theta.ab.p,
rmsq,
alpha.score1,
alpha.score2,
alpha.score3,
swtheta.p,
swsum.p,
swthetahat.p,
score1.skew,
score2.skew,
score3.skew,
score1.kurtosis,
score2.kurtosis,
score3.kurtosis,
theta.ab1skew,
theta.ab2skew,
theta.ab3skew,
theta.ab1kurtosis,
theta.ab2kurtosis,
theta.ab3kurtosis))

names(iter.results)<-NULL
sink(results.file,append=TRUE)
print(iter.results,digits=4,quote=FALSE)
sink()
good.iter<-good.iter+1
}

}
#-------------------- End loop structure ----------------------#

#=====================================================================#
# Begin looping individual conditions #
#=====================================================================#

options(width=2000)

{

results.file25<-"C:/Documents and Settings/Admin/Desktop/DissModel/C25.txt"
study1(seednum = 25,
numSubj = 750,
Numiter = n.it,
b.mean = -2.5,
b.sd = 0.70,
a.low = .31,
a.high = .58,
```

```
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file25)

results.file27<-"C:/Documents and Settings/Admin/Desktop/DissModel/C27.txt"
study1(seednum = 27,
numSubj = 750,
Numiter = n.it,
b.mean = -2.5,
b.sd = 0.70,
a.low = .31,
a.high = .58,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file27)

results.file29<-"C:/Documents and Settings/Admin/Desktop/DissModel/C29.txt"
study1(seednum = 29,
numSubj = 750,
Numiter = n.it,
b.mean = -2.5,
b.sd = 0.70,
a.low = .58,
a.high = 1.13,
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file29)

results.file31 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C31.txt"
study1(seednum = 31,
numSubj = 750,
Numiter = n.it,
b.mean = -2.5,
b.sd = 0.70,
a.low = .58,
a.high = 1.13,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file31)

results.file33 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C33.txt"
study1(seednum = 33,
numSubj = 750,
Numiter = n.it,
b.mean = -1.0,
b.sd = 0.70,
a.low = .31,
a.high = .58,
w1 = .3,
w2 = .3,
```

```
numItem = 15,
results.file = results.file33)

results.file35 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C35.txt"
study1(seednum = 35,
numSubj = 750,
Numiter = n.it,
b.mean = -1.0,
b.sd = 0.70,
a.low = .31,
a.high = .58,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file35)

results.file37<-"C:/Documents and Settings/Admin/Desktop/DissModel/C37.txt"
study1(seednum = 37,
numSubj = 750,
Numiter = n.it,
b.mean = -1.0,
b.sd = 0.70,
a.low = .58,
a.high = 1.13,
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file37)

results.file39 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C39.txt"
study1(seednum = 39,
numSubj = 750,
Numiter = n.it,
b.mean = -1.0,
b.sd = 0.70,
a.low = .58,
a.high = 1.13,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file39)

results.file41 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C41.txt"
study1(seednum = 41,
numSubj = 750,
Numiter = n.it,
b.mean = 0.5,
b.sd = 0.70,
a.low = .31,
a.high = .58,
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file41)
```

```
431:
432: results.file43 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C43.txt"
433: study1(seednum = 43,
434: numSubj = 750,
435: Numiter = n.it,
436: b.mean = 0.5,
437: b.sd = 0.70,
438: a.low = .31,
439: a.high = .58,
440: w1 = .5,
441: w2 = .5,
442: numItem = 15,
443: results.file = results.file43)
444:
445: results.file45 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C45.txt"
446: study1(seednum = 45,
447: numSubj = 750,
448: Numiter = n.it,
449: b.mean = 0.5,
450: b.sd = 0.70,
451: a.low = .58,
452: a.high = 1.13,
453: w1 = .3,
454: w2 = .3,
455: numItem = 15,
456: results.file = results.file45)
457:
458: results.file47 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C47.txt"
459: study1(seednum = 47,
460: numSubj = 750,
461: Numiter = n.it,
462: b.mean = 0.5,
463: b.sd = 0.70,
464: a.low = .58,
465: a.high = 1.13,
466: w1 = .5,
467: w2 = .5,
468: numItem = 15,
469: results.file = results.file47)
470:
471: }
472:
473: #=================================================================#
474: #Simulation loops for spurious interactions (n=750,k=30, normal)#
475: #=================================================================#
476:
477: {
478:
479: setwd("C:/Program Files/PARSCALE4")
480:
481: #Generate raw response matrix for IV1, IV2, and DV
482: score.item.prg<-function(numItem,numSubj,Ptheta,a,b,score,theta)
483: {
484: b1<-b
```

```
485: b2<-b1+.70
486: b3<-b2+.70
487: b4<-b3+.70
488:
489: for(i in 1:numItem){
490:
491: Ptheta1[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b1[i]))) #CBRF 1
492: Ptheta2[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b2[i]))) #CBRF 2
493: Ptheta3[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b3[i]))) #CBRF 3
494: Ptheta4[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b4[i]))) #CBRF 4
495:
496: Ptheta1a[, i]<-1.0 - Ptheta1[, i] #CRF for option 1
497: Ptheta2b[, i]<-Ptheta1[, i] - Ptheta2[, i] #CRF for option 2
498: Ptheta3c[, i]<-Ptheta2[, i] - Ptheta3[, i] #CRF for option 3
499: Ptheta4d[, i]<-Ptheta3[, i] - Ptheta4[, i] #CRF for option 4
500: Ptheta5e[, i]<-Ptheta4[, i] #CRF for option 5
501:
502: #Generating a response matrix by comparing a random value from a
503: #uniform distribution U(0,1) to the relative score categories
504: r<-runif(numSubj)
505: response1[,i]<-ifelse(r < Ptheta1a[,i],1,0)
506: response2[,i]<-ifelse(r < Ptheta1a[,i]+Ptheta2b[,i] & r >=
507: Ptheta1a[,i],2,0)
508: response3[,i]<-ifelse(r<Ptheta1a[,i]+Ptheta2b[,i]+Ptheta3c[,i]&r>=
509: Ptheta1a[,i] + Ptheta2b[,i],3,0)
510: response4[,i]<-ifelse(r<Ptheta1a[,i]+Ptheta2b[,i] + Ptheta3c[,i] +
511: Ptheta4d[,i]&r>=Ptheta1a[,i]+ Ptheta2b[,i] + Ptheta3c[,i],4,0)
512: response5[,i]<-ifelse(r>=Ptheta1a[,i]+Ptheta2b[,i]+Ptheta3c[,i] +
513: Ptheta4d[,i],5,0)
514:
515: #Compiling the response matrix to object 'score'
516: score<-response1+response2+response3+response4+response5
517: }
518: return(score)
519: }
520:
521: #Function to calculate skewness
522: skew <- function (x)
523: {
524: sk <- function(xx) {
525: n <- length(xx)
526: mn <- mean(xx)
527: dif.x <- xx - mn
528: m2 <- sum(dif.x^2)/n
529: m3 <- sum(dif.x^3)/n
530: m3/(m2^(3/2))
531: }
532: if (ncol(x) == 1 || is.null(dim(x)))
533: return(sk(x))
534: else return(apply(x, 2, sk))
535: }
536:
537: #Function to calculate kurtosis
538: kurtosis <-function (x)
```

```
{
kt <- function(xx) {
n <- length(xx)
mn <- mean(xx)
dif.x <- xx - mn
m2 <- sum(dif.x^2)/n
m4 <- sum(dif.x^4)/n
(m4/m2^2) - 3
}
if (ncol(x) == 1 || is.null(dim(x)))
return(kt(x))
else return(apply(x, 2, kt))
}

#------------------ fixed conditions -------------------------#

result <- matrix(0, nrow = Numiter, ncol = 9)

#--------------------- starting for-loop ------------------------#

iter<-0
good.iter<-1
while(good.iter <= Numiter) {

iter<-iter+1
set.seed(seednum+iter)

#------------------- initializing values ------------------------#

#Create a person by item matrix for the scores of CBRF 1 through 4
Ptheta1 <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta2 <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta3 <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta4 <- matrix(0, nrow = numSubj, ncol = numItem)

#Create a person x item matrix for the scores of CRF 1 through 5
Ptheta1a <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta2b <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta3c <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta4d <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta5e <- matrix(0, nrow = numSubj, ncol = numItem)

#Create a person x item matrix for raw scores of cat 1 through 5
response1 <- matrix(0, nrow = numSubj, ncol = numItem)
response2 <- matrix(0, nrow = numSubj, ncol = numItem)
response3 <- matrix(0, nrow = numSubj, ncol = numItem)
response4 <- matrix(0, nrow = numSubj, ncol = numItem)
response5 <- matrix(0, nrow = numSubj, ncol = numItem)

#Create the final person by item matrix of raw responses
score <- matrix(0, nrow = numSubj, ncol = numItem)

#------------- Generating IV1, IV2, and DV ---------------------#

```

```
theta1<-scale(rnorm(numSubj))
theta2<-scale(rnorm(numSubj))
theta3<-scale(w1*theta1+w2*theta2+sqrt(1-(w1^2+w2^2))*scale(rnorm(numSubj)))

#------------ Specifying item parameters ------------------------#

a1 <- runif(numItem, a.low, a.high)
a2 <- runif(numItem, a.low, a.high)
a3 <- runif(numItem, a.low, a.high)
b1a <- rnorm(numItem, b.mean, b.sd)
b2a <- rnorm(numItem, b.mean, b.sd)
b3a <- rnorm(numItem, b.mean, b.sd)

#------------ Generating reponse patterns -----------------------#

score1<-score.item.prg(numItem,numSubj,Ptheta1,a1,b1a,score,theta1)
score2<-score.item.prg(numItem,numSubj,Ptheta1,a2,b2a,score, theta2)
score3<-score.item.prg(numItem,numSubj,Ptheta1,a3,b3a,score, theta3)

#--------- Compute Cronbach's Alpha for reliability ------------#

alpha1<-cronbach.alpha(score1)
alpha2<-cronbach.alpha(score2)
alpha3<-cronbach.alpha(score3)
alpha.score1<-alpha1$alpha
alpha.score2<-alpha2$alpha
alpha.score3<-alpha3$alpha

#---------- Estimating parameters using PARSCALE4.1 -------------#

#Command to invoke PARSCALE to generate theta estimates
#Note that all files must be located in the PARSCALE directory

#-------------------------n=750, k=30----------------------------#
#----------------------------score 1-----------------------------#
score1psl<-data.frame(1001:1750,score1)
write.table(score1psl,"C:\\Program Files\\PARSCALE4\\testscore_30-750.dat",
sep="",row.names=FALSE,col.names=FALSE)
system("score30-750.bat",show.output.on.console = FALSE)
theta.ab1<-read.table("C:\\Program Files\\PARSCALE4\\testscore30-750.SCO",
head=F,fill=T)[(1:750)*2,7]
theta.ab1<-as.matrix(theta.ab1)
#----------------------------score 2-----------------------------#
score2psl<-data.frame(1001:1750,score2)
write.table(score2psl,"C:\\Program Files\\PARSCALE4\\testscore_30-750.dat",
sep="",row.names=FALSE,col.names=FALSE)
system("score30-750.bat",show.output.on.console = FALSE)
theta.ab2<-read.table("C:\\Program Files\\PARSCALE4\\testscore30-750.SCO",
head=F,fill=T)[(1:750)*2,7]
theta.ab2<-as.matrix(theta.ab2)
#----------------------------score 3-----------------------------#
score3psl<-data.frame(1001:1750,score3)
write.table(score3psl,"C:\\Program Files\\PARSCALE4\\testscore_30-750.dat",
sep="",row.names=FALSE,col.names=FALSE)
```

```
647: system("score30-750.bat",show.output.on.console = FALSE)
648: theta.ab3<-read.table("C:\\Program Files\\PARSCALE4\\testscore30-750.SCO",
649: head=F,fill=T)[(1:750)*2,7]
650: theta.ab3<-as.matrix(theta.ab3)
651: #-----------------------------------------------------------------#
652:
653: # Computing the rmsq, total scores, skew and kurtosis -----------#
654:
655: rmsq<- sqrt( mean( (theta1 - theta.ab1)^2 ) )
656:
657: score1 <- apply(score1, 1, mean)
658: score2 <- apply(score2, 1, mean)
659: score3 <- apply(score3, 1, mean)
660:
661: score1.skew<-skew(score1)
662: score1.kurtosis<-kurtosis(score1)
663: score2.skew<-skew(score2)
664: score2.kurtosis<-kurtosis(score2)
665: score3.skew<-skew(score3)
666: score3.kurtosis<-kurtosis(score3)
667:
668: theta.ab1skew<-skew(theta.ab1)
669: theta.ab1kurtosis<-kurtosis(theta.ab1)
670: theta.ab2skew<-skew(theta.ab2)
671: theta.ab2kurtosis<-kurtosis(theta.ab2)
672: theta.ab3skew<-skew(theta.ab3)
673: theta.ab3kurtosis<-kurtosis(theta.ab3)
674:
675: #-- Applying additive and multiplicative regression models ------#
676:
677: #Actual theta scores
678: theta.add<-lm(theta3~theta1+theta2)
679: theta.mul<-lm(theta3~theta1+theta2+I(theta1*theta2))
680:
681: #Raw scores
682: sum.add<-lm(score3~score1+score2)
683: sum.mul<-lm(score3~score1+score2+I(score1*score2))
684:
685: #Estimated theta scores
686: theta.ab.add<-lm(theta.ab3~theta.ab1+theta.ab2)
687: theta.ab.mul<-lm(theta.ab3~theta.ab1+theta.ab2+I(theta.ab1*theta.ab2))
688:
689: #-------- Shapiro-Wilk test for checking normality --------------#
690:
691: theta.orderres <- summary(theta.add)$res
692: sum.orderres <- summary(sum.add)$res
693: thetahat.orderres <- summary(theta.ab.add)$res
694:
695: swtheta.p <- round(shapiro.test(theta.orderres)$p,5)
696: swsum.p <- round(shapiro.test(sum.orderres)$p,5)
697: swthetahat.p <- round(shapiro.test(thetahat.orderres)$p,5)
698:
699: #report r-square, r-square change sig., & rm squared deviations #
700:
```

```
701: cat("Working on sample",seednum,"iteration",iter,good.iter, "\n")
702: theta.add.rsq <- summary(theta.add)$r.squared
703: theta.mul.rsq <- summary(theta.mul)$r.squared
704: theta.p <- anova(theta.add, theta.mul)$"Pr(>F)"[2]
705: theta.p[is.na(theta.p)] <- 1.00
706:
707: sum.add.rsq <- summary(sum.add)$r.squared
708: sum.mul.rsq <- summary(sum.mul)$r.squared
709: sum.p <- round(anova(sum.add, sum.mul)$"Pr(>F)"[2], 4)
710: sum.p[is.na(sum.p)] <- 1.00
711:
712: theta.ab.add.rsq <- summary(theta.ab.add)$r.squared
713: theta.ab.mul.rsq <- summary(theta.ab.mul)$r.squared
714: theta.ab.p<-round(anova(theta.ab.add,theta.ab.mul)$"Pr(>F)"[2], 4)
715: theta.ab.p[is.na(theta.ab.p)] <- 1.00
716:
717: #------------ Summarize results of each loop --------------------#
718:
719: iter.results<-as.vector(c(iter,
720: seednum,
721: numItem,
722: a.low,
723: a.high,
724: b.mean,
725: b.sd,
726: w1,
727: w2,
728: theta.add.rsq,
729: theta.mul.rsq,
730: theta.p,
731: sum.add.rsq,
732: sum.mul.rsq,
733: sum.p,
734: theta.ab.add.rsq,
735: theta.ab.mul.rsq,
736: theta.ab.p,
737: rmsq,
738: alpha.score1,
739: alpha.score2,
740: alpha.score3,
741: swtheta.p,
742: swsum.p,
743: swthetahat.p,
744: score1.skew,
745: score2.skew,
746: score3.skew,
747: score1.kurtosis,
748: score2.kurtosis,
749: score3.kurtosis,
750: theta.ab1skew,
751: theta.ab2skew,
752: theta.ab3skew,
753: theta.ab1kurtosis,
754: theta.ab2kurtosis,
```

```
755: theta.ab3kurtosis))
756:
757: names(iter.results)<-NULL
758: sink(results.file,append=TRUE)
759: print(iter.results,digits=4,quote=FALSE)
760: sink()
761: good.iter<-good.iter+1
762: }
763:
764: }
765: #-------------------- End loop structure ----------------------#
766:
767: #================================================================#
768: # Begin looping individual conditions #
769: #================================================================#
770:
771: options(width=2000)
772:
773: {
774:
775: results.file26<-"C:/Documents and Settings/Admin/Desktop/DissModel/C26.txt"
776: study1(seednum = 26,
777: numSubj = 750,
778: Numiter = n.it,
779: b.mean = -2.5,
780: b.sd = 0.7,
781: a.low = .31,
782: a.high = .58,
783: w1 = .3,
784: w2 = .3,
785: numItem = 30,
786: results.file = results.file26)
787:
788: results.file28<-"C:/Documents and Settings/Admin/Desktop/DissModel/C28.txt"
789: study1(seednum = 28,
790: numSubj = 750,
791: Numiter = n.it,
792: b.mean = -2.5,
793: b.sd = 0.7,
794: a.low = .31,
795: a.high = .58,
796: w1 = .5,
797: w2 = .5,
798: numItem = 30,
799: results.file = results.file28)
800:
801: results.file30 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C30.txt"
802: study1(seednum = 30,
803: numSubj = 750,
804: Numiter = n.it,
805: b.mean = -2.5,
806: b.sd = 0.7,
807: a.low = .58,
808: a.high = 1.13,
```

```
w1 = .3,
w2 = .3,
numItem = 30,
results.file = results.file30)

results.file32 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C32.txt"
study1(seednum = 32,
numSubj = 750,
Numiter = n.it,
b.mean = -2.5,
b.sd = 0.7,
a.low = .58,
a.high = 1.13,
w1 = .5,
w2 = .5,
numItem = 30,
results.file = results.file32)

results.file34<-"C:/Documents and Settings/Admin/Desktop/DissModel/C34.txt"
study1(seednum = 34,
numSubj = 750,
Numiter = n.it,
b.mean = -1.0,
b.sd = 0.7,
a.low = .31,
a.high = .58,
w1 = .3,
w2 = .3,
numItem = 30,
results.file = results.file34)

results.file36 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C36.txt"
study1(seednum = 36,
numSubj = 750,
Numiter = n.it,
b.mean = -1.0,
b.sd = 0.7,
a.low = .31,
a.high = .58,
w1 = .5,
w2 = .5,
numItem = 30,
results.file = results.file36)

results.file38<-"C:/Documents and Settings/Admin/Desktop/DissModel/C38.txt"
study1(seednum = 38,
numSubj = 750,
Numiter = n.it,
b.mean = -1.0,
b.sd = 0.7,
a.low = .58,
a.high = 1.13,
w1 = .3,
w2 = .3,
```

```
863: numItem = 30,
864: results.file = results.file38)
865:
866: results.file40 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C40.txt"
867: study1(seednum = 40,
868: numSubj = 750,
869: Numiter = n.it,
870: b.mean = -1.0,
871: b.sd = 0.7,
872: a.low = .58,
873: a.high = 1.13,
874: w1 = .5,
875: w2 = .5,
876: numItem = 30,
877: results.file = results.file40)
878:
879: results.file42 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C42.txt"
880: study1(seednum = 42,
881: numSubj = 750,
882: Numiter = n.it,
883: b.mean = 0.5,
884: b.sd = 0.7,
885: a.low = .31,
886: a.high = .58,
887: w1 = .3,
888: w2 = .3,
889: numItem = 30,
890: results.file = results.file42)
891:
892: results.file44<-"C:/Documents and Settings/Admin/Desktop/DissModel/C44.txt"
893: study1(seednum = 44,
894: numSubj = 750,
895: Numiter = n.it,
896: b.mean = 0.5,
897: b.sd = 0.7,
898: a.low = .31,
899: a.high = .58,
900: w1 = .5,
901: w2 = .5,
902: numItem = 30,
903: results.file = results.file44)
904:
905: results.file46 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C46.txt"
906: study1(seednum = 46,
907: numSubj = 750,
908: Numiter = n.it,
909: b.mean = 0.5,
910: b.sd = 0.7,
911: a.low = .58,
912: a.high = 1.13,
913: w1 = .3,
914: w2 = .3,
915: numItem = 30,
916: results.file = results.file46)
```

```
917:
918: results.file48 <-"C:/Documents and Settings/Admin/Desktop/DissModel/C48.txt"
919: study1(seednum = 48,
920: numSubj = 750,
921: Numiter = n.it,
922: b.mean = 0.5,
923: b.sd = 0.7,
924: a.low = .58,
925: a.high = 1.13,
926: w1 = .5,
927: w2 = .5,
928: numItem = 30,
929: results.file = results.file48)
930:
931: }
932:
933: #=================================================================#
934: # Summarize Results for Table 2 (n=250, normal) #
935: #=================================================================#
936:
937: n.it<-1000
938:
939: results.list<-paste("C:/Documents and Settings/Admin/Desktop/DissModel/C",
940: 25:48,sep="")
941: results.list<-paste(results.list,".txt",sep="")
942:
943: type1.theta<-rep(0,24)
944: type1.sum <- rep(0,24)
945: type1.thetahat <- rep(0,24)
946: rdiff.theta <- rep(0,24)
947: rdiff.sum <- rep(0,24)
948: rdiff.thetahat <- rep(0,24)
949: mn.rmsq<-rep(0,24)
950: pvalue.score1.mn<-rep(0,24)
951: pvalue.score2.mn<-rep(0,24)
952: pvalue.score3.mn<-rep(0,24)
953: pvalue.score1.sd<-rep(0,24)
954: pvalue.score2.sd<-rep(0,24)
955: pvalue.score3.sd<-rep(0,24)
956: alpha.score1<-rep(0,24)
957: alpha.score2<-rep(0,24)
958: alpha.score3<-rep(0,24)
959: sw.theta.p<-rep(0,24)
960: sw.sum.p<-rep(0,24)
961: sw.thetahat.p<-rep(0,24)
962:
963: skew.score1<-rep(0,24)
964: skew.score2<-rep(0,24)
965: skew.score3<-rep(0,24)
966:
967: kurtosis.score1<-rep(0,24)
968: kurtosis.score2<-rep(0,24)
969: kurtosis.score3<-rep(0,24)
970:
```

```
skew.theta.ab1<-rep(0,24)
skew.theta.ab2<-rep(0,24)
skew.theta.ab3<-rep(0,24)

kurtosis.theta.ab1<-rep(0,24)
kurtosis.theta.ab2<-rep(0,24)
kurtosis.theta.ab3<-rep(0,24)

for(i in 1:24) {
infile <- read.table(results.list[[i]], header=FALSE)
infile<- infile[,2:ncol(infile)]
names(infile)<-list( "iter",
"seednum",
"numItem",
"a.low",
"a.high",
"b.mean",
"b.sd",
"w1",
"w2",
"theta.add.rsq",
"theta.mul.rsq",
"theta.p",
"sum.add.rsq",
"sum.mul.rsq",
"sum.p",
"theta.ab.add.rsq",
"theta.ab.mul.rsq",
"theta.ab.p",
"rmsq",
"alpha.score1",
"alpha.score2",
"alpha.score3",
"swtheta.p",
"swsum.p",
"swthetahat.p",
"skew.score1",
"skew.score2",
"skew.score3",
"kurtosis.score1",
"kurtosis.score2",
"kurtosis.score3",
"skew.theta.ab1",
"skew.theta.ab2",
"skew.theta.ab3",
"kurtosis.theta.ab1",
"kurtosis.theta.ab2",
"kurtosis.theta.ab3")

write.table(infile,
"C:/Documents and Settings/Admin/Desktop/DissModel/norm 750 full.txt")

type1.theta[i]<-sum(infile["theta.p"] <= .05)/n.it
type1.sum[i]<-sum(infile["sum.p"] <=.05)/n.it
```

```
type1.thetahat[i]<-sum(infile["theta.ab.p"] <=.05)/n.it

rdiff.theta[i]<-round(sum(infile["theta.mul.rsq"]-infile["theta.add.rsq"])/
n.it,2)
rdiff.sum[i]<-round(sum(infile["sum.mul.rsq"]-infile["sum.add.rsq"])/n.it,2)
rdiff.thetahat[i]<-round(sum(infile["theta.ab.mul.rsq"] -
infile["theta.ab.add.rsq"])/n.it,2)
mn.rmsq[i]<-round(mean(infile["rmsq"]),2)
alpha.score1[i]<-round(mean(infile["alpha.score1"]),2)
alpha.score2[i]<-round(mean(infile["alpha.score2"]),2)
alpha.score3[i]<-round(mean(infile["alpha.score3"]),2)

sw.theta.p[i]<-round(sum(infile["swtheta.p"] > .05)/n.it,5)
sw.sum.p[i]<-round(sum(infile["swsum.p"] > .05)/n.it,5)
sw.thetahat.p[i]<-round(sum(infile["swthetahat.p"] > .05)/n.it,5)

skew.score1[i]<-round(mean(infile["skew.score1"]),5)
skew.score2[i]<-round(mean(infile["skew.score2"]),5)
skew.score3[i]<-round(mean(infile["skew.score3"]),5)
kurtosis.score1[i]<-round(mean(infile["kurtosis.score1"]),5)
kurtosis.score2[i]<-round(mean(infile["kurtosis.score2"]),5)
kurtosis.score3[i]<-round(mean(infile["kurtosis.score3"]),5)
skew.theta.ab1[i]<-round(mean(infile["skew.theta.ab1"]),5)
skew.theta.ab2[i]<-round(mean(infile["skew.theta.ab2"]),5)
skew.theta.ab3[i]<-round(mean(infile["skew.theta.ab3"],na.rm=TRUE),5)
kurtosis.theta.ab1[i]<-round(mean(infile["kurtosis.theta.ab1"]),5)
kurtosis.theta.ab2[i]<-round(mean(infile["kurtosis.theta.ab2"]),5)
kurtosis.theta.ab3[i]<-round(mean(infile["kurtosis.theta.ab3"],na.rm=TRUE),5)

}

n <- c(rep(750,24))
b <- c(rep("N(-1.5,1.0)",8),rep("N(0,1)",8),rep("N(1.5,1.0)",8))
a <- c(rep("U(0.31, 0.58)",4),rep("U(0.58, 1.13)",4))
a <- rep(a,3)
B1B2 <- rep(c(.3,.3,.5,.5),6)
Items<-rep(c(15,30),12)

mean.alpha<-round(apply(cbind(alpha.score1,alpha.score2,alpha.score3),
1,mean),2)

type1.theta<-round(type1.theta,2)
type1.sum<-round(type1.sum,2)
type1.thetahat<-round(type1.thetahat,2)

sw.theta.p<-round(sw.theta.p,2)
sw.sum.p<-round(sw.sum.p,2)
sw.thetahat<-round(sw.thetahat.p,2)

sktab1<-round(data.frame(skew.score3, kurtosis.score3, skew.theta.ab3,
kurtosis.theta.ab3),2)

table2<-data.frame(n,b,a,B1B2,Items,type1.theta,type1.sum,type1.thetahat,
mean.alpha,sw.theta.p,sw.sum.p,sw.thetahat,sktab1)
```

```
1079:
1080: print(table2)
1081: write.table(table2,
1082: "C:/Documents and Settings/Admin/Desktop/DissModel/Table2 norm 750.txt")
1083: #======End simulation for Table 2 (n=750, normal)===============#
```

APPENDIX C: R CODE FOR SIMULATION 3

```
#Morse Dissertation Table 3 (n=250, restricted)

#Load latent trait model library
library("ltm")

#Set number of iterations per condition
n.it<-1000

#Individual Monte Carlo loop structure
study1<-function(seednum, numSubj=numSubj, Numiter=n.it,
b.mean, b.sd, a.low, a.high, w1, w2, numItem, results.file)

#=================================================================#
#Simulation loops for spurious interactions (n=250, k=15, restr.)#
#=================================================================#

{

setwd("C:/Program Files/PARSCALE4")

#Generate raw response matrix for IV1, IV2, and DV
score.item.prg<-function(numItem,numSubj,Ptheta,a, b, score, theta)
{
b1<-b
b2<-b1+.35
b3<-b2+.35
b4<-b3+.35

for(i in 1:numItem){

Ptheta1[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b1[i]))) #CBRF 1
Ptheta2[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b2[i]))) #CBRF 2
Ptheta3[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b3[i]))) #CBRF 3
Ptheta4[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b4[i]))) #CBRF 4

Ptheta1a[, i]<-1.0 - Ptheta1[, i] #CRF for option 1
Ptheta2b[, i]<-Ptheta1[, i] - Ptheta2[, i] #CRF for option 2
Ptheta3c[, i]<-Ptheta2[, i] - Ptheta3[, i] #CRF for option 3
Ptheta4d[, i]<-Ptheta3[, i] - Ptheta4[, i] #CRF for option 4
Ptheta5e[, i]<-Ptheta4[, i] #CRF for option 5

#Generating a response matrix by comparing a random value from a
#uniform distribution U(0,1) to the relative score categories
r<-runif(numSubj)
response1[,i]<-ifelse(r < Ptheta1a[,i],1,0)
response2[,i]<-ifelse(r < Ptheta1a[,i] + Ptheta2b[,i] & r >=
Ptheta1a[,i],2,0)
response3[,i]<-ifelse(r<Ptheta1a[,i]+Ptheta2b[,i]+Ptheta3c[,i]&r >=
Ptheta1a[,i]+Ptheta2b[,i],3,0)
response4[,i]<-ifelse(r < Ptheta1a[,i]+Ptheta2b[,i]+ Ptheta3c[,i] +
Ptheta4d[,i] & r >= Ptheta1a[,i] + Ptheta2b[,i] + Ptheta3c[,i],4,0)
response5[,i]<-ifelse(r >= Ptheta1a[,i]+Ptheta2b[,i]+Ptheta3c[,i] +
```

```
53: Ptheta4d[,i],5,0)
54:
55: #Compiling the response matrix to object 'score'
56: score<-response1+response2+response3+response4+response5
57: }
58: return(score)
59: }
60:
61: #Function to calculate skewness
62: skew <- function (x)
63: {
64: sk <- function(xx) {
65: n <- length(xx)
66: mn <- mean(xx)
67: dif.x <- xx - mn
68: m2 <- sum(dif.x^2)/n
69: m3 <- sum(dif.x^3)/n
70: m3/(m2^(3/2))
71: }
72: if (ncol(x) == 1 || is.null(dim(x)))
73: return(sk(x))
74: else return(apply(x, 2, sk))
75: }
76:
77: #Function to calculate kurtosis
78: kurtosis <-function (x)
79: {
80: kt <- function(xx) {
81: n <- length(xx)
82: mn <- mean(xx)
83: dif.x <- xx - mn
84: m2 <- sum(dif.x^2)/n
85: m4 <- sum(dif.x^4)/n
86: (m4/m2^2) - 3
87: }
88: if (ncol(x) == 1 || is.null(dim(x)))
89: return(kt(x))
90: else return(apply(x, 2, kt))
91: }
92:
93: #---------------------- fixed conditions -------------------------#
94:
95: result <- matrix(0, nrow = Numiter, ncol = 9)
96:
97: #---------------------- starting for-loop ------------------------#
98:
99: iter<-0
100: good.iter<-1
101: while(good.iter <= Numiter) {
102:
103: iter<-iter+1
104: set.seed(seednum+iter)
105:
106: #-------------------- initializing values ----------------------#
```

```
107:
108: #Create a person by item matrix for the scores of CBRF 1 through 4
109: Ptheta1 <- matrix(0, nrow = numSubj, ncol = numItem)
110: Ptheta2 <- matrix(0, nrow = numSubj, ncol = numItem)
111: Ptheta3 <- matrix(0, nrow = numSubj, ncol = numItem)
112: Ptheta4 <- matrix(0, nrow = numSubj, ncol = numItem)
113:
114: #Create a person x item matrix for the scores of CRF 1 through 5
115: Ptheta1a <- matrix(0, nrow = numSubj, ncol = numItem)
116: Ptheta2b <- matrix(0, nrow = numSubj, ncol = numItem)
117: Ptheta3c <- matrix(0, nrow = numSubj, ncol = numItem)
118: Ptheta4d <- matrix(0, nrow = numSubj, ncol = numItem)
119: Ptheta5e <- matrix(0, nrow = numSubj, ncol = numItem)
120:
121: #Create a person x item matrix for raw scores of cat 1 through 5
122: response1 <- matrix(0, nrow = numSubj, ncol = numItem)
123: response2 <- matrix(0, nrow = numSubj, ncol = numItem)
124: response3 <- matrix(0, nrow = numSubj, ncol = numItem)
125: response4 <- matrix(0, nrow = numSubj, ncol = numItem)
126: response5 <- matrix(0, nrow = numSubj, ncol = numItem)
127:
128: #Create the final person by item matrix of raw responses
129: score <- matrix(0, nrow = numSubj, ncol = numItem)
130:
131: #------------- Generating IV1, IV2, and DV ----------------------#
132:
133: theta1<-scale(rnorm(numSubj))
134: theta2<-scale(rnorm(numSubj))
135: theta3<-scale(w1*theta1+w2*theta2+sqrt(1-(w1^2+w2^2))*scale(rnorm(numSubj)))
136:
137: #------------ Specifying item parameters ----------------------#
138:
139: a1 <- runif(numItem, a.low, a.high)
140: a2 <- runif(numItem, a.low, a.high)
141: a3 <- runif(numItem, a.low, a.high)
142: b1a <- rnorm(numItem, b.mean, b.sd)
143: b2a <- rnorm(numItem, b.mean, b.sd)
144: b3a <- rnorm(numItem, b.mean, b.sd)
145:
146: #------------ Generating reponse patterns ----------------------#
147:
148: score1<-score.item.prg(numItem,numSubj,Ptheta1,a1,b1a,score, theta1)
149: score2<-score.item.prg(numItem,numSubj, Ptheta1, a2, b2a, score, theta2)
150: score3<-score.item.prg(numItem,numSubj, Ptheta1, a3, b3a, score, theta3)
151:
152: #-------- Compute Cronbach's Alpha for reliability -------------#
153:
154: alpha1<-cronbach.alpha(score1)
155: alpha2<-cronbach.alpha(score2)
156: alpha3<-cronbach.alpha(score3)
157: alpha.score1<-alpha1$alpha
158: alpha.score2<-alpha2$alpha
159: alpha.score3<-alpha3$alpha
160:
```

```
#---------- Estimating parameters using PARSCALE4.1 ------------#

#Command to invoke PARSCALE to generate theta estimates
#Note that all files must be located in the PARSCALE directory

#-------------------------n=250, k=15----------------------------#
#---------------------------score 1------------------------------#
score1psl<-data.frame(1001:1250,score1)
write.table(score1psl,"C:\\Program Files\\PARSCALE4\\testscore_15-250.dat",
sep="",row.names=FALSE,col.names=FALSE)
system("score15-250.bat",show.output.on.console = FALSE)
theta.ab1<-read.table("C:\\Program Files\\PARSCALE4\\testscore15-250.SCO",
head=F,fill=T)[(1:250)*2,7]
theta.ab1<-as.matrix(theta.ab1)
#---------------------------score 2------------------------------#
score2psl<-data.frame(1001:1250,score2)
write.table(score2psl,"C:\\Program Files\\PARSCALE4\\testscore_15-250.dat",
sep="",row.names=FALSE,col.names=FALSE)
system("score15-250.bat",show.output.on.console = FALSE)
theta.ab2<-read.table("C:\\Program Files\\PARSCALE4\\testscore15-250.SCO",
head=F,fill=T)[(1:250)*2,7]
theta.ab2<-as.matrix(theta.ab2)
#---------------------------score 3------------------------------#
score3psl<-data.frame(1001:1250,score3)
write.table(score3psl,"C:\\Program Files\\PARSCALE4\\testscore_15-250.dat",
sep="",row.names=FALSE,col.names=FALSE)
system("score15-250.bat",show.output.on.console = FALSE)
theta.ab3<-read.table("C:\\Program Files\\PARSCALE4\\testscore15-250.SCO",
head=F,fill=T)[(1:250)*2,7]
theta.ab3<-as.matrix(theta.ab3)
#----------------------------------------------------------------#

#-- Computing the rmsq, total scores, skew and kurtosis----------#

rmsq<- sqrt( mean( (theta1 - theta.ab1)^2 ) )

score1 <- apply(score1, 1, mean)
score2 <- apply(score2, 1, mean)
score3 <- apply(score3, 1, mean)

score1.skew<-skew(score1)
score1.kurtosis<-kurtosis(score1)
score2.skew<-skew(score2)
score2.kurtosis<-kurtosis(score2)
score3.skew<-skew(score3)
score3.kurtosis<-kurtosis(score3)

theta.ab1skew<-skew(theta.ab1)
theta.ab1kurtosis<-kurtosis(theta.ab1)
theta.ab2skew<-skew(theta.ab2)
theta.ab2kurtosis<-kurtosis(theta.ab2)
theta.ab3skew<-skew(theta.ab3)
theta.ab3kurtosis<-kurtosis(theta.ab3)

```

```
#-- Applying additive and multiplicative regression models ------#

#Actual theta scores
theta.add<-lm(theta3~theta1+theta2)
theta.mul<-lm(theta3~theta1+theta2+I(theta1*theta2))

#Raw scores
sum.add<-lm(score3~score1+score2)
sum.mul<-lm(score3~score1+score2+I(score1*score2))

#Estimated theta scores
theta.ab.add<-lm(theta.ab3~theta.ab1+theta.ab2)
theta.ab.mul<-lm(theta.ab3~theta.ab1+theta.ab2+I(theta.ab1*theta.ab2))

#-------- Shapiro-Wilk test for checking normality --------------#

theta.orderres <- summary(theta.add)$res
sum.orderres <- summary(sum.add)$res
thetahat.orderres <- summary(theta.ab.add)$res

swtheta.p <- round(shapiro.test(theta.orderres)$p,5)
swsum.p <- round(shapiro.test(sum.orderres)$p,5)
swthetahat.p <- round(shapiro.test(thetahat.orderres)$p,5)

#report r-square, r-square change sig., and rm squared deviations#

cat("Working on sample",seednum,"iteration",iter, good.iter, "\n")
theta.add.rsq <- summary(theta.add)$r.squared
theta.mul.rsq <- summary(theta.mul)$r.squared
theta.p <- anova(theta.add, theta.mul)$"Pr(>F)"[2]
theta.p[is.na(theta.p)] <- 1.00

sum.add.rsq <- summary(sum.add)$r.squared
sum.mul.rsq <- summary(sum.mul)$r.squared
sum.p <- round(anova(sum.add, sum.mul)$"Pr(>F)"[2], 4)
sum.p[is.na(sum.p)] <- 1.00

theta.ab.add.rsq <- summary(theta.ab.add)$r.squared
theta.ab.mul.rsq <- summary(theta.ab.mul)$r.squared
theta.ab.p<-round(anova(theta.ab.add,theta.ab.mul)$"Pr(>F)"[2], 4)
theta.ab.p[is.na(theta.ab.p)] <- 1.00

#------------ Summarize results of each loop -------------------#

iter.results<-as.vector(c(iter,
seednum,
numItem,
a.low,
a.high,
b.mean,
b.sd,
w1,
w2,
theta.add.rsq,
```

```
theta.mul.rsq,
theta.p,
sum.add.rsq,
sum.mul.rsq,
sum.p,
theta.ab.add.rsq,
theta.ab.mul.rsq,
theta.ab.p,
rmsq,
alpha.score1,
alpha.score2,
alpha.score3,
swtheta.p,
swsum.p,
swthetahat.p,
score1.skew,
score2.skew,
score3.skew,
score1.kurtosis,
score2.kurtosis,
score3.kurtosis,
theta.ab1skew,
theta.ab2skew,
theta.ab3skew,
theta.ab1kurtosis,
theta.ab2kurtosis,
theta.ab3kurtosis))

names(iter.results)<-NULL
sink(results.file,append=TRUE)
print(iter.results,digits=4,quote=FALSE)
sink()
good.iter<-good.iter+1
}

}
#-------------------- End loop structure ----------------------#

#=====================================================================#
# Begin looping individual conditions #
#=====================================================================#

options(width=2000)

{

results.file1<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C1.txt"
study1(seednum = 1,
numSubj = 250,
Numiter = n.it,
b.mean = -2.0,
b.sd = 0.35,
a.low = .31,
a.high = .58,
```

```
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file1)

results.file3<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C3.txt"
study1(seednum = 3,
numSubj = 250,
Numiter = n.it,
b.mean = -2.0,
b.sd = 0.35,
a.low = .31,
a.high = .58,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file3)

results.file5<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C5.txt"
study1(seednum = 5,
numSubj = 250,
Numiter = n.it,
b.mean = -2.0,
b.sd = 0.35,
a.low = .58,
a.high = 1.13,
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file5)

results.file7 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C7.txt"
study1(seednum = 7,
numSubj = 250,
Numiter = n.it,
b.mean = -2.0,
b.sd = 0.35,
a.low = .58,
a.high = 1.13,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file7)

results.file9 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C9.txt"
study1(seednum = 9,
numSubj = 250,
Numiter = n.it,
b.mean = -0.5,
b.sd = 0.35,
a.low = .31,
a.high = .58,
w1 = .3,
w2 = .3,
```

```
numItem = 15,
results.file = results.file9)

results.file11 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C11.txt"
study1(seednum = 11,
numSubj = 250,
Numiter = n.it,
b.mean = -0.5,
b.sd = 0.35,
a.low = .31,
a.high = .58,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file11)

results.file13<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C13.txt"
study1(seednum = 13,
numSubj = 250,
Numiter = n.it,
b.mean = -0.5,
b.sd = 0.35,
a.low = .58,
a.high = 1.13,
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file13)

results.file15 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C15.txt"
study1(seednum = 15,
numSubj = 250,
Numiter = n.it,
b.mean = -0.5,
b.sd = 0.35,
a.low = .58,
a.high = 1.13,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file15)

results.file17 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C17.txt"
study1(seednum = 17,
numSubj = 250,
Numiter = n.it,
b.mean = 1.0,
b.sd = 0.35,
a.low = .31,
a.high = .58,
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file17)
```

```
431:
432: results.file19 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C19.txt"
433: study1(seednum = 19,
434: numSubj = 250,
435: Numiter = n.it,
436: b.mean = 1.0,
437: b.sd = 0.35,
438: a.low = .31,
439: a.high = .58,
440: w1 = .5,
441: w2 = .5,
442: numItem = 15,
443: results.file = results.file19)
444:
445: results.file21 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C21.txt"
446: study1(seednum = 21,
447: numSubj = 250,
448: Numiter = n.it,
449: b.mean = 1.0,
450: b.sd = 0.35,
451: a.low = .58,
452: a.high = 1.13,
453: w1 = .3,
454: w2 = .3,
455: numItem = 15,
456: results.file = results.file21)
457:
458: results.file23 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C23.txt"
459: study1(seednum = 23,
460: numSubj = 250,
461: Numiter = n.it,
462: b.mean = 1.0,
463: b.sd = 0.35,
464: a.low = .58,
465: a.high = 1.13,
466: w1 = .5,
467: w2 = .5,
468: numItem = 15,
469: results.file = results.file23)
470:
471: }
472:
473: #=================================================================#
474: #Simulation loops for spurious interactions (n=250,k=30,restr.)#
475: #=================================================================#
476:
477: {
478:
479: setwd("C:/Program Files/PARSCALE4")
480:
481: #Generate raw response matrix for IV1, IV2, and DV
482: score.item.prg<-function(numItem,numSubj,Ptheta,a,b,score, theta)
483: {
484: b1<-b
```

```
485: b2<-b1+.35
486: b3<-b2+.35
487: b4<-b3+.35
488:
489: for(i in 1:numItem){
490:
491: Ptheta1[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b1[i]))) #CBRF 1
492: Ptheta2[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b2[i]))) #CBRF 2
493: Ptheta3[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b3[i]))) #CBRF 3
494: Ptheta4[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b4[i]))) #CBRF 4
495:
496: Ptheta1a[, i]<-1.0 - Ptheta1[, i] #CRF for option 1
497: Ptheta2b[, i]<-Ptheta1[, i] - Ptheta2[, i] #CRF for option 2
498: Ptheta3c[, i]<-Ptheta2[, i] - Ptheta3[, i] #CRF for option 3
499: Ptheta4d[, i]<-Ptheta3[, i] - Ptheta4[, i] #CRF for option 4
500: Ptheta5e[, i]<-Ptheta4[, i] #CRF for option 5
501:
502: #Generating a response matrix by comparing a random value from a
503: #uniform distribution U(0,1) to the relative score categories
504: r<-runif(numSubj)
505: response1[,i]<-ifelse(r < Ptheta1a[,i],1,0)
506: response2[,i]<-ifelse(r < Ptheta1a[,i]+Ptheta2b[,i] & r >=
507: Ptheta1a[,i],2,0)
508: response3[,i]<-ifelse(r<Ptheta1a[,i]+Ptheta2b[,i]+Ptheta3c[,i]&r>=
509: Ptheta1a[,i] + Ptheta2b[,i],3,0)
510: response4[,i]<-ifelse(r<Ptheta1a[,i]+Ptheta2b[,i] + Ptheta3c[,i] +
511: Ptheta4d[,i]&r >= Ptheta1a[,i] + Ptheta2b[,i] + Ptheta3c[,i],4,0)
512: response5[,i]<-ifelse(r>=Ptheta1a[,i]+Ptheta2b[,i]+Ptheta3c[,i] +
513: Ptheta4d[,i],5,0)
514:
515: #Compiling the response matrix to object 'score'
516: score<-response1+response2+response3+response4+response5
517: }
518: return(score)
519: }
520:
521: #Function to calculate skewness
522: skew <- function (x)
523: {
524: sk <- function(xx) {
525: n <- length(xx)
526: mn <- mean(xx)
527: dif.x <- xx - mn
528: m2 <- sum(dif.x^2)/n
529: m3 <- sum(dif.x^3)/n
530: m3/(m2^(3/2))
531: }
532: if (ncol(x) == 1 || is.null(dim(x)))
533: return(sk(x))
534: else return(apply(x, 2, sk))
535: }
536:
537: #Function to calculate kurtosis
538: kurtosis <-function (x)
```

```
{
kt <- function(xx) {
n <- length(xx)
mn <- mean(xx)
dif.x <- xx - mn
m2 <- sum(dif.x^2)/n
m4 <- sum(dif.x^4)/n
(m4/m2^2) - 3
}
if (ncol(x) == 1 || is.null(dim(x)))
return(kt(x))
else return(apply(x, 2, kt))
}

#---------------------- fixed conditions -------------------------#

result <- matrix(0, nrow = Numiter, ncol = 9)

#---------------------- starting for-loop ------------------------#

iter<-0
good.iter<-1
while(good.iter <= Numiter) {

iter<-iter+1
set.seed(seednum+iter)

#------------------- initializing values ------------------------#

#Create a person by item matrix for the scores of CBRF 1 through 4
Ptheta1 <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta2 <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta3 <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta4 <- matrix(0, nrow = numSubj, ncol = numItem)

#Create a person x item matrix for the scores of CRF 1 through 5
Ptheta1a <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta2b <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta3c <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta4d <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta5e <- matrix(0, nrow = numSubj, ncol = numItem)

#Create a person x item matrix for raw scores of cat 1 through 5
response1 <- matrix(0, nrow = numSubj, ncol = numItem)
response2 <- matrix(0, nrow = numSubj, ncol = numItem)
response3 <- matrix(0, nrow = numSubj, ncol = numItem)
response4 <- matrix(0, nrow = numSubj, ncol = numItem)
response5 <- matrix(0, nrow = numSubj, ncol = numItem)

#Create the final person by item matrix of raw responses
score <- matrix(0, nrow = numSubj, ncol = numItem)

#------------- Generating IV1, IV2, and DV ---------------------#

```

```
593: theta1<-scale(rnorm(numSubj))
594: theta2<-scale(rnorm(numSubj))
595: theta3<-scale(w1*theta1+w2*theta2+sqrt(1-(w1^2+w2^2))*scale(rnorm(numSubj)))
596:
597: #------------ Specifying item parameters ------------------------#
598:
599: a1 <- runif(numItem, a.low, a.high)
600: a2 <- runif(numItem, a.low, a.high)
601: a3 <- runif(numItem, a.low, a.high)
602: b1a <- rnorm(numItem, b.mean, b.sd)
603: b2a <- rnorm(numItem, b.mean, b.sd)
604: b3a <- rnorm(numItem, b.mean, b.sd)
605:
606: #--------- Generating reponse patterns -------------------------#
607:
608: score1<-score.item.prg(numItem,numSubj,Ptheta1,a1,b1a,score, theta1)
609: score2<-score.item.prg(numItem,numSubj,Ptheta1,a2,b2a,score, theta2)
610: score3<-score.item.prg(numItem,numSubj,Ptheta1,a3,b3a,score, theta3)
611:
612: #------ Compute Cronbach's Alpha for reliability ---------------#
613:
614: alpha1<-cronbach.alpha(score1)
615: alpha2<-cronbach.alpha(score2)
616: alpha3<-cronbach.alpha(score3)
617: alpha.score1<-alpha1$alpha
618: alpha.score2<-alpha2$alpha
619: alpha.score3<-alpha3$alpha
620:
621: #------------- Estimating parameters using PARSCALE4.1 -----------
-----#
622:
623: #Command to invoke PARSCALE to generate theta estimates
624: #Note that all files must be located in the PARSCALE directory
625:
626: #----------------------n=250, k=30-------------------------------#
627: #-------------------------score 1---------------------------------#
628: score1psl<-data.frame(1001:1250,score1)
629: write.table(score1psl,"C:\\Program Files\\PARSCALE4\\testscore_30-250.dat",
630: sep="",row.names=FALSE,col.names=FALSE)
631: system("score30-250.bat",show.output.on.console = FALSE)
632: theta.ab1<-read.table("C:\\Program Files\\PARSCALE4\\testscore30-250.SCO",
633: head=F,fill=T)[(1:250)*2,7]
634: theta.ab1<-as.matrix(theta.ab1)
635: #-------------------------score 2---------------------------------#
636: score2psl<-data.frame(1001:1250,score2)
637: write.table(score2psl,"C:\\Program Files\\PARSCALE4\\testscore_30-250.dat",
638: sep="",row.names=FALSE,col.names=FALSE)
639: system("score30-250.bat",show.output.on.console = FALSE)
640: theta.ab2<-read.table("C:\\Program Files\\PARSCALE4\\testscore30-250.SCO",
641: head=F,fill=T)[(1:250)*2,7]
642: theta.ab2<-as.matrix(theta.ab2)
643: #-------------------------score 3---------------------------------#
644: score3psl<-data.frame(1001:1250,score3)
645: write.table(score3psl,"C:\\Program Files\\PARSCALE4\\testscore_30-250.dat",
```

```
646: sep="",row.names=FALSE,col.names=FALSE)
647: system("score30-250.bat",show.output.on.console = FALSE)
648: theta.ab3<-read.table("C:\\Program Files\\PARSCALE4\\testscore30-250.SCO",
649: head=F,fill=T)[(1:250)*2,7]
650: theta.ab3<-as.matrix(theta.ab3)
651: #-----------------------------------------------------------------#
652:
653: #- Computing the rmsq, total scores, skew and kurtosis ----------#
654:
655: rmsq<- sqrt( mean( (theta1 - theta.ab1)^2 ) )
656:
657: score1 <- apply(score1, 1, mean)
658: score2 <- apply(score2, 1, mean)
659: score3 <- apply(score3, 1, mean)
660:
661: score1.skew<-skew(score1)
662: score1.kurtosis<-kurtosis(score1)
663: score2.skew<-skew(score2)
664: score2.kurtosis<-kurtosis(score2)
665: score3.skew<-skew(score3)
666: score3.kurtosis<-kurtosis(score3)
667:
668: theta.ab1skew<-skew(theta.ab1)
669: theta.ab1kurtosis<-kurtosis(theta.ab1)
670: theta.ab2skew<-skew(theta.ab2)
671: theta.ab2kurtosis<-kurtosis(theta.ab2)
672: theta.ab3skew<-skew(theta.ab3)
673: theta.ab3kurtosis<-kurtosis(theta.ab3)
674:
675: #-- Applying additive and multiplicative regression models ------#
676:
677: #Actual theta scores
678: theta.add<-lm(theta3~theta1+theta2)
679: theta.mul<-lm(theta3~theta1+theta2+I(theta1*theta2))
680:
681: #Raw scores
682: sum.add<-lm(score3~score1+score2)
683: sum.mul<-lm(score3~score1+score2+I(score1*score2))
684:
685: #Estimated theta scores
686: theta.ab.add<-lm(theta.ab3~theta.ab1+theta.ab2)
687: theta.ab.mul<-lm(theta.ab3~theta.ab1+theta.ab2+I(theta.ab1*theta.ab2))
688:
689: #-------- Shapiro-Wilk test for checking normality --------------#
690:
691: theta.orderres <- summary(theta.add)$res
692: sum.orderres <- summary(sum.add)$res
693: thetahat.orderres <- summary(theta.ab.add)$res
694:
695: swtheta.p <- round(shapiro.test(theta.orderres)$p,5)
696: swsum.p <- round(shapiro.test(sum.orderres)$p,5)
697: swthetahat.p <- round(shapiro.test(thetahat.orderres)$p,5)
698:
699: #report r-square, r-square change sig., & rm squared deviations #
```

```
700:
701: cat("Working on sample", seednum,"iteration",iter,good.iter, "\n")
702: theta.add.rsq <- summary(theta.add)$r.squared
703: theta.mul.rsq <- summary(theta.mul)$r.squared
704: theta.p <- anova(theta.add, theta.mul)$"Pr(>F)"[2]
705: theta.p[is.na(theta.p)] <- 1.00
706:
707: sum.add.rsq <- summary(sum.add)$r.squared
708: sum.mul.rsq <- summary(sum.mul)$r.squared
709: sum.p <- round(anova(sum.add, sum.mul)$"Pr(>F)"[2], 4)
710: sum.p[is.na(sum.p)] <- 1.00
711:
712: theta.ab.add.rsq <- summary(theta.ab.add)$r.squared
713: theta.ab.mul.rsq <- summary(theta.ab.mul)$r.squared
714: theta.ab.p<-round(anova(theta.ab.add,theta.ab.mul)$"Pr(>F)"[2], 4)
715: theta.ab.p[is.na(theta.ab.p)] <- 1.00
716:
717: #------------ Summarize results of each loop --------------------#
718:
719: iter.results<-as.vector(c(iter,
720: seednum,
721: numItem,
722: a.low,
723: a.high,
724: b.mean,
725: b.sd,
726: w1,
727: w2,
728: theta.add.rsq,
729: theta.mul.rsq,
730: theta.p,
731: sum.add.rsq,
732: sum.mul.rsq,
733: sum.p,
734: theta.ab.add.rsq,
735: theta.ab.mul.rsq,
736: theta.ab.p,
737: rmsq,
738: alpha.score1,
739: alpha.score2,
740: alpha.score3,
741: swtheta.p,
742: swsum.p,
743: swthetahat.p,
744: score1.skew,
745: score2.skew,
746: score3.skew,
747: score1.kurtosis,
748: score2.kurtosis,
749: score3.kurtosis,
750: theta.ab1skew,
751: theta.ab2skew,
752: theta.ab3skew,
753: theta.ab1kurtosis,
```

```
754: theta.ab2kurtosis,
755: theta.ab3kurtosis))
756:
757: names(iter.results)<-NULL
758: sink(results.file,append=TRUE)
759: print(iter.results,digits=4,quote=FALSE)
760: sink()
761: good.iter<-good.iter+1
762: }
763:
764: }
765: #-------------------- End loop structure ----------------------#
766:
767: #=====================================================================#
768: # Begin looping individual conditions #
769: #=====================================================================#
770:
771: options(width=2000)
772:
773: {
774:
775: results.file2<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C2.txt"
776: study1(seednum = 2,
777: numSubj = 250,
778: Numiter = n.it,
779: b.mean = -2.0,
780: b.sd = 0.35,
781: a.low = .31,
782: a.high = .58,
783: w1 = .3,
784: w2 = .3,
785: numItem = 30,
786: results.file = results.file2)
787:
788: results.file4<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C4.txt"
789: study1(seednum = 4,
790: numSubj = 250,
791: Numiter = n.it,
792: b.mean = -2.0,
793: b.sd = 0.35,
794: a.low = .31,
795: a.high = .58,
796: w1 = .5,
797: w2 = .5,
798: numItem = 30,
799: results.file = results.file4)
800:
801: results.file6 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C6.txt"
802: study1(seednum = 6,
803: numSubj = 250,
804: Numiter = n.it,
805: b.mean = -2.0,
806: b.sd = 0.35,
807: a.low = .58,
```

```
a.high = 1.13,
w1 = .3,
w2 = .3,
numItem = 30,
results.file = results.file6)

results.file8 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C8.txt"
study1(seednum = 8,
numSubj = 250,
Numiter = n.it,
b.mean = -2.0,
b.sd = 0.35,
a.low = .58,
a.high = 1.13,
w1 = .5,
w2 = .5,
numItem = 30,
results.file = results.file8)

results.file10<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C10.txt"
study1(seednum = 10,
numSubj = 250,
Numiter = n.it,
b.mean = -0.5,
b.sd = 0.35,
a.low = .31,
a.high = .58,
w1 = .3,
w2 = .3,
numItem = 30,
results.file = results.file10)

results.file12 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C12.txt"
study1(seednum = 12,
numSubj = 250,
Numiter = n.it,
b.mean = -0.5,
b.sd = 0.35,
a.low = .31,
a.high = .58,
w1 = .5,
w2 = .5,
numItem = 30,
results.file = results.file12)

results.file14<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C14.txt"
study1(seednum = 14,
numSubj = 250,
Numiter = n.it,
b.mean = -0.5,
b.sd = 0.35,
a.low = .58,
a.high = 1.13,
w1 = .3,
```

```
862: w2 = .3,
863: numItem = 30,
864: results.file = results.file14)
865:
866: results.file16 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C16.txt"
867: study1(seednum = 16,
868: numSubj = 250,
869: Numiter = n.it,
870: b.mean = -0.5,
871: b.sd = 0.35,
872: a.low = .58,
873: a.high = 1.13,
874: w1 = .5,
875: w2 = .5,
876: numItem = 30,
877: results.file = results.file16)
878:
879: results.file18 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C18.txt"
880: study1(seednum = 18,
881: numSubj = 250,
882: Numiter = n.it,
883: b.mean = 1.0,
884: b.sd = 0.35,
885: a.low = .31,
886: a.high = .58,
887: w1 = .3,
888: w2 = .3,
889: numItem = 30,
890: results.file = results.file18)
891:
892: results.file20<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C20.txt"
893: study1(seednum = 20,
894: numSubj = 250,
895: Numiter = n.it,
896: b.mean = 1.0,
897: b.sd = 0.35,
898: a.low = .31,
899: a.high = .58,
900: w1 = .5,
901: w2 = .5,
902: numItem = 30,
903: results.file = results.file20)
904:
905: results.file22 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C22.txt"
906: study1(seednum = 22,
907: numSubj = 250,
908: Numiter = n.it,
909: b.mean = 1.0,
910: b.sd = 0.35,
911: a.low = .58,
912: a.high = 1.13,
913: w1 = .3,
914: w2 = .3,
915: numItem = 30,
```

```
916: results.file = results.file22)
917:
918: results.file24 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C24.txt"
919: study1(seednum = 24,
920: numSubj = 250,
921: Numiter = n.it,
922: b.mean = 1.0,
923: b.sd = 0.35,
924: a.low = .58,
925: a.high = 1.13,
926: w1 = .5,
927: w2 = .5,
928: numItem = 30,
929: results.file = results.file24)
930:
931: }
932:
933: #=================================================================#
934: # Summarize Results for Table 3 (n=250, restricted) #
935: #=================================================================#
936:
937: n.it<-1000
938:
939: results.list<-paste("C:/Documents and Settings/Admin/Desktop/DissModel2/C",
940: 1:24,sep="")
941: results.list<-paste(results.list,".txt",sep="")
942:
943: type1.theta<-rep(0,24)
944: type1.sum <- rep(0,24)
945: type1.thetahat <- rep(0,24)
946: rdiff.theta <- rep(0,24)
947: rdiff.sum <- rep(0,24)
948: rdiff.thetahat <- rep(0,24)
949: mn.rmsq<-rep(0,24)
950: pvalue.score1.mn<-rep(0,24)
951: pvalue.score2.mn<-rep(0,24)
952: pvalue.score3.mn<-rep(0,24)
953: pvalue.score1.sd<-rep(0,24)
954: pvalue.score2.sd<-rep(0,24)
955: pvalue.score3.sd<-rep(0,24)
956: alpha.score1<-rep(0,24)
957: alpha.score2<-rep(0,24)
958: alpha.score3<-rep(0,24)
959: sw.theta.p<-rep(0,24)
960: sw.sum.p<-rep(0,24)
961: sw.thetahat.p<-rep(0,24)
962:
963: skew.score1<-rep(0,24)
964: skew.score2<-rep(0,24)
965: skew.score3<-rep(0,24)
966:
967: kurtosis.score1<-rep(0,24)
968: kurtosis.score2<-rep(0,24)
969: kurtosis.score3<-rep(0,24)
```

```
970:
971: skew.theta.ab1<-rep(0,24)
972: skew.theta.ab2<-rep(0,24)
973: skew.theta.ab3<-rep(0,24)
974:
975: kurtosis.theta.ab1<-rep(0,24)
976: kurtosis.theta.ab2<-rep(0,24)
977: kurtosis.theta.ab3<-rep(0,24)
978:
979: for(i in 1:24) {
980: infile <- read.table(results.list[[i]], header=FALSE)
981: infile<- infile[,2:ncol(infile)]
982: names(infile)<-list( "iter",
983: "seednum",
984: "numItem",
985: "a.low",
986: "a.high",
987: "b.mean",
988: "b.sd",
989: "w1",
990: "w2",
991: "theta.add.rsq",
992: "theta.mul.rsq",
993: "theta.p",
994: "sum.add.rsq",
995: "sum.mul.rsq",
996: "sum.p",
997: "theta.ab.add.rsq",
998: "theta.ab.mul.rsq",
999: "theta.ab.p",
1000: "rmsq",
1001: "alpha.score1",
1002: "alpha.score2",
1003: "alpha.score3",
1004: "swtheta.p",
1005: "swsum.p",
1006: "swthetahat.p",
1007: "skew.score1",
1008: "skew.score2",
1009: "skew.score3",
1010: "kurtosis.score1",
1011: "kurtosis.score2",
1012: "kurtosis.score3",
1013: "skew.theta.ab1",
1014: "skew.theta.ab2",
1015: "skew.theta.ab3",
1016: "kurtosis.theta.ab1",
1017: "kurtosis.theta.ab2",
1018: "kurtosis.theta.ab3")
1019:
1020: write.table(infile,
1021: "C:/Documents and Settings/Admin/Desktop/DissModel2/restr 250 full.txt")
1022:
1023: type1.theta[i]<-sum(infile["theta.p"] <= .05)/n.it
```

```
type1.sum[i]<-sum(infile["sum.p"] <=.05)/n.it
type1.thetahat[i]<-sum(infile["theta.ab.p"] <=.05)/n.it

rdiff.theta[i]<-round(sum(infile["theta.mul.rsq"]-infile["theta.add.rsq"])/
n.it,2)
rdiff.sum[i]<-round(sum(infile["sum.mul.rsq"]-infile["sum.add.rsq"])/n.it,2)
rdiff.thetahat[i]<-round(sum(infile["theta.ab.mul.rsq"] -
infile["theta.ab.add.rsq"])/n.it,2)
mn.rmsq[i]<-round(mean(infile["rmsq"]),2)
alpha.score1[i]<-round(mean(infile["alpha.score1"]),2)
alpha.score2[i]<-round(mean(infile["alpha.score2"]),2)
alpha.score3[i]<-round(mean(infile["alpha.score3"]),2)

sw.theta.p[i]<-round(sum(infile["swtheta.p"] > .05)/n.it,5)
sw.sum.p[i]<-round(sum(infile["swsum.p"] > .05)/n.it,5)
sw.thetahat.p[i]<-round(sum(infile["swthetahat.p"] > .05)/n.it,5)

skew.score1[i]<-round(mean(infile["skew.score1"]),5)
skew.score2[i]<-round(mean(infile["skew.score2"]),5)
skew.score3[i]<-round(mean(infile["skew.score3"]),5)
kurtosis.score1[i]<-round(mean(infile["kurtosis.score1"]),5)
kurtosis.score2[i]<-round(mean(infile["kurtosis.score2"]),5)
kurtosis.score3[i]<-round(mean(infile["kurtosis.score3"]),5)
skew.theta.ab1[i]<-round(mean(infile["skew.theta.ab1"]),5)
skew.theta.ab2[i]<-round(mean(infile["skew.theta.ab2"]),5)
skew.theta.ab3[i]<-round(mean(infile["skew.theta.ab3"],na.rm=TRUE),5)
kurtosis.theta.ab1[i]<-round(mean(infile["kurtosis.theta.ab1"]),5)
kurtosis.theta.ab2[i]<-round(mean(infile["kurtosis.theta.ab2"]),5)
kurtosis.theta.ab3[i]<-round(mean(infile["kurtosis.theta.ab3"],na.rm=TRUE),5)

}

n <- c(rep(250,24))
b <- c(rep("N(-1.5,1.0)",8),rep("N(0,1)",8),rep("N(1.5,1.0)",8))
a <- c(rep("U(0.31, 0.58)",4),rep("U(0.58, 1.13)",4))
a <- rep(a,3)
B1B2 <- rep(c(.3,.3,.5,.5),6)
Items<-rep(c(15,30),12)

mean.alpha<-round(apply(cbind(alpha.score1,alpha.score2,alpha.score3),
1,mean),2)

type1.theta<-round(type1.theta,2)
type1.sum<-round(type1.sum,2)
type1.thetahat<-round(type1.thetahat,2)

sw.theta.p<-round(sw.theta.p,2)
sw.sum.p<-round(sw.sum.p,2)
sw.thetahat<-round(sw.thetahat.p,2)

sktab1<-round(data.frame(skew.score3, kurtosis.score3, skew.theta.ab3,
kurtosis.theta.ab3),2)

table3<-data.frame(n,b,a,B1B2,Items,type1.theta,type1.sum,type1.thetahat,
```

```
1078: mean.alpha,sw.theta.p,sw.sum.p,sw.thetahat,sktab1)
1079:
1080: print(table3)
1081: write.table(table3,
1082: "C:/Documents and Settings/Admin/Desktop/DissModel2/Table3 restr 250.txt")
1083: #======End simulation for Table 3 (n=250, restricted)==========#
```

APPENDIX D: R CODE FOR SIMULATION 4

```
#Morse Dissertation Table 4 (n=750, restricted)

#Load latent trait model library
library("ltm")

#Set number of iterations per condition
n.it<-1000

#Individual Monte Carlo loop structure
study1<-function(seednum, numSubj=numSubj, Numiter=n.it,
b.mean, b.sd, a.low, a.high, w1, w2, numItem, results.file)

#=================================================================#
#Simulation loops for spurious interactions (n=750, k=15, restr.)#
#=================================================================#

{

setwd("C:/Program Files/PARSCALE4")

#Generate raw response matrix for IV1, IV2, and DV
score.item.prg<-function(numItem,numSubj,Ptheta,a,b,score, theta)
{
b1<-b
b2<-b1+.35
b3<-b2+.35
b4<-b3+.35

for(i in 1:numItem){

Ptheta1[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b1[i]))) #CBRF 1
Ptheta2[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b2[i]))) #CBRF 2
Ptheta3[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b3[i]))) #CBRF 3
Ptheta4[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b4[i]))) #CBRF 4

Ptheta1a[, i]<-1.0 - Ptheta1[, i] #CRF for option 1
Ptheta2b[, i]<-Ptheta1[, i] - Ptheta2[, i] #CRF for option 2
Ptheta3c[, i]<-Ptheta2[, i] - Ptheta3[, i] #CRF for option 3
Ptheta4d[, i]<-Ptheta3[, i] - Ptheta4[, i] #CRF for option 4
Ptheta5e[, i]<-Ptheta4[, i] #CRF for option 5

#Generating a response matrix by comparing a random value from a
#uniform distribution U(0,1) to the relative score categories
r<-runif(numSubj)
response1[,i]<-ifelse(r < Ptheta1a[,i],1,0)
response2[,i]<-ifelse(r < Ptheta1a[,i] + Ptheta2b[,i] & r >=
Ptheta1a[,i],2,0)
response3[,i]<-ifelse(r<Ptheta1a[,i]+Ptheta2b[,i]+Ptheta3c[,i]&r >=
Ptheta1a[,i]+Ptheta2b[,i],3,0)
response4[,i]<-ifelse(r<Ptheta1a[,i]+ Ptheta2b[,i] + Ptheta3c[,i] +
Ptheta4d[,i] & r >= Ptheta1a[,i] + Ptheta2b[,i] + Ptheta3c[,i],4,0)
response5[,i]<-ifelse(r >=Ptheta1a[,i]+Ptheta2b[,i]+ Ptheta3c[,i] +
```

```
53: Ptheta4d[,i],5,0)
54:
55: #Compiling the response matrix to object 'score'
56: score<-response1+response2+response3+response4+response5
57: }
58: return(score)
59: }
60:
61: #Function to calculate skewness
62: skew <- function (x)
63: {
64: sk <- function(xx) {
65: n <- length(xx)
66: mn <- mean(xx)
67: dif.x <- xx - mn
68: m2 <- sum(dif.x^2)/n
69: m3 <- sum(dif.x^3)/n
70: m3/(m2^(3/2))
71: }
72: if (ncol(x) == 1 || is.null(dim(x)))
73: return(sk(x))
74: else return(apply(x, 2, sk))
75: }
76:
77: #Function to calculate kurtosis
78: kurtosis <-function (x)
79: {
80: kt <- function(xx) {
81: n <- length(xx)
82: mn <- mean(xx)
83: dif.x <- xx - mn
84: m2 <- sum(dif.x^2)/n
85: m4 <- sum(dif.x^4)/n
86: (m4/m2^2) - 3
87: }
88: if (ncol(x) == 1 || is.null(dim(x)))
89: return(kt(x))
90: else return(apply(x, 2, kt))
91: }
92:
93: #---------------------- fixed conditions -------------------------#
94:
95: result <- matrix(0, nrow = Numiter, ncol = 9)
96:
97: #---------------------- starting for-loop ------------------------#
98:
99: iter<-0
100: good.iter<-1
101: while(good.iter <= Numiter) {
102:
103: iter<-iter+1
104: set.seed(seednum+iter)
105:
106: #-------------------- initializing values ----------------------#
```

```
107:
108: #Create a person by item matrix for the scores of CBRF 1 through 4
109: Ptheta1 <- matrix(0, nrow = numSubj, ncol = numItem)
110: Ptheta2 <- matrix(0, nrow = numSubj, ncol = numItem)
111: Ptheta3 <- matrix(0, nrow = numSubj, ncol = numItem)
112: Ptheta4 <- matrix(0, nrow = numSubj, ncol = numItem)
113:
114: #Create a person x item matrix for the scores of CRF 1 through 5
115: Ptheta1a <- matrix(0, nrow = numSubj, ncol = numItem)
116: Ptheta2b <- matrix(0, nrow = numSubj, ncol = numItem)
117: Ptheta3c <- matrix(0, nrow = numSubj, ncol = numItem)
118: Ptheta4d <- matrix(0, nrow = numSubj, ncol = numItem)
119: Ptheta5e <- matrix(0, nrow = numSubj, ncol = numItem)
120:
121: #Create a person x item matrix for raw scores of cat 1 through 5
122: response1 <- matrix(0, nrow = numSubj, ncol = numItem)
123: response2 <- matrix(0, nrow = numSubj, ncol = numItem)
124: response3 <- matrix(0, nrow = numSubj, ncol = numItem)
125: response4 <- matrix(0, nrow = numSubj, ncol = numItem)
126: response5 <- matrix(0, nrow = numSubj, ncol = numItem)
127:
128: #Create the final person by item matrix of raw responses
129: score <- matrix(0, nrow = numSubj, ncol = numItem)
130:
131: #------------- Generating IV1, IV2, and DV ---------------------#
132:
133: theta1<-scale(rnorm(numSubj))
134: theta2<-scale(rnorm(numSubj))
135: theta3<-scale(w1*theta1+w2*theta2+sqrt(1-(w1^2+w2^2))*scale(rnorm(numSubj)))
136:
137: #--------------- Specifying item parameters ---------------------#
138:
139: a1 <- runif(numItem, a.low, a.high)
140: a2 <- runif(numItem, a.low, a.high)
141: a3 <- runif(numItem, a.low, a.high)
142: b1a <- rnorm(numItem, b.mean, b.sd)
143: b2a <- rnorm(numItem, b.mean, b.sd)
144: b3a <- rnorm(numItem, b.mean, b.sd)
145:
146: #--------- Generating reponse patterns --------------------------#
147:
148: score1<-score.item.prg(numItem,numSubj, Ptheta1, a1, b1a, score, theta1)
149: score2<-score.item.prg(numItem,numSubj, Ptheta1, a2, b2a, score, theta2)
150: score3<-score.item.prg(numItem,numSubj, Ptheta1, a3, b3a, score, theta3)
151:
152: #------ Compute Cronbach's Alpha for reliability ---------------#
153:
154: alpha1<-cronbach.alpha(score1)
155: alpha2<-cronbach.alpha(score2)
156: alpha3<-cronbach.alpha(score3)
157: alpha.score1<-alpha1$alpha
158: alpha.score2<-alpha2$alpha
159: alpha.score3<-alpha3$alpha
160:
```

```
#---------- Estimating parameters using PARSCALE4.1 ------------#

#Command to invoke PARSCALE to generate theta estimates
#Note that all files must be located in the PARSCALE directory

#----------------------n=750, k=15-------------------------------#
#-----------------------score 1---------------------------------#
score1psl<-data.frame(1001:1750,score1)
write.table(score1psl,"C:\\Program Files\\PARSCALE4\\testscore_15-750.dat",
sep="",row.names=FALSE,col.names=FALSE)
system("score15-750.bat",show.output.on.console = FALSE)
theta.ab1<-read.table("C:\\Program Files\\PARSCALE4\\testscore15-750.SCO",
head=F,fill=T)[(1:750)*2,7]
theta.ab1<-as.matrix(theta.ab1)
#-----------------------score 2---------------------------------#
score2psl<-data.frame(1001:1750,score2)
write.table(score2psl,"C:\\Program Files\\PARSCALE4\\testscore_15-750.dat",
sep="",row.names=FALSE,col.names=FALSE)
system("score15-750.bat",show.output.on.console = FALSE)
theta.ab2<-read.table("C:\\Program Files\\PARSCALE4\\testscore15-750.SCO",
head=F,fill=T)[(1:750)*2,7]
theta.ab2<-as.matrix(theta.ab2)
#-----------------------score 3---------------------------------#
score3psl<-data.frame(1001:1750,score3)
write.table(score3psl,"C:\\Program Files\\PARSCALE4\\testscore_15-750.dat",
sep="",row.names=FALSE,col.names=FALSE)
system("score15-750.bat",show.output.on.console = FALSE)
theta.ab3<-read.table("C:\\Program Files\\PARSCALE4\\testscore15-750.SCO",
head=F,fill=T)[(1:750)*2,7]
theta.ab3<-as.matrix(theta.ab3)
#-----------------------------------------------------------------#

# Computing the rmsq, total scores, skew and kurtosis------------#

rmsq<- sqrt( mean( (theta1 - theta.ab1)^2 ) )

score1 <- apply(score1, 1, mean)
score2 <- apply(score2, 1, mean)
score3 <- apply(score3, 1, mean)

score1.skew<-skew(score1)
score1.kurtosis<-kurtosis(score1)
score2.skew<-skew(score2)
score2.kurtosis<-kurtosis(score2)
score3.skew<-skew(score3)
score3.kurtosis<-kurtosis(score3)

theta.ab1skew<-skew(theta.ab1)
theta.ab1kurtosis<-kurtosis(theta.ab1)
theta.ab2skew<-skew(theta.ab2)
theta.ab2kurtosis<-kurtosis(theta.ab2)
theta.ab3skew<-skew(theta.ab3)
theta.ab3kurtosis<-kurtosis(theta.ab3)

```

```
#-- Applying additive and multiplicative regression models ------#

#Actual theta scores
theta.add<-lm(theta3~theta1+theta2)
theta.mul<-lm(theta3~theta1+theta2+I(theta1*theta2))

#Raw scores
sum.add<-lm(score3~score1+score2)
sum.mul<-lm(score3~score1+score2+I(score1*score2))

#Estimated theta scores
theta.ab.add<-lm(theta.ab3~theta.ab1+theta.ab2)
theta.ab.mul<-lm(theta.ab3~theta.ab1+theta.ab2+I(theta.ab1*theta.ab2))

#-------- Shapiro-Wilk test for checking normality --------------#

theta.orderres <- summary(theta.add)$res
sum.orderres <- summary(sum.add)$res
thetahat.orderres <- summary(theta.ab.add)$res

swtheta.p <- round(shapiro.test(theta.orderres)$p,5)
swsum.p <- round(shapiro.test(sum.orderres)$p,5)
swthetahat.p <- round(shapiro.test(thetahat.orderres)$p,5)

#report r-square, r-square change sig., and rm squared deviations#

cat("Working on sample",seednum,"iteration",iter, good.iter, "\n")
theta.add.rsq <- summary(theta.add)$r.squared
theta.mul.rsq <- summary(theta.mul)$r.squared
theta.p <- anova(theta.add, theta.mul)$"Pr(>F)"[2]
theta.p[is.na(theta.p)] <- 1.00

sum.add.rsq <- summary(sum.add)$r.squared
sum.mul.rsq <- summary(sum.mul)$r.squared
sum.p <- round(anova(sum.add, sum.mul)$"Pr(>F)"[2], 4)
sum.p[is.na(sum.p)] <- 1.00

theta.ab.add.rsq <- summary(theta.ab.add)$r.squared
theta.ab.mul.rsq <- summary(theta.ab.mul)$r.squared
theta.ab.p<-round(anova(theta.ab.add,theta.ab.mul)$"Pr(>F)"[2], 4)
theta.ab.p[is.na(theta.ab.p)] <- 1.00

#------------ Summarize results of each loop --------------------#

iter.results<-as.vector(c(iter,
seednum,
numItem,
a.low,
a.high,
b.mean,
b.sd,
w1,
w2,
theta.add.rsq,
```

```
theta.mul.rsq,
theta.p,
sum.add.rsq,
sum.mul.rsq,
sum.p,
theta.ab.add.rsq,
theta.ab.mul.rsq,
theta.ab.p,
rmsq,
alpha.score1,
alpha.score2,
alpha.score3,
swtheta.p,
swsum.p,
swthetahat.p,
score1.skew,
score2.skew,
score3.skew,
score1.kurtosis,
score2.kurtosis,
score3.kurtosis,
theta.ab1skew,
theta.ab2skew,
theta.ab3skew,
theta.ab1kurtosis,
theta.ab2kurtosis,
theta.ab3kurtosis))

names(iter.results)<-NULL
sink(results.file,append=TRUE)
print(iter.results,digits=4,quote=FALSE)
sink()
good.iter<-good.iter+1
}

}
#-------------------- End loop structure ----------------------#

#=====================================================================#
# Begin looping individual conditions #
#=====================================================================#

options(width=2000)

{

results.file25<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C25.txt"
study1(seednum = 25,
numSubj = 750,
Numiter = n.it,
b.mean = -2.0,
b.sd = 0.35,
a.low = .31,
a.high = .58,
```

```
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file25)

results.file27<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C27.txt"
study1(seednum = 27,
numSubj = 750,
Numiter = n.it,
b.mean = -2.0,
b.sd = 0.35,
a.low = .31,
a.high = .58,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file27)

results.file29<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C29.txt"
study1(seednum = 29,
numSubj = 750,
Numiter = n.it,
b.mean = -2.0,
b.sd = 0.35,
a.low = .58,
a.high = 1.13,
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file29)

results.file31 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C31.txt"
study1(seednum = 31,
numSubj = 750,
Numiter = n.it,
b.mean = -2.0,
b.sd = 0.35,
a.low = .58,
a.high = 1.13,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file31)

results.file33 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C33.txt"
study1(seednum = 33,
numSubj = 750,
Numiter = n.it,
b.mean = -0.5,
b.sd = 0.35,
a.low = .31,
a.high = .58,
w1 = .3,
w2 = .3,
```

```
numItem = 15,
results.file = results.file33)

results.file35 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C35.txt"
study1(seednum = 35,
numSubj = 750,
Numiter = n.it,
b.mean = -0.5,
b.sd = 0.35,
a.low = .31,
a.high = .58,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file35)

results.file37<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C37.txt"
study1(seednum = 37,
numSubj = 750,
Numiter = n.it,
b.mean = -0.5,
b.sd = 0.35,
a.low = .58,
a.high = 1.13,
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file37)

results.file39 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C39.txt"
study1(seednum = 39,
numSubj = 750,
Numiter = n.it,
b.mean = -0.5,
b.sd = 0.35,
a.low = .58,
a.high = 1.13,
w1 = .5,
w2 = .5,
numItem = 15,
results.file = results.file39)

results.file41 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C41.txt"
study1(seednum = 41,
numSubj = 750,
Numiter = n.it,
b.mean = 1.0,
b.sd = 0.35,
a.low = .31,
a.high = .58,
w1 = .3,
w2 = .3,
numItem = 15,
results.file = results.file41)
```

```
431:
432: results.file43 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C43.txt"
433: study1(seednum = 43,
434: numSubj = 750,
435: Numiter = n.it,
436: b.mean = 1.0,
437: b.sd = 0.35,
438: a.low = .31,
439: a.high = .58,
440: w1 = .5,
441: w2 = .5,
442: numItem = 15,
443: results.file = results.file43)
444:
445: results.file45 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C45.txt"
446: study1(seednum = 45,
447: numSubj = 750,
448: Numiter = n.it,
449: b.mean = 1.0,
450: b.sd = 0.35,
451: a.low = .58,
452: a.high = 1.13,
453: w1 = .3,
454: w2 = .3,
455: numItem = 15,
456: results.file = results.file45)
457:
458: results.file47 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C47.txt"
459: study1(seednum = 47,
460: numSubj = 750,
461: Numiter = n.it,
462: b.mean = 1.0,
463: b.sd = 0.35,
464: a.low = .58,
465: a.high = 1.13,
466: w1 = .5,
467: w2 = .5,
468: numItem = 15,
469: results.file = results.file47)
470:
471: }
472:
473: #=================================================================#
474: #Simulation loops for spurious interactions (n=750,k=30, restr.)#
475: #=================================================================#
476:
477: {
478:
479: setwd("C:/Program Files/PARSCALE4")
480:
481: #Generate raw response matrix for IV1, IV2, and DV
482: score.item.prg<-function(numItem,numSubj,Ptheta,a,b,score,theta)
483: {
484: b1<-b
```

```
485: b2<-b1+.35
486: b3<-b2+.35
487: b4<-b3+.35
488:
489: for(i in 1:numItem){
490:
491: Ptheta1[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b1[i]))) #CBRF 1
492: Ptheta2[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b2[i]))) #CBRF 2
493: Ptheta3[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b3[i]))) #CBRF 3
494: Ptheta4[, i]<-1/(1 + exp(-1.702 * a[i] * (theta - b4[i]))) #CBRF 4
495:
496: Ptheta1a[, i]<-1.0 - Ptheta1[, i] #CRF for option 1
497: Ptheta2b[, i]<-Ptheta1[, i] - Ptheta2[, i] #CRF for option 2
498: Ptheta3c[, i]<-Ptheta2[, i] - Ptheta3[, i] #CRF for option 3
499: Ptheta4d[, i]<-Ptheta3[, i] - Ptheta4[, i] #CRF for option 4
500: Ptheta5e[, i]<-Ptheta4[, i] #CRF for option 5
501:
502: #Generating a response matrix by comparing a random value from a
503: #uniform distribution U(0,1) to the relative score categories
504: r<-runif(numSubj)
505: response1[,i]<-ifelse(r < Ptheta1a[,i],1,0)
506: response2[,i]<-ifelse(r < Ptheta1a[,i]+Ptheta2b[,i] & r >=
507: Ptheta1a[,i],2,0)
508: response3[,i]<-ifelse(r<Ptheta1a[,i]+Ptheta2b[,i]+Ptheta3c[,i]&r>=
509: Ptheta1a[,i] + Ptheta2b[,i],3,0)
510: response4[,i]<-ifelse(r<Ptheta1a[,i]+Ptheta2b[,i] + Ptheta3c[,i] +
511: Ptheta4d[,i]&r>=Ptheta1a[,i] + Ptheta2b[,i] + Ptheta3c[,i],4,0)
512: response5[,i]<-ifelse(r>=Ptheta1a[,i]+Ptheta2b[,i]+Ptheta3c[,i] +
513: Ptheta4d[,i],5,0)
514:
515: #Compiling the response matrix to object 'score'
516: score<-response1+response2+response3+response4+response5
517: }
518: return(score)
519: }
520:
521: #Function to calculate skewness
522: skew <- function (x)
523: {
524: sk <- function(xx) {
525: n <- length(xx)
526: mn <- mean(xx)
527: dif.x <- xx - mn
528: m2 <- sum(dif.x^2)/n
529: m3 <- sum(dif.x^3)/n
530: m3/(m2^(3/2))
531: }
532: if (ncol(x) == 1 || is.null(dim(x)))
533: return(sk(x))
534: else return(apply(x, 2, sk))
535: }
536:
537: #Function to calculate kurtosis
538: kurtosis <-function (x)
```

```
{
kt <- function(xx) {
n <- length(xx)
mn <- mean(xx)
dif.x <- xx - mn
m2 <- sum(dif.x^2)/n
m4 <- sum(dif.x^4)/n
(m4/m2^2) - 3
}
if (ncol(x) == 1 || is.null(dim(x)))
return(kt(x))
else return(apply(x, 2, kt))
}

#---------------------- fixed conditions -------------------------#

result <- matrix(0, nrow = Numiter, ncol = 9)

#---------------------- starting for-loop ------------------------#

iter<-0
good.iter<-1
while(good.iter <= Numiter) {

iter<-iter+1
set.seed(seednum+iter)

#------------------- initializing values ------------------------#

#Create a person by item matrix for the scores of CBRF 1 through 4
Ptheta1 <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta2 <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta3 <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta4 <- matrix(0, nrow = numSubj, ncol = numItem)

#Create a person x item matrix for the scores of CRF 1 through 5
Ptheta1a <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta2b <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta3c <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta4d <- matrix(0, nrow = numSubj, ncol = numItem)
Ptheta5e <- matrix(0, nrow = numSubj, ncol = numItem)

#Create a person x item matrix for raw scores of cat 1 through 5
response1 <- matrix(0, nrow = numSubj, ncol = numItem)
response2 <- matrix(0, nrow = numSubj, ncol = numItem)
response3 <- matrix(0, nrow = numSubj, ncol = numItem)
response4 <- matrix(0, nrow = numSubj, ncol = numItem)
response5 <- matrix(0, nrow = numSubj, ncol = numItem)

#Create the final person by item matrix of raw responses
score <- matrix(0, nrow = numSubj, ncol = numItem)

#------------- Generating IV1, IV2, and DV ---------------------#

```

```
theta1<-scale(rnorm(numSubj))
theta2<-scale(rnorm(numSubj))
theta3<-scale(w1*theta1+w2*theta2+sqrt(1-(w1^2+w2^2))*scale(rnorm(numSubj)))

#------------ Specifying item parameters ------------------------#

a1 <- runif(numItem, a.low, a.high)
a2 <- runif(numItem, a.low, a.high)
a3 <- runif(numItem, a.low, a.high)
b1a <- rnorm(numItem, b.mean, b.sd)
b2a <- rnorm(numItem, b.mean, b.sd)
b3a <- rnorm(numItem, b.mean, b.sd)

#------------ Generating reponse patterns -----------------------#

score1<-score.item.prg(numItem,numSubj,Ptheta1,a1,b1a, score, theta1)
score2<-score.item.prg(numItem,numSubj,Ptheta1,a2,b2a, score, theta2)
score3<-score.item.prg(numItem,numSubj,Ptheta1,a3,b3a, score, theta3)

#--------- Compute Cronbach's Alpha for reliability ------------#

alpha1<-cronbach.alpha(score1)
alpha2<-cronbach.alpha(score2)
alpha3<-cronbach.alpha(score3)
alpha.score1<-alpha1$alpha
alpha.score2<-alpha2$alpha
alpha.score3<-alpha3$alpha

#---------- Estimating parameters using PARSCALE4.1 -------------#

#Command to invoke PARSCALE to generate theta estimates
#Note that all files must be located in the PARSCALE directory

#-------------------------n=750, k=30----------------------------#
#----------------------------score 1-----------------------------#
score1psl<-data.frame(1001:1750,score1)
write.table(score1psl,"C:\\Program Files\\PARSCALE4\\testscore_30-750.dat",
sep="",row.names=FALSE,col.names=FALSE)
system("score30-750.bat",show.output.on.console = FALSE)
theta.ab1<-read.table("C:\\Program Files\\PARSCALE4\\testscore30-750.SCO",
head=F,fill=T)[(1:750)*2,7]
theta.ab1<-as.matrix(theta.ab1)
#----------------------------score 2-----------------------------#
score2psl<-data.frame(1001:1750,score2)
write.table(score2psl,"C:\\Program Files\\PARSCALE4\\testscore_30-750.dat",
sep="",row.names=FALSE,col.names=FALSE)
system("score30-750.bat",show.output.on.console = FALSE)
theta.ab2<-read.table("C:\\Program Files\\PARSCALE4\\testscore30-750.SCO",
head=F,fill=T)[(1:750)*2,7]
theta.ab2<-as.matrix(theta.ab2)
#----------------------------score 3-----------------------------#
score3psl<-data.frame(1001:1750,score3)
write.table(score3psl,"C:\\Program Files\\PARSCALE4\\testscore_30-750.dat",
sep="",row.names=FALSE,col.names=FALSE)
```

```
647: system("score30-750.bat",show.output.on.console = FALSE)
648: theta.ab3<-read.table("C:\\Program Files\\PARSCALE4\\testscore30-750.SCO",
649: head=F,fill=T)[(1:750)*2,7]
650: theta.ab3<-as.matrix(theta.ab3)
651: #-----------------------------------------------------------------#
652:
653: #- Computing the rmsq, total scores, skew and kurtosis ----------#
654:
655: rmsq<- sqrt( mean( (theta1 - theta.ab1)^2 ) )
656:
657: score1 <- apply(score1, 1, mean)
658: score2 <- apply(score2, 1, mean)
659: score3 <- apply(score3, 1, mean)
660:
661: score1.skew<-skew(score1)
662: score1.kurtosis<-kurtosis(score1)
663: score2.skew<-skew(score2)
664: score2.kurtosis<-kurtosis(score2)
665: score3.skew<-skew(score3)
666: score3.kurtosis<-kurtosis(score3)
667:
668: theta.ab1skew<-skew(theta.ab1)
669: theta.ab1kurtosis<-kurtosis(theta.ab1)
670: theta.ab2skew<-skew(theta.ab2)
671: theta.ab2kurtosis<-kurtosis(theta.ab2)
672: theta.ab3skew<-skew(theta.ab3)
673: theta.ab3kurtosis<-kurtosis(theta.ab3)
674:
675: #-- Applying additive and multiplicative regression models ------#
676:
677: #Actual theta scores
678: theta.add<-lm(theta3~theta1+theta2)
679: theta.mul<-lm(theta3~theta1+theta2+I(theta1*theta2))
680:
681: #Raw scores
682: sum.add<-lm(score3~score1+score2)
683: sum.mul<-lm(score3~score1+score2+I(score1*score2))
684:
685: #Estimated theta scores
686: theta.ab.add<-lm(theta.ab3~theta.ab1+theta.ab2)
687: theta.ab.mul<-lm(theta.ab3~theta.ab1+theta.ab2+I(theta.ab1*theta.ab2))
688:
689: #-------- Shapiro-Wilk test for checking normality --------------#
690:
691: theta.orderres <- summary(theta.add)$res
692: sum.orderres <- summary(sum.add)$res
693: thetahat.orderres <- summary(theta.ab.add)$res
694:
695: swtheta.p <- round(shapiro.test(theta.orderres)$p,5)
696: swsum.p <- round(shapiro.test(sum.orderres)$p,5)
697: swthetahat.p <- round(shapiro.test(thetahat.orderres)$p,5)
698:
699: #report r-square, r-square change sig., & rm squared deviations #
700:
```

```
701: cat("Working on sample",seednum,"iteration",iter, good.iter, "\n")
702: theta.add.rsq <- summary(theta.add)$r.squared
703: theta.mul.rsq <- summary(theta.mul)$r.squared
704: theta.p <- anova(theta.add, theta.mul)$"Pr(>F)"[2]
705: theta.p[is.na(theta.p)] <- 1.00
706:
707: sum.add.rsq <- summary(sum.add)$r.squared
708: sum.mul.rsq <- summary(sum.mul)$r.squared
709: sum.p <- round(anova(sum.add, sum.mul)$"Pr(>F)"[2], 4)
710: sum.p[is.na(sum.p)] <- 1.00
711:
712: theta.ab.add.rsq <- summary(theta.ab.add)$r.squared
713: theta.ab.mul.rsq <- summary(theta.ab.mul)$r.squared
714: theta.ab.p<-round(anova(theta.ab.add,theta.ab.mul)$"Pr(>F)"[2], 4)
715: theta.ab.p[is.na(theta.ab.p)] <- 1.00
716:
717: #------------ Summarize results of each loop --------------------#
718:
719: iter.results<-as.vector(c(iter,
720: seednum,
721: numItem,
722: a.low,
723: a.high,
724: b.mean,
725: b.sd,
726: w1,
727: w2,
728: theta.add.rsq,
729: theta.mul.rsq,
730: theta.p,
731: sum.add.rsq,
732: sum.mul.rsq,
733: sum.p,
734: theta.ab.add.rsq,
735: theta.ab.mul.rsq,
736: theta.ab.p,
737: rmsq,
738: alpha.score1,
739: alpha.score2,
740: alpha.score3,
741: swtheta.p,
742: swsum.p,
743: swthetahat.p,
744: score1.skew,
745: score2.skew,
746: score3.skew,
747: score1.kurtosis,
748: score2.kurtosis,
749: score3.kurtosis,
750: theta.ab1skew,
751: theta.ab2skew,
752: theta.ab3skew,
753: theta.ab1kurtosis,
754: theta.ab2kurtosis,
```

```
755: theta.ab3kurtosis))
756:
757: names(iter.results)<-NULL
758: sink(results.file,append=TRUE)
759: print(iter.results,digits=4,quote=FALSE)
760: sink()
761: good.iter<-good.iter+1
762: }
763:
764: }
765: #-------------------- End loop structure ----------------------#
766:
767: #================================================================#
768: # Begin looping individual conditions #
769: #================================================================#
770:
771: options(width=2000)
772:
773: {
774:
775: results.file26<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C26.txt"
776: study1(seednum = 26,
777: numSubj = 750,
778: Numiter = n.it,
779: b.mean = -2.0,
780: b.sd = 0.35,
781: a.low = .31,
782: a.high = .58,
783: w1 = .3,
784: w2 = .3,
785: numItem = 30,
786: results.file = results.file26)
787:
788: results.file28<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C28.txt"
789: study1(seednum = 28,
790: numSubj = 750,
791: Numiter = n.it,
792: b.mean = -2.0,
793: b.sd = 0.35,
794: a.low = .31,
795: a.high = .58,
796: w1 = .5,
797: w2 = .5,
798: numItem = 30,
799: results.file = results.file28)
800:
801: results.file30 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C30.txt"
802: study1(seednum = 30,
803: numSubj = 750,
804: Numiter = n.it,
805: b.mean = -2.0,
806: b.sd = 0.35,
807: a.low = .58,
808: a.high = 1.13,
```

```
w1 = .3,
w2 = .3,
numItem = 30,
results.file = results.file30)

results.file32 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C32.txt"
study1(seednum = 32,
numSubj = 750,
Numiter = n.it,
b.mean = -2.0,
b.sd = 0.35,
a.low = .58,
a.high = 1.13,
w1 = .5,
w2 = .5,
numItem = 30,
results.file = results.file32)

results.file34<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C34.txt"
study1(seednum = 34,
numSubj = 750,
Numiter = n.it,
b.mean = -0.5,
b.sd = 0.35,
a.low = .31,
a.high = .58,
w1 = .3,
w2 = .3,
numItem = 30,
results.file = results.file34)

results.file36 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C36.txt"
study1(seednum = 36,
numSubj = 750,
Numiter = n.it,
b.mean = -0.5,
b.sd = 0.35,
a.low = .31,
a.high = .58,
w1 = .5,
w2 = .5,
numItem = 30,
results.file = results.file36)

results.file38<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C38.txt"
study1(seednum = 38,
numSubj = 750,
Numiter = n.it,
b.mean = -0.5,
b.sd = 0.35,
a.low = .58,
a.high = 1.13,
w1 = .3,
w2 = .3,
```

```
numItem = 30,
results.file = results.file38)

results.file40 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C40.txt"
study1(seednum = 40,
numSubj = 750,
Numiter = n.it,
b.mean = -0.5,
b.sd = 0.35,
a.low = .58,
a.high = 1.13,
w1 = .5,
w2 = .5,
numItem = 30,
results.file = results.file40)

results.file42 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C42.txt"
study1(seednum = 42,
numSubj = 750,
Numiter = n.it,
b.mean = 1.0,
b.sd = 0.35,
a.low = .31,
a.high = .58,
w1 = .3,
w2 = .3,
numItem = 30,
results.file = results.file42)

results.file44<-"C:/Documents and Settings/Admin/Desktop/DissModel2/C44.txt"
study1(seednum = 44,
numSubj = 750,
Numiter = n.it,
b.mean = 1.0,
b.sd = 0.35,
a.low = .31,
a.high = .58,
w1 = .5,
w2 = .5,
numItem = 30,
results.file = results.file44)

results.file46 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C46.txt"
study1(seednum = 46,
numSubj = 750,
Numiter = n.it,
b.mean = 1.0,
b.sd = 0.35,
a.low = .58,
a.high = 1.13,
w1 = .3,
w2 = .3,
numItem = 30,
results.file = results.file46)
```

```
917:
918: results.file48 <-"C:/Documents and Settings/Admin/Desktop/DissModel2/C48.txt"
919: study1(seednum = 48,
920: numSubj = 750,
921: Numiter = n.it,
922: b.mean = 1.0,
923: b.sd = 0.35,
924: a.low = .58,
925: a.high = 1.13,
926: w1 = .5,
927: w2 = .5,
928: numItem = 30,
929: results.file = results.file48)
930:
931: }
932:
933: #=================================================================#
934: # Summarize Results for Table 4 (n=750, restricted) #
935: #=================================================================#
936:
937: n.it<-1000
938:
939: results.list<-paste("C:/Documents and Settings/Admin/Desktop/DissModel2/C",
940: 25:48,sep="")
941: results.list<-paste(results.list,".txt",sep="")
942:
943: type1.theta<-rep(0,24)
944: type1.sum <- rep(0,24)
945: type1.thetahat <- rep(0,24)
946: rdiff.theta <- rep(0,24)
947: rdiff.sum <- rep(0,24)
948: rdiff.thetahat <- rep(0,24)
949: mn.rmsq<-rep(0,24)
950: pvalue.score1.mn<-rep(0,24)
951: pvalue.score2.mn<-rep(0,24)
952: pvalue.score3.mn<-rep(0,24)
953: pvalue.score1.sd<-rep(0,24)
954: pvalue.score2.sd<-rep(0,24)
955: pvalue.score3.sd<-rep(0,24)
956: alpha.score1<-rep(0,24)
957: alpha.score2<-rep(0,24)
958: alpha.score3<-rep(0,24)
959: sw.theta.p<-rep(0,24)
960: sw.sum.p<-rep(0,24)
961: sw.thetahat.p<-rep(0,24)
962:
963: skew.score1<-rep(0,24)
964: skew.score2<-rep(0,24)
965: skew.score3<-rep(0,24)
966:
967: kurtosis.score1<-rep(0,24)
968: kurtosis.score2<-rep(0,24)
969: kurtosis.score3<-rep(0,24)
970:
```

```
skew.theta.ab1<-rep(0,24)
skew.theta.ab2<-rep(0,24)
skew.theta.ab3<-rep(0,24)

kurtosis.theta.ab1<-rep(0,24)
kurtosis.theta.ab2<-rep(0,24)
kurtosis.theta.ab3<-rep(0,24)

for(i in 1:24) {
infile <- read.table(results.list[[i]], header=FALSE)
infile<- infile[,2:ncol(infile)]
names(infile)<-list( "iter",
"seednum",
"numItem",
"a.low",
"a.high",
"b.mean",
"b.sd",
"w1",
"w2",
"theta.add.rsq",
"theta.mul.rsq",
"theta.p",
"sum.add.rsq",
"sum.mul.rsq",
"sum.p",
"theta.ab.add.rsq",
"theta.ab.mul.rsq",
"theta.ab.p",
"rmsq",
"alpha.score1",
"alpha.score2",
"alpha.score3",
"swtheta.p",
"swsum.p",
"swthetahat.p",
"skew.score1",
"skew.score2",
"skew.score3",
"kurtosis.score1",
"kurtosis.score2",
"kurtosis.score3",
"skew.theta.ab1",
"skew.theta.ab2",
"skew.theta.ab3",
"kurtosis.theta.ab1",
"kurtosis.theta.ab2",
"kurtosis.theta.ab3")

write.table(infile,
"C:/Documents and Settings/Admin/Desktop/DissModel2/restr 750 full.txt")

type1.theta[i]<-sum(infile["theta.p"] <= .05)/n.it
type1.sum[i]<-sum(infile["sum.p"] <=.05)/n.it
```

```
type1.thetahat[i]<-sum(infile["theta.ab.p"] <=.05)/n.it

rdiff.theta[i]<-round(sum(infile["theta.mul.rsq"]-infile["theta.add.rsq"])/
n.it,2)
rdiff.sum[i]<-round(sum(infile["sum.mul.rsq"]-infile["sum.add.rsq"])/n.it,2)
rdiff.thetahat[i]<-round(sum(infile["theta.ab.mul.rsq"] -
infile["theta.ab.add.rsq"])/n.it,2)
mn.rmsq[i]<-round(mean(infile["rmsq"]),2)
alpha.score1[i]<-round(mean(infile["alpha.score1"]),2)
alpha.score2[i]<-round(mean(infile["alpha.score2"]),2)
alpha.score3[i]<-round(mean(infile["alpha.score3"]),2)

sw.theta.p[i]<-round(sum(infile["swtheta.p"] > .05)/n.it,5)
sw.sum.p[i]<-round(sum(infile["swsum.p"] > .05)/n.it,5)
sw.thetahat.p[i]<-round(sum(infile["swthetahat.p"] > .05)/n.it,5)

skew.score1[i]<-round(mean(infile["skew.score1"]),5)
skew.score2[i]<-round(mean(infile["skew.score2"]),5)
skew.score3[i]<-round(mean(infile["skew.score3"]),5)
kurtosis.score1[i]<-round(mean(infile["kurtosis.score1"]),5)
kurtosis.score2[i]<-round(mean(infile["kurtosis.score2"]),5)
kurtosis.score3[i]<-round(mean(infile["kurtosis.score3"]),5)
skew.theta.ab1[i]<-round(mean(infile["skew.theta.ab1"]),5)
skew.theta.ab2[i]<-round(mean(infile["skew.theta.ab2"]),5)
skew.theta.ab3[i]<-round(mean(infile["skew.theta.ab3"],na.rm=TRUE),5)
kurtosis.theta.ab1[i]<-round(mean(infile["kurtosis.theta.ab1"]),5)
kurtosis.theta.ab2[i]<-round(mean(infile["kurtosis.theta.ab2"]),5)
kurtosis.theta.ab3[i]<-round(mean(infile["kurtosis.theta.ab3"],na.rm=TRUE),5)

}

n <- c(rep(750,24))
b <- c(rep("N(-1.5,1.0)",8),rep("N(0,1)",8),rep("N(1.5,1.0)",8))
a <- c(rep("U(0.31, 0.58)",4),rep("U(0.58, 1.13)",4))
a <- rep(a,3)
B1B2 <- rep(c(.3,.3,.5,.5),6)
Items<-rep(c(15,30),12)

mean.alpha<-round(apply(cbind(alpha.score1,alpha.score2,alpha.score3),
1,mean),2)

type1.theta<-round(type1.theta,2)
type1.sum<-round(type1.sum,2)
type1.thetahat<-round(type1.thetahat,2)

sw.theta.p<-round(sw.theta.p,2)
sw.sum.p<-round(sw.sum.p,2)
sw.thetahat<-round(sw.thetahat.p,2)

sktab1<-round(data.frame(skew.score3, kurtosis.score3, skew.theta.ab3,
kurtosis.theta.ab3),2)

table4<-data.frame(n,b,a,B1B2,Items,type1.theta,type1.sum,type1.thetahat,
mean.alpha,sw.theta.p,sw.sum.p,sw.thetahat,sktab1)
```

```
1079:
1080: print(table4)
1081: write.table(table4,
1082: "C:/Documents and Settings/Admin/Desktop/DissModel2/Table4 restr 750.txt")
1083: #======End simulation for Table 4 (n=750, restricted)==========#
```

APPENDIX E: BATCH FILES FOR PARSCALE INTEGRATION

Filename: testscore_15-250.bat

```
1: C:\Program Files\PARSCALE4
2: psl0 testscore_15-250
3: psl1 testscore_15-250
4: psl2 testscore_15-250
5: psl3 testscore_15-250
6: exit
```

Filename: testscore_30-250.bat

```
1: C:\Program Files\PARSCALE4
2: psl0 testscore_30-250
3: psl1 testscore_30-250
4: psl2 testscore_30-250
5: psl3 testscore_30-250
6: exit
```

Filename: testscore_15-750.bat

```
1: C:\Program Files\PARSCALE4
2: psl0 testscore_15-750
3: psl1 testscore_15-750
4: psl2 testscore_15-750
5: psl3 testscore_15-750
6: exit
```

Filename: testscore_30-750.bat

```
1: C:\Program Files\PARSCALE4
2: psl0 testscore_30-750
3: psl1 testscore_30-750
4: psl2 testscore_30-750
5: psl3 testscore_30-750
6: exit
```

APPENDIX F: SYNTAX FILES FOR PARSCALE INTEGRATION

Filename: testscore_15-250.PSL

```
1: >COMMENT batch run (n=250, k=15)
2: >FILES       DFNAME='testscore_15-250.dat',SAVE;
3: >SAVE        SCORE='testscore15-250.SCO';
4: >INPUT       NTEST=1, NIDCH=4, NTOTAL=15, LENGTH=15;
5: (4A1,15A1)
6: >SCALE1      ITEM=(1(1)15), NBLOCK=1;
7: >BLOCKS      NITEMS=15, NCAT=5, ORIGINAL=(1,2,3,4,5);
8: >CALIB       GRADED,NORMAL,SPRIOR,TPRIOR,ITEMFIT=20;
9: >SCORE       EAP,DIST=2,ITERATION=(0.001,50);
```

Filename: testscore_30-250.PSL

```
1: >COMMENT batch run (k=30, n=250)
2: >FILES       DFNAME='testscore_30-250.dat',SAVE;
3: >SAVE        SCORE='testscore30-250.SCO';
4: >INPUT       NTEST=1, NIDCH=4, NTOTAL=30, LENGTH=30;
5: (4A1,30A1)
6: >SCALE1      ITEM=(1(1)30), NBLOCK=1;
7: >BLOCKS      NITEMS=30, NCAT=5, ORIGINAL=(1,2,3,4,5);
8: >CALIB       GRADED,NORMAL,SPRIOR,TPRIOR,ITEMFIT=20;
9: >SCORE       EAP,DIST=2,ITERATION=(0.001,50);
```

Filename: testscore_15-750.PSL

```
1: >COMMENT batch run (k=15, n=750)
2: >FILES       DFNAME='testscore_15-750.dat',SAVE;
3: >SAVE        SCORE='testscore15-750.SCO';
4: >INPUT       NTEST=1, NIDCH=4, NTOTAL=15, LENGTH=15;
5: (4A1,15A1)
6: >SCALE1      ITEM=(1(1)15), NBLOCK=1;
7: >BLOCKS      NITEMS=15, NCAT=5, ORIGINAL=(1,2,3,4,5);
8: >CALIB       GRADED,NORMAL,SPRIOR,TPRIOR,ITEMFIT=20;
9: >SCORE       EAP,DIST=2,ITERATION=(0.001,50);
```

Filename: testscore_30-750.PSL

```
1: >COMMENT batch run (k=30, n=250)
2: >FILES       DFNAME='testscore_30-750.dat',SAVE;
3: >SAVE        SCORE='testscore30-750.SCO';
4: >INPUT       NTEST=1, NIDCH=4, NTOTAL=30, LENGTH=30;
5: (4A1,30A1)
6: >SCALE1      ITEM=(1(1)30), NBLOCK=1;
7: >BLOCKS      NITEMS=30, NCAT=5, ORIGINAL=(1,2,3,4,5);
8: >CALIB       GRADED,NORMAL,SPRIOR,TPRIOR,ITEMFIT=20;
9: >SCORE       EAP,DIST=2,ITERATION=(0.001,50);
```