5-2012

Utility of a goodness-of-fit index for the graded response model with small sample sizes : a Monte Carlo investigation.

Christina Ruth Studts 1971-
University of Louisville

UTILITY OF A GOODNESS-OF-FIT INDEX FOR THE GRADED RESPONSE MODEL WITH SMALL SAMPLE SIZES: A MONTE CARLO INVESTIGATION

by

Christina Ruth Studts
B.A., University of Notre Dame, 1993
M.S.W., University of Kentucky, 1997
Ph.D., University of Louisville, 2008

A Thesis
Submitted to the Faculty of the
School of Public Health and Information Sciences of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Master of Science

Department of Bioinformatics and Biostatistics
School of Public Health and Information Sciences
University of Louisville
Louisville, Kentucky

May 2012

UTILITY OF A GOODNESS-OF-FIT INDEX FOR THE

GRADED RESPONSE MODEL WITH SMALL SAMPLE SIZES:

A MONTE CARLO INVESTIGATION

by

Christina Ruth Studts
B.A., University of Notre Dame, 1993
M.S.W., University of Kentucky, 1997
Ph.D., University of Louisville, 2008

A Thesis Approved on

March 28, 2012

by the following Thesis Committee:

Guy Brock, Ph.D., Thesis Director

Somnath Datta, Ph.D.

L. Jane Goldsmith, Ph.D.

Michiel A. van Zyl, Ph.D.

DEDICATION

To Jamie, Shannon Kathleen, and Cait Marie…

…mo mhíle grá,

tá mo chroí istigh ionat.

ACKNOWLEDGEMENTS

I cannot adequately thank my advisor, Dr. Guy Brock, for his support, patience, and guidance as I completed this thesis. This project was undertaken in the midst of major life changes, and its completion is owed in great part to Dr. Brock's encouragement and positive attitude. My appreciation also goes to Dr. Somnath Datta and Dr. Jane Goldsmith, committee members from the Department of Bioinformatics and Biostatistics, who readily agreed to serve on my committee and offer their feedback to improve this thesis. Dr. Riaan van Zyl was my dissertation chair in the Kent School of Social Work; he first introduced me to item response theory and is, therefore, partly responsible for the direction I took with this project. His gracious agreement to serve as an outside reader on my committee was a relief and a welcome link to my previous training. Finally, my love and gratitude go to my husband, Jamie; my daughters, Shannon and Cait (both of whom arrived while this project was a work in progress); my sister, Jeannie Mabey, and her husband, Jonas, and son, Caleb; and my parents, Bob and Kathy Clark. Without my family, I would not have had the motivation to persevere.

ABSTRACT

UTILITY OF A GOODNESS-OF-FIT INDEX FOR THE GRADED RESPONSE MODEL WITH SMALL SAMPLE SIZES: A MONTE CARLO INVESTIGATION

Christina R. Studts

March 28, 2012

Item response theory (IRT) is expanding to diverse research settings, without accompanying access to easily implemented model fit methods. One simple model fit approach involves χ^2/df ratios. However, its utility is not known across several conditions salient to recent applied IRT research. A Monte Carlo simulation was implemented to investigate the effects of several factors (sample size, adjustment condition, type of misfit, and proportion of misfitting items) on χ^2/df ratios in the context of the Graded Response Model. Results suggested that: (a) adjusted χ^2/df ratios were appropriate for the largest sample size condition (N=10000), but were extremely inflated for small (N=400) and medium (N=1500) conditions; (b) χ^2/df ratios were differentially affected across sample sizes by type and amount of misfit; and (c) sensitivity of the $\chi^2/df > 3$ cut point for identifying misfit in single items was notably low across all study conditions. Implications, limitations, and future directions are discussed.

TABLE OF CONTENTS

PAGE

CHAPTER

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Item response theory (IRT) models, developed in the educational and psychological testing fields, are gaining prominence in health research and the social sciences. These models are used to analyze the response patterns of individuals to sets of items which are scored categorically (i.e., each item or question is scored using either binary or polytomous ordered or unordered response options). When a given set of items is intended to measure a latent variable of interest (e.g., intelligence, depression, health-related quality of life), IRT models can be fit to provide estimates of the measurement properties of individual items, groupings of items, and the set of items as a whole, as well as to estimate individual respondents' levels of the latent variable of interest.

Most IRT models are generalized linear fixed-effect or mixed-effect models, incorporating parameters which characterize certain qualities of each item in the given set (Hambleton & Swaminathan, 1985). Item *difficulty* (i.e., location) and *discrimination* (i.e., slope) are the most frequently included item-level parameters in these models. By estimating item difficulty and discrimination parameters, IRT models facilitate comparisons of the amount of measurement information provided by items at specific levels of the latent variable of interest (Baker & Kim, 2004). In addition, the consistency of item parameter estimates can be evaluated for disparate groups of respondents, allowing for the investigation of *differential item functioning* (DIF), or item bias, among different groups of respondents (Teresi, 2001). These and other products of the fitting of IRT models are employed in a wide range of practical applications, described further in Chapter II.

Because the development of IRT models was primarily within the fields of educational and psychological standardized testing, sample sizes exceeding 10,000 respondents are common in many applications. However, recent advances in the availability of user-friendly statistical software capable of fitting IRT models, paired with increased interest in the use of these models in a wide range of research settings and fields, have yielded many applications of IRT methods employing sample sizes as low as 200. While the precision and reliability of parameter estimation has been examined for a range of sample sizes and IRT models (Tay-Lim & Harwell, 1997), one area relevant to the use of small samples which has not yet benefitted from extensive, systematic investigation is the evaluation of model fit, for which there is no consensus in the literature regarding best approaches.

For IRT models designed for polytomous items, one fairly simple index of model fit is Drasgow and colleagues' (1995) chi-square to degrees of freedom ratio (χ^2/df). This method of investigating item-fit to a given IRT model involves calculations of χ^2/df ratios for all single items, pairs of items, and triplets of items in a given set (i.e., in the measurement instrument of interest), comparing observed response pattern counts to those expected based on the IRT model fitted. In general, χ^2/df ratios exceeding 3 are described by Drasgow et al. (1995) as indicating moderately large to large degrees of misfit, and a rule of thumb setting 3 as a cut point for misfit has been employed by several authors. Another frequently used convention suggested by Drasgow et al. is the adjustment of sample sizes used to calculate the χ^2/df ratios from the actual sample size to a standard sample size (N = 3000), thus allowing comparisons of model fit across studies with differing sample sizes.

While Drasgow and colleagues developed this approach in the context of large-scale educational testing applications ($N > 10{,}000$), others have recently used the χ^2/df ratio index of model fit in studies examining such diverse issues as health-related quality of life (Fryback, Palta, Cherepanov, Bolt, & Kim, 2010); attention-deficit hyperactivity disorder in children (Gomez, 2008); cultural equivalence of measures of depression (Kim, Chiriboga, & Jang, 2009); forensic psychopathy (Bolt, Hare, & Neumann, 2007); spiritual wellbeing (Gomez & Fisher, 2005); business leadership (Zagorsek, Stough, & Jaklic, 2006); financial risk-taking (Lampenius & Zickar, 2005); emotional intelligence (Cooper & Petrides, 2010); sexual harassment in the military (Estrada, Probst, Brown, & Graso, 2011; Stark, Chernyshenko, Lancaster, Drasgow, & Fitzgerald, 2002); military attrition (Stark, Chernyshenko, Drasgow, Lee, White, & Young, 2011); and personality assessment (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Maydeu-Olivares, 2005; Robie, Zickar, & Schmit, 2001; Schmidt, Kihm, & Robie, 2000; Zickar & Drasgow, 1996). Sample sizes in these investigations ranged from under 300 (Lampenius & Zickar, 2005) to nearly 72,000 (Estrada et al., 2011), illustrating the multitude of settings and designs characterizing current applied research utilizing IRT methods.

Notably, however, the performance of the χ^2/df index of model fit has not been investigated systematically across the sets of conditions likely to be encountered in practical IRT research. Factors including sample size, type of misfit, and percentage of misfitting items within a given set may influence the performance of χ^2/df ratios calculated for single items, pairs, and triplets of items. Clarification of these issues may facilitate appropriate use and interpretation of the χ^2/df ratios approach to assess model fit in future research employing Samejima's (1969) graded response model (GRM), the primary focus of this study.

CHAPTER II

REVIEW OF THE LITERATURE

As introduced briefly in Chapter I, item response theory (IRT) comprises a class of latent variable models that utilize a given set of observed variables (i.e., item responses) to measure a single underlying latent variable of interest (Hambleton & Swaminathan, 1985). In this chapter, a brief overview of IRT will be provided, addressing the assumptions of IRT models, several models developed for use with binary items, and several models developed for use with polytomous items. A more detailed description of Samejima's (1969) graded response model (GRM) will be provided, as item fit within the GRM comprises the focus of this study. In addition, examples of practical applications of IRT will be offered. Methods for assessing model fit in IRT will be reviewed, distinguishing between person-fit and item-fit approaches. The χ^2/df ratio item-fit method, developed by Drasgow and colleagues (1995) and used in a variety of research settings and conditions, will be described in more detail, particularly as it relates to IRT applications with small samples. Finally, several research questions of interest will be delineated, along with associated hypotheses.

Brief Overview of Item Response Theory

Though the foundations of IRT can be traced to Thurstone's conceptualization of latent traits in the 1920s, the development of this class of models is generally attributed to pioneering work by Lord (1953). Throughout the 1950s and 1960s, psychometric researchers including Lord, Birnbaum, Rasch, and Wright introduced logistic latent variable models and methods for model parameter estimation, highlighting potential applications of IRT methods in education, industry, and psychology (Bock, 1997; Hambleton & Swaminathan, 1985). By the 1980s, advances in computer technology and software expanded the accessibility of IRT methods to researchers and practitioners in measurement-oriented fields (Hambleton & Jones, 1993).

At its core, IRT consists of a set of generalized linear models which estimate the probability of a particular response to an item based upon (a) the level of the latent trait possessed by the respondent, and (b) certain stable characteristics of the item (Embretson & Reise, 2000). For a given item with ordered response options measuring a latent variable, the probability of endorsement of a higher response category should rise as a respondent's level of the latent variable increases. The simplest application of IRT modeling is to *binary* items. In knowledge-based testing, such items may be scored as *correct* or *incorrect*, while in trait- or symptom-type testing, they may be scored as *endorsed* or *not endorsed* (Embretson & Reise, 2000). A more complex application is to *polytomous* items, including items with either ordered (e.g., Likert-type) or unordered (e.g., nominal multiple choice) response options (Hambleton & Swaminathan, 1985). For most types of items, the probability of a randomly selected individual's response to an item is represented as a logistic monotonic function of the level of the latent variable, determined by certain item characteristics. This relationship is graphically represented by the *item characteristic curve* (ICC) for dichotomous items, and by *option characteristic curves* (OCCs) for polytomous items (sometimes referred to as category response curves; Hambleton & Swaminathan, 1985).

Assumptions of IRT

Item response theory models typically rely on three assumptions: (a) unidimensionality, (b) conditional independence, and (c) monotonicity. For the following discussion, X_{vi} is the response of individual $v \in \{1, \ldots, N\}$ to item $i \in \{1, \ldots, J\}$, and each item is scored on a categorical scale from $m = 0, \ldots, K_i$.

The *unidimensionality* assumption requires that there is a single, one-dimensional latent trait possessed by each respondent in the sample that fully accounts for each individual respondent's propensity to select a particular response to a given item. This propensity, or the level of the latent variable in individual v, is customarily denoted by θ_v.

Given θ_v, the assumption of *conditional independence* requires that the elements of respondent v's item response vector, $\boldsymbol{X}_v = (X_{v1}, \ldots, X_{vJ})^{\mathrm{T}}$, are independent. Thus, θ_v alone determines a respondent's pattern of responses to items i_1 through i_J.

The *monotonicity* assumption requires that $Pr\{X_{vi} > t \mid \theta_v\}$ be a non-decreasing function of the individual respondent's propensity θ_v, for all i and for all $t \in \Re$. Thus, respondents with high θ_v are more likely to select higher item response options than those with low θ_v.

Models for Binary Items

A simple example of a basic IRT model is one frequently applied with binary items: the two-parameter logistic model (2PL), originally proposed by Birnbaum (1968). This model illustrates several common features of most IRT models:

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

(1) provides the 2PL *item characteristic function* for a binary item (i.e., correct/incorrect, true/false, etc.). In the 2PL, $P_i(\theta)$ represents the probability of the endorsement of item *i*, given a particular level of the latent variable, distributed as $\theta \sim N(0,1)$. The mathematical constant *e* is the base of the natural logarithm. The mathematical constant *D* represents an optional scaling factor, generally set to 1.7; this value makes the item parameters from logistic IRT models very similar to the item parameters that would be obtained in normal-ogive IRT models (Hambleton & Swaminathan, 1985). The difficulty of item *i* is represented by b_i, and refers to the level of the latent variable (θ) at which the probability of item endorsement is equal to 0.5. The discrimination of item *i* is represented by a_i, a value proportional to the slope of the tangent line to the item characteristic function at its steepest point, which is at its difficulty level (i.e., at b_i). Steeper slope of the curve at this point is associated with greater precision of discrimination between respondents at similar levels of θ; flatter slopes suggest weaker item capacity to discriminate between respondents.

When the item characteristic function depicted in (1) is graphed for a single item *i* with particular item parameters b_i and a_i over a range of values of θ, the result is the ICC, illustrated for a hypothetical binary item in Figure 1. Several features of the ICC graph are notable. First, the range of the latent variable θ depicted on the *x*-axis generally extends from -3.0 to +3.0, where θ is arbitrarily scaled to have a mean of 0 and standard deviation of 1.0 (i.e., $\theta \sim \mathrm{N}(0,1)$). The probability of item endorsement asymptotically approaches 0 at decreasing levels of θ and 1.0 at increasing levels of θ. For the illustrated hypothetical item with difficulty level $b_i = 0.25$ and discrimination level $a_i = 1.0$, the probability of item endorsement for respondents with a latent trait level 1 standard deviation below the mean is approximately 0.20; for respondents with latent trait levels 2 standard deviations above the mean, the probability of endorsement is approximately 0.85; and for respondents at the mean latent trait level, the probability of endorsement is approximately 0.45.

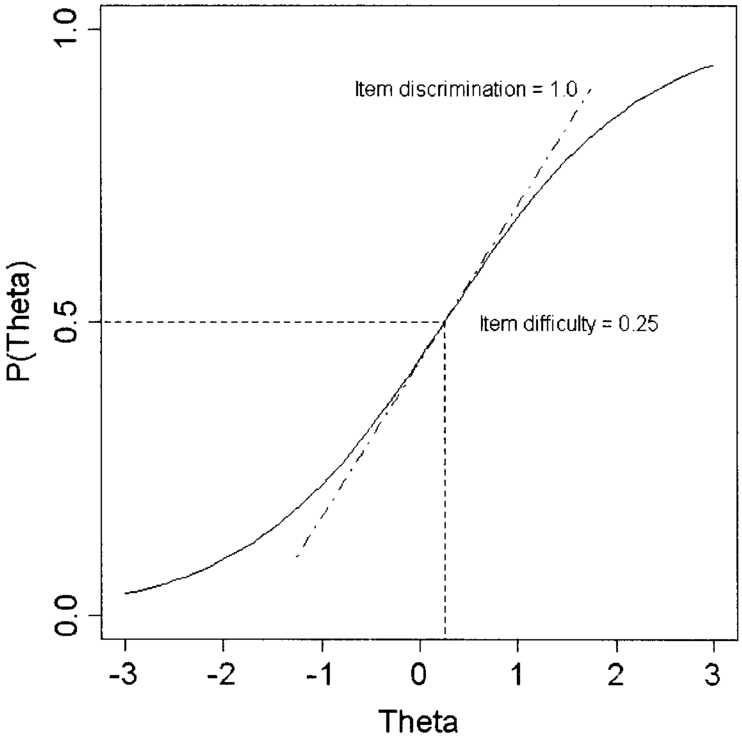

***Figure 1.* Item characteristic curve (ICC) for a hypothetical item in the two-parameter logistic model (2PL; b_i = 0.25, a_i = 1.00).**

In a two-parameter model such as the 2PL, both item parameters can vary between items. Thus, items can differ in their difficulty levels (i.e., location), as well as in their discrimination levels (i.e., slope). One-parameter models exist which constrain the discrimination levels of all items to be equal (usually at a = 1.0), and these models are often referred to as *Rasch* models, for their developer (Hambleton & Swaminathan, 1985). In addition, three-parameter models are possible, which include an additional parameter (c_i) allowing the lower asymptote of the ICC to be greater than 0; these models are often applied to knowledge-testing items, in which the probability of guessing correctly increases the base level of probability of a correct response (Embretson & Reise, 2000).

In Figure 2, three hypothetical ICCs in the 2PL are depicted with differing difficulty and discrimination parameters. In creating a measurement instrument, if one were interested in including items which precisely measured respondents with levels of the latent trait between 1 and 2 standard deviations above the mean, of these three items, Item 3 would be the most informative. For Item 1 (b_1 = -2.0, a_1 = 1.2), all respondents with θ levels above the mean would share high probabilities of endorsing the item. For Item 2 (b_2 = 0.0, a_2 = 0.5), the probabilities of item endorsement change very slowly for the θ levels of interest, obscuring distinctions between respondents at similar, but not identical, levels of θ. In contrast, Item 3 (b_3 = 1.5, a_3 = 1.8) can discriminate well between respondents at the desired levels of θ. This example illustrates the applicability of IRT modeling to the identification and selection of items with specific, desired measurement properties.

Models for Polytomous Items

For polytomous items, multiple functions characterize each item, each representing the probability of choosing a particular item response option given a specific level of the latent variable (Hambleton & Swaminathan, 1985). In a polytomous item, the probability of choosing a particular response option is a function of the levels of the latent variable; if response options are ordered, respondents with higher levels of θ are

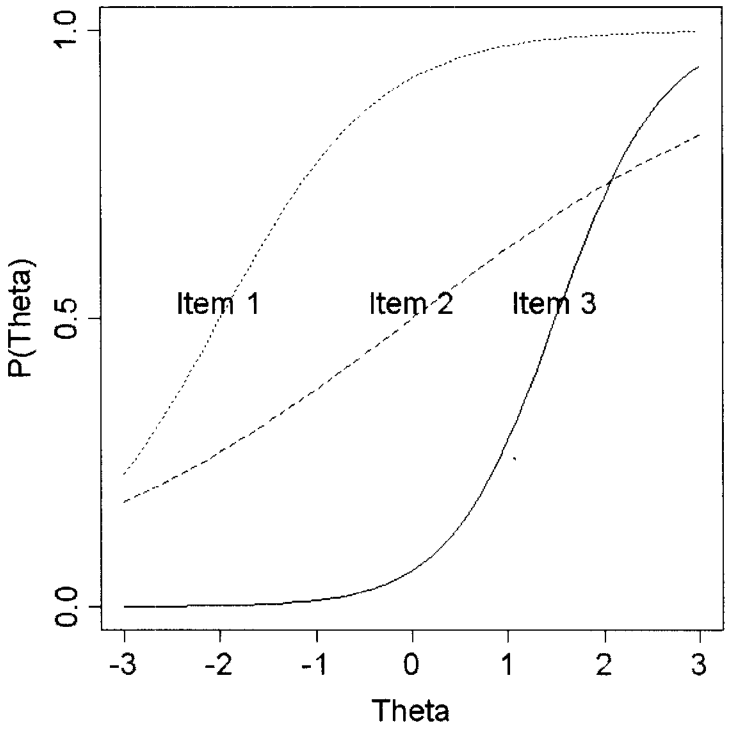

Figure 2. **Three hypothetical item characteristic curves (ICCs) with differing item parameters (b_1 = -2.0, a_1 = 1.2; b_2 = 0.0, a_2 = 0.5; and b_3 = 1.5, a_3 = 1.8).**

more likely to choose higher response options. These *option characteristic functions* can be graphically represented by OCCs, just as binary item characteristic functions are depicted by ICCs. The points of intersection of the OCCs for a single polytomous item indicate the levels of θ at which shifts in selection of response options are most likely for that item. Points of intersection of OCCs are referred to as difficulty thresholds, of which there are always one fewer than response options.

Many IRT models have been developed which can be applied to items with multiple nominal response categories (Bock, 1972), as well as to items with Likert-type polytomous ratings (i.e., those with ordered response options). Models for polytomous items with ordered response options include the graded response model (Samejima, 1969), the partial credit model (Masters, 1982), the ordinal model (Thissen & Steinberg, 1986), and the generalized partial credit model (Muraki, 1992). The current study will focus on the graded response model (Samejima, 1969), described in detail below.

The Graded Response Model. When item responses can be ordered into more than two categories along a continuum, Samejima's (1969) graded response model (GRM) may be an appropriate polytomous IRT model. While dichotomization of polytomous item responses is often conducted to allow fitting of simpler IRT models (e.g., the Rasch or 2PL models), preservation of the ordinal nature of item responses provides more psychometric information than is yielded by binary models with comparable item parameters (Agresti, 2002; Samejima, 1977). The two-parameter polytomous GRM is an extension of the 2PL described earlier in this chapter, and, as with the 2PL, use of the logistic function in the model is generally preferred to the cumulative normal function to preserve computational efficiency.

In this overview of the GRM, hypothetical items with three ordered response options are used for illustration. Each hypothetical item, therefore, has $K = 3$ ordered response options, coded $k = 0$, 1, and 2. Parallel to the manner in which item characteristic functions are estimated for binary items, in the GRM, *option characteristic functions* must be estimated for each response option in an item (Samejima, 1969). The option characteristic functions are derived from the 2PL presented in (1), by estimating item responses as one of the two dichotomies captured in the cumulative response thresholds: (a) response option 0 versus options 1 and 2; and (b) response options 0 and 1 versus option 2. The probability of endorsing option 0 or higher is defined as 1.0, and the probability of endorsing an option higher than option 2 is defined as 0, since no option higher than 2 is provided. The option characteristic functions associated with a hypothetical item with $K = 3$ ordered response options ($k = 0, 1, 2$) are as follows:

$$P(k_i|\Theta) = \begin{cases} 1 - \frac{e^{Da_i(\Theta - b_{i1})}}{1+e^{Da_i(\Theta - b_{i1})}} & \text{if } k_i = 0 \\ \frac{e^{Da_i(\Theta - b_{i1})}}{1+e^{Da_i(\Theta - b_{i1})}} - \frac{e^{Da_i(\Theta - b_{i2})}}{1+e^{Da_i(\Theta - b_{i2})}} & \text{if } k_i = 1 \\ \frac{e^{Da_i(\Theta - b_{i2})}}{1+e^{Da_i(\Theta - b_{i2})}} & \text{if } k_i = 2 \end{cases} \qquad (2)$$

In (2), $P(k_i \mid \theta)$ represents the probability of the endorsement of response option k for item i, given a particular level of the latent variable, represented by θ. The mathematical constants e and D (the scaling factor which may or may not be used) are identical to their values in the 2PL. The parameter b_{i1} represents the value of θ at the threshold (i.e., intersection) between response options 0 and 1, and the parameter b_{i2} represents the value of θ at the threshold between response options 1 and 2. In the two-parameter polytomous GRM, item discrimination is constrained as constant within item response options, but may vary between items; thus, the parameter a_i refers to the discrimination level of all response options of item i.

A graphical illustration of the GRM for a hypothetical item with three ordered response options clarifies the interpretation of the option characteristic functions presented above. Figure 3 is a graph of the probabilities of endorsement of the response options associated with one such item, conditional on the level of the latent trait being measured. Note that for the lowest levels of θ, the most likely response option to be selected is option 0 (often labeled as *not at all* or *never* in symptom-type items). As the

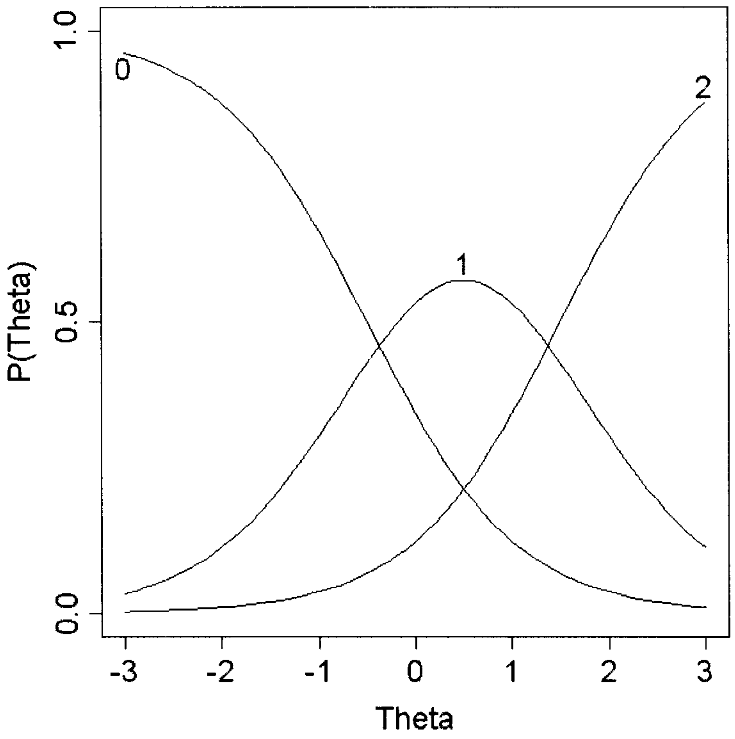

Figure 3. **Graded response model option characteristic curves (OCCs) for a hypothetical item with three response options (*a_i* = 1.3, *$b_{i,1}$* = -0.5, *$b_{i,2}$* = 1.5).**

level of θ increases, the probability that option 0 will be selected gradually lowers, until at θ = -0.5, the probability of endorsing option 0 is equal to the probability of endorsing option 1 (often labeled *sometimes* or *somewhat true* in symptom-type items). This level of θ is equal to the parameter b_{i1}, the threshold between response options 0 and 1. As the level of θ increases, the probability of endorsement of option 1 initially increases but gradually begins to decrease, until at θ = 1.5, the probability of endorsing option 1 is equal to the probability of endorsing option 2 (often labeled *always* or *often true* in symptom-type items). This level of θ is equal to the parameter b_{i2}, the threshold between response options 1 and 2. From this level of θ on, the probability of endorsement of option 2 increases, asymptotically approaching 1.0 as θ increases.

Model-fitting and estimation of the item parameters b_{ik} and a_i can be efficiently achieved using marginal maximum likelihood estimation procedures with an expectation maximization algorithm (Bock & Aitkin, 1981). These procedures are available in the R package *ltm* (Rizopoulus, 2006), which has been demonstrated to recover stable and accurate parameters using the GRM. Once a model is fit to the response patterns of a group of respondents to a set of items with ordinal response options, each item can be described in terms of the difficulty levels associated with the points of intersection between option characteristic functions, as well as in terms of the item's ability to discriminate between respondents at different levels of θ. In addition, the item parameter estimates obtained by fitting the GRM can be used for each of the practical applications of IRT methods described later in this chapter.

This discussion of binary and polytomous IRT models, including the GRM, highlights their potential utility in evaluating the quality of measurement provided by a given item at specific levels of a latent trait. The process of estimating item parameters using a given set of data capturing many individuals' response patterns to a set of items is referred to in IRT applications as *item calibration*, and the resulting parameter estimates provide valuable information for item and scale evaluation (Hambleton & Swaminathan, 1985), as well as for methods of quantifying respondents' levels of the latent variable of interest. A number of practical applications of IRT stem directly from this process.

Practical Applications of Item Response Theory

Calibrating items in IRT applications (i.e., obtaining parameter estimates via fitting an appropriate IRT model) facilitates a range of practical applications in the development, evaluation, refinement, and use of measurement instruments (Embretson &

Reise, 2000). The most obvious use of IRT methods is in allowing detailed descriptions of the performance of individual test items; item difficulty and discrimination parameter estimates characterize the levels of θ at which a given item measures most precisely. Such information allows a test (or other measurement instrument) developer to determine how well a given set of items measures the full range, or desired sub-ranges, of the latent variable of interest; if sections of the θ continuum are not adequately measured by included items, the test developer can locate or develop items to fill those gaps. Similarly, if multiple items measure the same section of the θ continuum, redundant items can be deleted, promoting parsimony and reducing respondent burden. Sets of items can be tailored to measure specific ranges of the latent variable of interest, either broadly or narrowly, eliminating ceiling and floor effects if desired (Hambleton & Swaminathan, 1985). This method can also be used to build so-called "parallel measures," in which different sets of items are used to develop multiple versions of a single measurement instrument. In this context, sets of items with matching difficulty and discrimination parameter estimates are selected and compiled into multiple versions of a test or other measurement instrument; such parallel measures are especially useful in the administration of repeated measures, in which test-retest effects can complicate interpretation of findings. Similarly, multiple existing instruments designed to measure the same latent variable (e.g., the myriad of measures of depression) can be "equated," allowing for cross-instrument comparisons of scores by placing them on the same θ metric (Baker, 1992).

Another important application of IRT methods is in the assessment of differential item functioning (DIF), in which a given item or set of items is characterized by differing difficulty and/or discrimination parameters for disparate groups of respondents who share the same θ levels (Holland & Wainer, 1993). In this scenario, the same set of items can be administered to distinct groups of respondents who differ on some key characteristic (e.g., race, sex, etc.). Next, an appropriate IRT model is fit to the data from each group separately. Parameter estimates can be constrained to be equal for items known to function similarly across groups, while parameter estimates for items under investigation for DIF are free to vary. The estimates obtained from fitting the same IRT model with the subgroups of interest can be tested for differences, and items yielding unequal parameter estimates may be determined to function differently based on the key characteristic defining the subgroups. Differential item functioning has been assessed in numerous measurement instruments targeting a range of constructs (e.g., Teresi, 2001).

In addition to assessing item and test measurement properties, IRT methods can also be applied to measure individual respondents' response-profile quality and consistency, based upon the "known" item characteristics obtained in previous item calibration efforts. The application of IRT methods allows the simultaneous consideration of responses to multiple items, in light of the item parameter estimates previously obtained via item calibration, in determining a respondent's "score," or level of θ (Birnbaum, 1968). In addition, estimates of the likelihood of a given observed response pattern across items can be obtained (e.g., Drasgow, Levine, & Mclaughlin, 1987), to determine both the consistency of individuals' responses and the degree to which the IRT model employed fits the observed response patterns.

Finally, computer-adaptive test (CAT) administration is a rapidly growing field in the practical application of IRT methods (Wainer, 2000; Ware, 2003). CAT development employs item calibration to develop an "item bank" of possible items to be used in the computerized administration of a test or measurement instrument. Algorithms are programmed to determine the item-by-item selection of questions to be posed to an individual respondent in order to iteratively estimate his or her θ level to a predetermined level of precision. This application combines the capacity of IRT methods to obtain detailed descriptive data about individual items and sets of items with their ability to estimate respondents' levels of θ, given known characteristics of each item. The use of CAT administration has been reported to reduce respondent burden and time needed to obtain precise estimates of θ by half (Weiss & Kingsbury, 1984), making this a valuable tool in the development and efficient administration of tests and other measurement instruments.

All of the benefits and advantages of using IRT methods in the development, evaluation, refinement, and use of measurement instruments, however, depend on appropriate model fit. Several methods to assessing model fit have been proposed and used, and the most prevalent approaches are discussed below.

Model Fit in Item Response Theory

In IRT applications, the fit of the model to the data can be assessed in many ways. Relative IRT model fit can be compared between nested models using likelihood ratio tests or comparisons of Akaike's information criterion (Akaike, 1974); more typically, however, item-fit approaches are utilized. In most item-fit approaches to assessing model fit, expected and observed frequencies of an item's response options are compared for various binned levels of the latent variable (θ), based upon the particular IRT model employed. Several specific approaches to this method have been proposed, though the literature reveals no consensus regarding which approach to use with particular models in various research contexts (i.e., with differing sample sizes, potential proportions of misfitting items, etc.).

In the assessment of item fit to determine model fit, several steps are generally taken (Stone & Zhang, 2003). First, item parameter and latent variable estimates are obtained via item calibration. Next, the latent variable continuum is binned into a pre-set number of subgroups based on θ score estimates. Third, the distribution of observed responses is constructed, with respondents categorized into the appropriate binned subgroups along the latent variable continuum. Fourth, expected response distributions are computed for each item response option within each binned subgroup, using probabilities generated by the IRT model employed. Finally, the resulting data are subjected to a range of evaluative approaches, including (a) visual inspection of graphical plots of observed versus expected response frequencies for item response options, and (b) calculation of one or more of several chi-square-based model fit indices, such as Yen's Q_1 (1981), Bock's χ^2 (1972), or McKinley & Mills' likelihood ratio G^2 (1985). More recently, the calculation of likelihood-based item fit indices have been proposed (Orlando & Thissen, 2000, 2003), in which expected item response frequencies are formulated using summed scores for the latent variable (i.e., sums of the item responses across all items for a given respondent), rather than the θ estimates typically used. These indices include $S\text{-}\chi^2$ and $S\text{-}G^2$, both of which have been expanded recently from their development with binary IRT models to now address polytomous IRT models, such as the GRM (see Bjorner, Smith, Stone, & Sun, 2007, for a SAS macro designed to obtain these fit indices for most binary and polytomous IRT models).

Problems with each of these approaches have been noted in the literature, due to issues including the arbitrary choice of intervals of θ used in groupings (Reise, 1990); the effects of error in θ estimates on the calculation of expected frequencies (Stone, 2003); sparcity of data within latent variable groupings along the continuum of θ (Agresti, 2002); and the effects of sample size and associated degrees of freedom on χ^2 test statistics (Agresti, 2002). Further, in applied IRT research, reliance on stand-alone software packages (e.g., MULTILOG and PARSCALE for polytomous models) that do not generate the above model fit indices has posed a problem for some applied researchers.

Chi-square to Degrees of Freedom Ratio Method

A variation of the chi-square-based model fit indices discussed above was proposed by Drasgow and colleagues (1995) and can be implemented easily by applied researchers with a freely available Excel program called MODFIT (Stark, 2002). This method, often referred to as the χ^2/df ratio approach, has been reported in many diverse research applications (e.g., Bolt, Hare, & Neumann, 2007; Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Cooper & Petrides, 2010; Estrada, Probst, Brown, & Graso, 2011; Fryback, Palta, Cherepanov, Bolt, & Kim, 2010; Gomez, 2008; Gomez & Fisher, 2005; Kim, Chiriboga, & Jang, 2009; Lampenius & Zickar, 2005; Maydeu-Olivares, 2005; Robie, Zickar, & Schmidt, 2001; Schmidt, Kihm, & Robie, 2000; Stark, Chernyshenko, Drasgow, Lee, White, & Young, 2011; Stark, Chernyshenko, Lancaster, Drasgow, & Fitzgerald, 2002; Zagorsek, Stough, & Jaklic, 2006; Zickar & Drasgow, 1996). Its use has been recommended in a recent IRT textbook (De Ayala, 2009) as a relatively simple and accessible way to rectify the lack of model fit indices provided in the most commonly-used IRT software packages.

A key step in utilizing the χ^2/df ratio approach to assessing model fit is the use of expected and observed response frequencies from two sets of data: a *calibration* sample and a *validation* sample, respectively. The calibration sample is used to fit the IRT model of interest and obtain the expected response frequencies used in assessing model fit; the validation sample is a disjoint set of respondent response patterns which is used to determine the observed, or empirical, response frequencies. Thus, in most research applications using the χ^2/df ratio approach, samples are randomly split into calibration and validation subsamples to assess item fit.

The χ^2/df ratio approach relies on the calculation of the expected frequency of respondents selecting each response option for a particular individual item, using the calibration sample:

$$E_i(k) = N \int P(v_i = k|\Theta = t)f(t)dt \tag{3}$$

where k is the response option of interest for item i, f(t) is the θ density, ~ N(0,1), and probabilities are obtained from (2). Expected counts for each response option are summed across all values of the latent variable continuum, and observed counts are obtained via the frequencies of item response option choices in the validation sample. The χ^2 statistic is then obtained in the usual manner:

$$\chi_i^2 = \sum_{k=1}^{m} \frac{(O_k - E_k)^2}{E_k} \quad , \tag{4}$$

for item i with $k = (1, \ldots, m)$ response options. The χ^2 statistic is subsequently divided by its associated degrees of freedom to obtain the χ^2/df ratio. Drasgow and colleagues (1995) noted previous findings that χ^2 statistics for individual items are often insensitive to violations of the unidimensionality assumption in IRT (van den Wollenberg, 1982); further, they observed that certain types of misfit cannot be detected in individual items by this method, such as when the observed and predicted response functions cross. In such cases, a χ^2/df ratio computed for a single item may approach zero, despite the existence of actual misfit. Thus, Drasgow and colleagues suggested computing the χ^2/df statistic for single items, pairs of items, and triplets of items within a given set of items, with the expectation that pairs and triples of items with similar misfits will have large χ^2/df values, revealing the misfit. Calculation of the expected frequency of respondents selecting option k for item i and option k' for item i' is achieved with:

$$E_{i,i'}(k,k') = N \int P(v_i = k|\Theta = t)P(v_{i'} = k'|\Theta = t)f(t)dt \tag{5}$$

Cell(s) with expected frequencies < 5 are combined with cell(s) with the next lowest frequencies until all cells contain expected counts ≥ 5. The observed frequencies are obtained from the validation sample via cell counts. The χ^2 statistic for a two-way contingency table is then calculated and divided by its associated degrees of freedom to obtain the χ^2/df ratio for that pair of items. A parallel procedure is conducted with triplets of items.

To facilitate comparisons of this index of model fit across applications using disparate sample sizes, Drasgow et al. (1995) recommended reporting χ^2/df ratios for single items, pairs of items, and triplets of items, adjusted to a standard sample size of N = 3000. They further suggested that mean adjusted χ^2/df ratios > 3 indicate poor model fit.

This method of assessing model fit, including the suggested reporting of frequencies and means of adjusted χ^2/df ratios for single items, pairs of items, and triplets of items, has been employed in a wide variety of research settings, primarily by applied IRT researchers. However, no published resources are available that investigate the performance of this approach across the conditions observed in such research, including variations in sample size, type of item misfit, and proportion of misfitting items in a given set. Because the availability of IRT software has facilitated the application of IRT methods in more research settings than ever before (many of which utilize smaller sample sizes than used in previous applications of IRT in the areas of large-scale educational and psychological assessment), a systematic investigation is warranted of the performance of adjusted and unadjusted χ^2/df ratios for single items, pairs of items, and triplets of items. This study will conduct such an investigation, focusing specifically on the use of the χ^2/df ratio index of model fit for the GRM, one of the most popular IRT models in applied research.

Summary and Research Questions

Item response theory analyses are expanding to a variety of research fields and settings conducting measurement instrument development, evaluation, refinement, and administration. Samejima's (1969) GRM is an IRT model frequently used to analyze response data for items with ordered categorical response options, as seen in Likert-type items typically utilized in health-related and other social science research. The extension of IRT methods to a wide variety of research settings has not been accompanied by easily implemented approaches to assessing model fit, a vital step in the appropriate application of any model-fitting analysis. The χ^2/df ratios method (Drasgow et al., 1995) is one relatively simple approach which has been used in a wide variety of settings, is easily implemented using a free program, and has been recommended as a strategy to remediate the dearth of goodness-of-fit indices provided by stand-alone IRT software (De Ayala, 2009). However, the utility of this method, developed in the context of large-scale educational research settings, has not been assessed across several conditions salient to recent applied IRT research. Three such conditions include applications of IRT with (a) relatively small sample sizes ($N \leq 1500$), (b) items which exhibit misfit for disparate reasons, and (c) sets of items incorporating differing proportions of misfitting items.

Two research questions stem directly from this discussion. The methods for addressing each will be described in Chapter III.

Research Question 1: Are adjusted (to N = 3000) or unadjusted χ^2/df ratios more appropriate for small-sample IRT research?

Research Question 2: As a means of assessing model fit for the GRM, how are the magnitude and utility of χ^2/df ratios affected by (a) sample size, (b) type of item misfit, and (c) proportion of misfitting items in a given set?

CHAPTER III

METHOD

Study Design

In this Monte Carlo study, two research questions were addressed. Research Question 1 addressed the implications of sample size (i.e., small, medium, or large) on the magnitude of unadjusted versus adjusted χ^2/df ratios used to assess item fit in applications of Samejima's (1969) GRM. Research Question 2 targeted the effects of several data characteristics—sample size, type of misfit, and proportion of misfitting items—on the magnitude of χ^2/df ratios, as well as on their ability to correctly identify misfitting items.

Research Question 1: Sample Size and Adjustment Condition

For this question, effects of sample size (factor *A*) and adjustment condition (factor *B*) on the magnitude of mean χ^2/df ratios used to assess item fit were examined, using simulated data. A two-factor experiment with repeated measures on factor *B* was designed. Factor *A*, sample size, included three levels (N = 400, 1500, and 10000), while factor *B*, adjustment condition, comprised two levels (unadjusted versus adjusted to N = 3000). No item misfit was present in the simulated data. See Appendix A for an illustration of the study design for Research Question 1.

Research Question 2: Sample Size, Type of Misfit, and Proportion of Misfitting Items

For this question, effects of sample size (factor *A*), type of misfit (factor *B*), and proportion of misfitting items (factor *C*) on mean χ^2/df ratios used to assess model fit were examined, using simulated data. A fully crossed factorial design with three factors was used. Three levels of sample size (N = 400, 1500, and 10000), three types of item misfit (misfit due to multidimensionality, to DIF, and to generation from a competing model), and two levels of proportion of misfitting items (10% and 33%) were manipulated. See Appendix A for an illustration of the study design for Research Question 2.

Data Simulation

Data were simulated using the *rmvordlogis* function in the *ltm* package (Rizopoulos, 2006) for R: A Language and Environment for Statistical Computing (R Development Core Team, 2012). See Appendix B for simulation code. The *rmvordlogis* function produces multinomial random variates under several polytomous IRT models, including the GRM. Given arguments for desired sample size *n*, a matrix of "true betas" (i.e., item difficulty threshold and discrimination parameters) for each of *p* "test items," and number of response categories *ncatg* for each item, *rmvordlogis* produces a matrix of item responses for the desired number of simulated respondents. For each simulated condition, 1000 replications were generated of *n* sets of item responses to 30 items with 5 response options.

Research Question 1: Sample Size and Adjustment Condition

For the investigation of the effects of sample size and adjustment condition on the χ^2/df ratios used to assess model fit, the *rmvordlogis* function was used to simulate item responses for three levels of sample size: N = 400, 1500, and 10000. These levels represent typical small, medium, and large sample sizes reported in applied IRT research.

Simulated responses were randomly drawn from a Gaussian distribution of the latent construct, θ~N(0,1). The "true betas," or defined parameters for the 30 simulated items, were taken from Bolt (2002), who generated parameters for a set of unidimensional, DIF-free items for use in a Monte Carlo investigation of DIF. These parameters are presented in Appendix C.

For the unadjusted condition, procedures were followed to obtain the χ^2/df ratios described by Drasgow and colleagues (1995), using the approach implemented in the MODFIT program (Stark, 2002). First, simulated respondents in each sample size condition were randomly split into calibration and cross-validation samples of equal size (n = 200 in the small sample size condition; 750 in the medium sample size condition; and 5000 in the large sample size condition). Item parameter estimates and standard errors (calculated using the delta method) for the calibration sample data were obtained by fitting the GRM model, using the *grm* function in the *ltm* package. Probabilities of responses to each item response category were calculated using the *iprob* internal function of *ltm* (D. Rizopoulos, personal communication, May 8, 2009), and expected frequencies were calculated using (3) from Chapter II. Observed frequencies in each cell were obtained from the cross-validation sample. In R, χ^2 statistics for differences in observed versus expected frequencies were obtained for all 30 single items, for all possible pairs of items (i.e., 30 choose 2: ${}_{30}C_2 = 435$), and for all possible triplets of items (i.e., ${}_{30}C_3 = 4060$)[1]. Each χ^2 statistic was then divided by its degrees of freedom, resulting

[1] In MODFIT, a subset of all possible triplets of items is used, in which sets of low-, medium-, and high-difficulty items are selected. This approach was implemented due to computer memory limitations at the time of program development, considering findings from Reckase et al. (1979) regarding systematic measurement differences between low-

in the unadjusted χ^2/df ratios. The mean and variance of the distributions of these ratios were calculated for all single items, pairs of items, and triplets of items in each simulated dataset.

For the adjusted condition, the actual sample size in each cell was proportionately adjusted to result in a total sample size of N = 3000, and the same calculations described above were repeated to generate the adjusted χ^2/df ratios. These were similarly averaged over all single items, pairs of items, and triplets of items in each simulated dataset.

Finally, the proportions of unadjusted and adjusted χ^2/df ratios > 3 within all single items, pairs of items, and triplets of items were determined for each dataset, to allow investigation of the "rule of thumb" often used as a cut point for indication of item misfit.

Research Question 2: Sample Size, Type of Misfit, and Proportion of Misfitting Items

To allow investigation of the effects of sample size, type of misfit, and proportion of misfitting items under analysis, the *rmvordlogis* function was again used to simulate response patterns under the GRM. Simulated responses were randomly drawn from a Gaussian distribution of the latent construct, θ~N(0,1). Sample size levels again included small (N = 400), medium (N = 1500), and large (N = 10000) conditions. Type of misfit comprised three categories: misfit due to (a) multidimensionality, (b) differential item functioning (DIF), and (c) generation from a different polytomous IRT model. Finally, proportion of misfitting items included two levels: 10% misfitting (i.e., 3 out of 30 items exhibited some type of misfit), and 33% misfitting (i.e., 10 of out 30 items exhibited

and high-difficulty items. In this study, it was feasible instead to use all possible triplets, an approach recommended by Drasgow (personal communication, January 7, 2009).

misfit). For this research question, the "true betas" used in data simulation differed under various conditions of the additional two factors, as described below. Item calibration (i.e., obtaining item parameter estimates and standard errors), as well as all calculations regarding χ^2/df ratios, proceeded as described for Research Question 1.

Misfit due to multidimensionality. The first type of misfit refers to the inclusion of items in a given test which measure something other than the latent construct of interest. While many IRT models are thought to be robust to violations of the assumption of unidimensionality, the effect of inclusion of such items on χ^2/df ratios used to assess item fit is unknown. To simulate items measuring a different construct than that measured by the items with parameters provided by Bolt (2002), parameters for a subset of items used in a different study represented the misfitting items, instead of the original "true betas." These parameters were taken from an investigation of GRM performance (Lautenschlager, Meade, & Kim, 2006) with items from the Minnesota Satisfaction Questionnaire (Weiss, 1967), a unidimensional scale measuring a construct disjoint from that measured by Bolt's (2002) items. These parameters are presented in Appendix D. For conditions in which 10% of items exhibited misfit due to multidimensionality, responses to the 27 fitting items were simulated as above, using the unidimensional, DIF-free "true betas" (as presented in Appendix C), while responses to the 3 misfitting items were simulated separately using the last three sets of "true betas" in Appendix D. Similarly, in conditions in which 33% of items exhibited such misfit, responses to the 20 fitting items were simulated as above, while responses to the 10 misfitting items were simulated separately, using the 10 sets of "true betas" in Appendix D.

Misfit due to differential item functioning. With DIF, group responses to a given item differ conditioned on some attribute other than the latent construct of interest. For example, male respondents may be more likely to endorse a lower response option than female respondents, even when they possess the same level of the latent construct. To simulate items exhibiting DIF, during each replication of the simulation, the sample was randomly split into equally sized *focus* and *reference* groups. "True betas" used for items misfitting due to DIF were systematically different for the two groups. These parameters were drawn from Bolt's (2002) investigation of items exhibiting DIF in the GRM, and are presented in Appendix E. For conditions in which 10% of items exhibited misfit due to DIF, the last 3 sets of unidimensional, DIF-free "true betas" (as presented in Appendix C) were replaced with either the last 3 sets of focus group "true betas" or the last 3 sets of reference group "true betas," as depicted in Appendix E. Similarly, in conditions in which 33% of items exhibited such misfit, the last 10 sets of "true betas" in Appendix C were replaced with either the 10 sets of focus group or reference group "true betas" in Appendix E. Following the splitting of the sample and data simulation based on different sets of parameters, the simulated response data were then combined into a single matrix for calculation of the χ^2/df ratios.

Misfit due to generation from a competing model. More than one polytomous IRT model exists for items with ordered response options, and selection of the appropriate model to use can be challenging in some situations. The effect of suboptimal model selection on χ^2/df ratios used to assess item fit is unknown. To simulate misfit due to incorrect model selection, a subset of item parameters were drawn from a competing polytomous IRT model. The *generalized partial credit model* (GPCM; Muraki, 1992) is defined by different cumulative category response functions than the GRM, while still estimating the same number of parameters as the GRM for a given item. Similarly to the GRM, the GPCM provides the probability of responding to a particular response option to a particular item, based upon item characteristics and the respondent's underlying level of the latent construct:

$$P_{ik}(z) = \frac{\exp(\sum_{c=0}^{k} \beta_i(z - \beta_{ic}^*)}{\sum_{r=0}^{m_i} \exp \sum_{c=0}^{r} \beta_i(z - \beta_{ic}^*)}, \tag{6}$$

where $P_{ik}(z)$ represents the probability of responding in category *k* for item *i*, given the level of the latent construct *z*; β_{ic}^* are the category threshold parameters for item *i*; β_i is the discrimination parameter for item *i*; m_i is the number of response categories for item *i*; and

$$\sum_{c=0}^{0} \beta_i(z - \beta_{ic}^*) == 0. \tag{7}$$

To simulate responses to items generated from a different model, GPCM parameters from Bolt (2002) were used as "true betas" for selected misfitting items. These parameters are presented in Appendix F. For conditions in which 10% of items exhibited misfit due to multidimensionality, the last 3 sets of unidimensional, DIF-free "true betas" (as presented in Appendix C) were replaced with the last 3 sets of "true betas" in Appendix F and responses to the 3 misfitting items were simulated separately. Similarly, in conditions in which 33% of items exhibited such misfit, the last 10 sets of "true betas" in Appendix C were replaced with the 10 sets of "true betas" in Appendix F and responses to the 10 misfitting items were simulated separately. Item responses were simulated for the misfitting items using the *rmvordlogis* function's GPCM option, rather than GRM.

Data Analysis

Several steps of data analysis were undertaken to answer the two research questions. First, descriptive statistics for the simulated data were obtained, characterizing the distribution of mean χ^2/df ratios across all single items, pairs of items, and triples of items. In addition, proportions of χ^2/df ratios > 3, suggesting item misfit, were computed across the entire set of data. Next, inferential analyses were conducted to test hypotheses associated with each Research Question, using graphical procedures to assist with interpretation of results. A Bonferroni-corrected level of significance of .003 was used in testing each hypothesis, to maintain a study-wide alpha of .05 ($\alpha = .05/17 = .003$). All analyses were conducted using R: A Language and Environment for Statistical Computing (R Development Core Team, 2012).

Research Question 1: Sample Size and Adjustment Condition

To answer Research Question 1, the fixed effects of two factors (sample size and adjustment condition) were tested by fitting generalized least squares linear models. To account for the repeated measures of adjustment condition on each simulated dataset, the covariance matrix for the residuals was specified as block diagonal with compound symmetric structure within subjects (where each simulated dataset is a subject). Three models were fit using maximum likelihood estimation, assessing the mean χ^2/df ratios for single items, pairs of items, and triplets of items. A Box-Cox transformation ($\lambda = 0$) was applied before fitting each model, to alleviate heteroskedasticity noted in residual plots when the untransformed response variable was used. Thus, the response variable in each case, y'_{ijk} , is the natural logarithm of the mean χ^2/df ratio for all single items, pairs of items, and triplets of items, respectively, for a particular simulated dataset. The full model in this case is:

$$y'_{ijk} = \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk} ,$$
$$\boldsymbol{\varepsilon}_{ij} = \begin{bmatrix} \varepsilon_{ij1} \\ \varepsilon_{ij2} \end{bmatrix} \sim N(0, \boldsymbol{\Lambda}_{ij}), \quad \boldsymbol{\Lambda}_{ij} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} , \tag{7}$$

where $i = \{1, \dots, 1000\}$ indexes the simulation number, $j = \{1,2,3\}$ indexes the sample size condition, and $k = \{1,2\}$ indexes the adjustment condition.

The first hypothesis tested addressed the potential interaction between sample size (factor A) and adjustment condition (factor B), where adjustment condition was a repeated measure (i.e., adjusted and unadjusted χ^2/df ratios). Thus, the primary effect of interest was the AB interaction effect, $(\alpha\beta)_{jk}$, where α_j denotes the factor A main effect and β_k denotes the factor B main effect. If no interaction effect was detected, the analysis plan included testing for two additional hypotheses regarding the main effects of factor A and factor B. The following formal hypotheses were posed:

Hypothesis 1.1: A significant interaction effect on the natural logarithm of mean χ^2/df ratios associated with single items, pairs of items, and triplets of items is expected between sample size and adjustment condition, when no item misfit is present.

$H_{0.1}$: all $(\alpha\beta)_{jk} = 0$

$H_{1.1}$: not all $(\alpha\beta)_{jk} = 0$

Hypothesis 1.2: A significant main effect of sample size on the natural logarithm of mean χ^2/df ratios associated with single items, pairs of items, and triplets of items is expected, when no item misfit is present.

$H_{0.2}$: all $\alpha_j = 0$

$H_{1.2}$: not all $\alpha_j = 0$

Hypothesis 1.3: A significant main effect of adjustment condition on the natural logarithm of mean χ^2/df ratios associated with single items, pairs of items, and triplets of items is expected, when no item misfit is present.

$H_{0.3}$: all $\beta_k = 0$

$H_{1.3}$: not all $\beta_k = 0$

The hypotheses were tested sequentially, so that if the *AB* interaction effect (Hypothesis 1.1) was significant, Hypotheses 1.2 and 1.3 were not tested.

To further inform interpretation of results, the proportion of χ^2/df ratios > 3 observed in each design cell was plotted using box plots. This approach allowed visualization of the frequency distributions of unadjusted versus adjusted ratios at each level of sample size, when no misfitting items were actually present in the simulated data.

Research Question 2: Sample Size, Type of Misfit, and Proportion of Misfitting Items

The outcomes of interest in Research Question 2 included (a) the mean χ^2/df ratios for single items, pairs of items, and triplets of items; (b) the proportion of χ^2/df ratios > 3 for single items, pairs of items, and triplets of items; and (c) the sensitivity and specificity of the "rule of thumb" for item misfit (i.e., χ^2/df ratios > 3) when applied to single items. (Sensitivity and specificity for pairs and triplets of items were not explored, because pairs and triplets could contain combinations of fitting and misfitting items simultaneously.)

The data used to answer Research Question 2 comprised the unadjusted χ^2/df ratios for the small and medium sample size conditions (n = 400 and 1500, respectively), and the adjusted χ^2/df ratios for the large sample size condition (n = 10000). The adjustment conditions for each sample size were selected based upon the results of Research Question 1. In addition, because the data simulation method allowed the truly misfitting items to be known in each simulated dataset, the sensitivity (percentage of misfitting items correctly identified) and specificity (percentage of fitting items correctly identified) could be calculated. In each simulated dataset, sensitivity of the χ^2/df ratios was obtained by dividing the number of correctly identified misfitting items by the total number of misfitting items in that dataset (i.e., either 3 or 10, depending on the condition). Similarly, specificity of the χ^2/df ratios was obtained by dividing the number of correctly identified fitting items by the total number of fitting items in that dataset (i.e., either 27 or 20, depending on the condition).

First, the fixed effects of three factors (sample size, type of misfit, and amount of misfit) were tested in generalized least squares linear models. To account for observed heteroskedasticity, weighted least squares were used, in which non-diagonal elements of the covariance matrix were zero and variances could differ by each combination of the three factors. Three models were fit, assessing the mean χ^2/df ratios for single items, pairs of items, and triplets of items. A Box-Cox transformation (λ = -1) was applied before fitting each model. Thus, the response variable in each case, y'_{ijk} , is the inverse of the mean χ^2/df ratio for all single items, pairs of items, and triplets of items, respectively, for a particular simulated dataset. The full model in this case is:

$$y'_{ijkl} = \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl}, \quad \varepsilon_{ijkl} \sim N(0, \sigma^2_{ijk}), \tag{8}$$

where $i = \{1, 2,3\}$ indexes the sample size condition, $j = \{1,2,3\}$ indexes the type of misfit, $k = \{1,2\}$ indexes the amount of misfit, and $l = \{1, \ldots, 1000\}$ indexes the simulation number.

The first hypothesis tested addressed the potential three-way interaction between sample size (factor A), type of misfit (factor B), and amount of misfit (factor C) on mean χ^2/df ratios. Thus, the primary effect of interest was the ABC interaction effect, $(\alpha\beta\gamma)_{ijk}$, where α_i denotes the factor A main effect, and β_j denotes the factor B main effect, and γ_k denotes the factor C main effect. If no three-way interaction effect was detected, the analysis plan included testing for each two-way interaction effect, and then similarly for each main effect. The following formal hypotheses were posed:

Hypothesis 2.1: A significant interaction effect on the inverse of mean χ^2/df ratios associated with single items, pairs of items, and triplets of items is expected between sample size, type of misfit, and amount of misfit.

$H_{0.1}$: all $(\alpha\beta\gamma)_{ijk} = 0$

$H_{1.1}$: not all $(\alpha\beta\gamma)_{ijk} = 0$

Hypothesis 2.2: A significant interaction effect on the inverse of mean χ^2/df ratios associated with single items, pairs of items, and triplets of items is expected between sample size and type of misfit.

$H_{0.2}$: all $(\alpha\beta)_{ij} = 0$

$H_{1.2}$: not all $(\alpha\beta)_{ij} = 0$

Hypothesis 2.3: A significant interaction effect on the inverse of mean χ^2/df ratios associated with single items, pairs of items, and triplets of items is expected between sample size and amount of misfit.

$H_{0.3}$: all $(\alpha\gamma)_{ik} = 0$

$H_{1.3}$: not all $(\alpha\gamma)_{ik} = 0$

Hypothesis 2.4: A significant interaction effect on the inverse of mean χ^2/df ratios associated with single items, pairs of items, and triplets of items is expected between type of misfit and amount of misfit.

$H_{0.4}$: all $(\beta\gamma)_{jk} = 0$

$H_{1.4}$: not all $(\beta\gamma)_{jk} = 0$

Hypothesis 2.5: A significant main effect of sample size on the inverse of mean χ^2/df ratios associated with single items, pairs of items, and triplets of items is expected.

$H_{0.5}$: all $\alpha_i = 0$

$H_{1.5}$: not all $\alpha_i = 0$

Hypothesis 2.6: A significant main effect of type of misfit on the inverse of mean χ^2/df ratios associated with single items, pairs of items, and triplets of items is expected.

$H_{0.6}$: all $\beta_j = 0$

$H_{1.6}$: not all $\beta_j = 0$

Hypothesis 2.7: A significant main effect of amount of misfit on inverse of mean χ^2/df ratios associated with single items, pairs of items, and triplets of items is expected.

$H_{0.7}$: all $\gamma_k = 0$

$H_{1.7}$: not all $\gamma_k = 0$

The hypotheses were tested sequentially, so that if the *ABC* interaction effect (Hypothesis 2.1) was significant, the two-way interactions and main effects of factors *A*, *B*, and *C* were not tested.

Next, the proportion of χ^2/df ratios > 3 observed in each design cell was plotted using box plots. This approach allowed visualization of the frequency distributions of ratios > 3 across conditions of sample size, type of misfit, and amount of misfit.

In addition, the sensitivity and specificity of using the χ^2/df ratios > 3 "rule of thumb" for identifying misfit in single items were investigated by fitting multiple logistic regression models. The main and interaction effects of sample size, type of misfit, and amount of misfit were tested. The full model in this case is:

$$\text{logit}[P(Y = 1)] = \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} \qquad (9)$$

where logit[(P(Y = 1)] is the log odds of the probability that an item is "correctly" identified (i.e., as misfitting, for sensitivity, or as fitting, for specificity), $i = \{1, 2,3\}$ indexes the sample size condition, $j = \{1,2,3\}$ indexes the type of misfit, and $k = \{1,2\}$ indexes the amount of misfit. Parallel hypotheses to Hypothesis 2.1-2.7 were posed:

Hypothesis 3.1: A significant interaction effect on the sensitivity and specificity of using χ^2/df ratios > 3 to identify single items with misfit is expected between sample size, type of misfit, and amount of misfit.

$H_{0.1}$: all $(\alpha\beta\gamma)_{ijk} = 0$

$H_{1.1}$: not all $(\alpha\beta\gamma)_{ijk} = 0$

Hypothesis 3.2: A significant interaction effect on the sensitivity and specificity of using χ^2/df ratios > 3 to identify single items with misfit is expected between sample size and type of misfit.

$H_{0.2}$: all $(\alpha\beta)_{ij} = 0$

$H_{1.2}$: not all $(\alpha\beta)_{ij} = 0$

Hypothesis 3.3: A significant interaction effect on the sensitivity and specificity of using χ^2/df ratios > 3 to identify single items with misfit is expected between sample size and amount of misfit.

$H_{0.3}$: all $(\alpha\gamma)_{ik} = 0$

$H_{1.3}$: not all $(\alpha\gamma)_{ik} = 0$

Hypothesis 3.4: A significant interaction effect on the sensitivity and specificity of using χ^2/df ratios > 3 to identify single items with misfit is expected between type of misfit and amount of misfit.

$H_{0.4}$: all $(\beta\gamma)_{jk} = 0$

$H_{1.4}$: not all $(\beta\gamma)_{jk} = 0$

Hypothesis 3.5: A significant main effect of sample size on the sensitivity and specificity of using χ^2/df ratios > 3 to identify single items with misfit is expected.

$H_{0.5}$: all $\alpha_i = 0$

$H_{1.5}$: not all $\alpha_i = 0$

Hypothesis 3.6: A significant main effect of type of misfit on the sensitivity and specificity of using χ^2/df ratios > 3 to identify single items with misfit is expected.

$H_{0.6}$: all $\beta_j = 0$

$H_{1.6}$: not all $\beta_j = 0$

Hypothesis 3.7: A significant main effect of amount of misfit on the sensitivity and specificity of using χ^2/df ratios > 3 to identify single items with misfit is expected.

$H_{0.7}$: all $\gamma_k = 0$

$H_{1.7}$: not all $\gamma_k = 0$

As in previous analyses, the hypotheses were tested sequentially, so that if the *ABC* interaction effect (Hypothesis 3.1) was significant, the two-way interactions and main effects of factors *A*, *B*, and *C* were not tested.

Finally, boxplots were used to visualize the distributions of sensitivity and specificity of using χ^2/df ratios > 3 to identify single items with misfit.

CHAPTER IV

RESULTS

Research Question 1: Sample Size and Adjustment Condition

Data Characteristics

The mean, standard deviation, minimum, and maximum of the χ^2/df ratios averaged across single items, pairs of items, and triplets of items for each level of sample size and adjustment condition are presented in Table 1. The simulated data included 1000 replications in each condition, with mean unadjusted and adjusted ratios calculated on the same datasets, for a total N=3000. The highest mean χ^2/df ratios were observed in the adjusted condition. Mean values peaked for adjusted χ^2/df ratios for single items in the smallest sample size condition [Mean (M) = 37.77, standard deviation (SD) = 12.55]. Values decreased for larger sample size conditions and for χ^2/df ratios averaged across pairs and triplets of items. In the unadjusted condition, the highest mean χ^2/df ratio was observed for single items in the largest sample size condition (M = 4.05, SD = 1.16), while mean unadjusted ratios were more similar across sample sizes when calculated for pairs and triplets of items.

The distributions of proportions of χ^2/df ratios exceeding the "rule of thumb" cut point of 3 for each level of sample size and adjustment condition are presented in Table 2. These values represent the percentage of single items, pairs of items, and triplets of items identified with misfit, although no misfitting items were simulated. In the adjusted

Table 1

Descriptive Statistics: Distribution of Mean χ^2/df Ratios by Adjustment and Sample Size Conditions, for Simulated Data with No Misfitting Items

	Single Items				Pairs of Items				Triplets of Items			
Condition	M	(SD)	Min	Max	M	(SD)	Min	Max	M	(SD)	Min	Max
Adjusted χ^2/df ratios												
N = 400	37.77	(12.55)	19.46	131.00	23.97	(3.97)	17.11	49.62	18.80	(1.61)	15.75	29.32
N = 1500	10.53	(3.63)	5.34	36.12	6.35	(1.09)	4.56	13.54	4.95	(0.43)	4.20	7.84
N = 10000	2.43	(0.69)	0.88	6.30	1.31	(0.23)	0.75	2.27	0.89	(0.09)	0.66	1.24
Unadjusted χ^2/df ratios												
N = 400	2.46	(0.84)	1.21	8.51	1.70	(0.35)	1.09	4.12	1.42	(0.24)	0.97	2.85
N = 1500	2.63	(0.91)	1.33	9.03	1.65	(0.31)	1.15	3.73	1.36	(0.17)	1.05	2.43
N = 10000	4.05	(1.16)	1.47	10.49	2.17	(0.37)	1.25	3.74	1.43	(0.13)	1.09	1.92

Table 2

Descriptive Statistics: Distribution of Proportion[a] of χ^2/df Ratios > 3 by Adjustment and Sample Size Conditions, for Simulated Data with No Misfitting Items

	Single Items				Pairs of Items				Triplets of Items			
Condition	M	(SD)	Min	Max	M	(SD)	Min	Max	M	(SD)	Min	Max
Adjusted χ^2/df ratios												
N = 400	98.79	(2.06)	90.00	100.00	100.00	(0.00)	100.00	100.00	100.00	(0.00)	100.00	100.00
N = 1500	87.86	(8.24)	60.00	100.00	97.29	(2.45)	86.21	100.00	99.52	(0.48)	97.04	100.00
N = 10000	29.37	(16.57)	0.00	90.00	0.17	(0.64)	0.00	6.44	0.00	(0.00)	0.00	0.00
Unadjusted χ^2/df ratios												
N = 400	28.61	(16.10)	3.33	100.00	39.94	(16.67)	8.97	98.39	22.18	(11.76)	3.60	85.12
N = 1500	32.83	(17.80)	3.33	100.00	3.38	(6.46)	0.00	68.51	0.11	(0.72)	0.00	17.54
N = 10000	58.71	(17.02)	6.67	100.00	12.41	(13.15)	0.00	85.29	0.00	(0.01)	0.00	0.25

[a]Shown as *100%

condition, the mean percentage of single, pairs, and triplets of items identified with misfit were uniformly high (means ranging from 86% to 100%) across the small and medium sample size conditions. In the large sample size condition, a non-negligible proportion of single items (M = 29%, SD = 17%) was identified as misfitting, while drastically fewer pairs and triplets exceeded the "rule of thumb" cut point. In the unadjusted condition, the smallest sample size condition still yielded substantial proportions of single, pairs, and triplets of items identified as misfitting (means ranging from 22% to 40%); the medium sample size condition had a high proportion of single items (M = 33%, SD = 18%) but lower proportions of pairs and triplets identified as misfitting; and the largest sample size had a large proportion of single items (M = 59%, SD = 17%), fewer pairs of items (M = 12%, SD = 13%), and virtually no triplets of items flagged for misfit.

Hypothesis Testing

Hypothesis 1.1 addressed the potential interaction between sample size and adjustment condition, where adjustment condition was a repeated measure (i.e., mean adjusted and unadjusted χ^2/df ratios). This hypothesis was tested by fitting generalized least squares linear models with block diagonal compound symmetrical residual covariance structure within subjects. Three separate models were fit for three response variables: the natural logarithm of the mean χ^2/df ratios for (a) single items, (b) pairs of items, and (c) triplets of items. For each model, residual plots were examined and were deemed appropriate. See Table 3 for detailed results. In each model, the interaction between sample size and adjustment condition was statistically significant ($p < .001$). Two reduced nested models (one including both main effects only and one including only the effect of adjustment condition) for each response variable were also fit and compared

Table 3

Linear Models: Effects of Sample Size and Adjustment Condition on Mean χ^2/df Ratios[a]

	Single Items				Pairs of Items				Triplets of Items			
Effect	*Coef.*	*se*	*t*	*p*	*Coef.*	*se*	*t*	*p*	*Coef.*	*se*	*t*	*p*
(Intercept)	0.85	0.01	94.61	<.001	0.26	0.00	48.86	<.001	-0.12	0.00	-35.25	<.001
Sample (n=1500)	1.45	0.01	114.43	<.001	1.58	0.01	210.77	<.001	1.71	0.00	363.31	<.001
Sample (n=400)	2.74	0.01	215.24	<.001	2.91	0.01	388.62	<.001	3.05	0.00	646.25	<.001
Adjustment (Unadjusted)	0.51	0.00	433.57	<.001	0.50	0.00	371.28	<.001	0.47	0.00	231.97	<.001
Sample (n=1500):Adjustment (Unadjusted)	-1.90	0.00	-1139.71	<.001	-1.85	0.00	-969.44	<.001	-1.77	0.00	-616.63	<.001
Sample (n=400):Adjustment (Unadjusted)	-3.24	0.00	-1946.74	<.001	-3.15	0.00	-1650.47	<.001	-3.06	0.00	-1029.42	<.001

[a]Response variables are the natural logarithms of the mean χ^2/df ratios.

Table 4

Comparing Nested Models: Full Model versus Reduced Models

		Single Items			Pairs of Items			Triplets of Items		
Model	*df*	AIC	BIC	LL	AIC	BIC	LL	AIC	BIC	LL
Sample:Adjustment	8	-10260	-10206	5138*	-12657	-12603	6336*	-12828	-12775	6422*
Sample + Adjustment	6	11195	11235	-5591*	7810	7850	-3899*	4736	4776	-2362*
Adjustment	4	15021	15048	-7507	15187	15214	-7590	15380	15406	-7686

Note. df = degrees of freedom. AIC = Akaike information criterion. BIC = Bayesian information criterion. LL = log likelihood.
*Model fits significantly better than next simpler model per likelihood ratio test, $p < .001$.

to the full model including the interaction term; AIC values and results of likelihood ratio tests indicated that in each case, the full model demonstrated significantly better model fit. See Table 4 for model fit results. Interaction plots are presented in Figure 4, 5, and 6, for mean ratios associated with single items, pairs of items, and triplets of items, respectively. Given the true structure of the data (i.e., no misfitting items), it appears clear from these results that the adjusted condition results in very inflated χ^2/df ratio values for small and medium sample sizes; unadjusted ratios, while still higher than warranted, are much more reflective of the true data structure in these sample size conditions. For example, the mean adjusted χ^2/df ratio values for single items in the small and medium sample size conditions were 37.77 and 10.53, respectively, compared to mean unadjusted ratios of 2.46 and 2.63, respectively. Conversely, in the large sample size condition, the mean adjusted χ^2/df ratios were closer to their expected value of 1 than the unadjusted ratios, as desired. For example, for single items, the mean adjusted χ^2/df ratio value in the large sample size condition was 2.43, while the unadjusted value was 4.05. Due to the detection of the significant interaction effect hypothesized in Hypothesis 1.1, Hypotheses 1.2 and 1.3 (regarding main effects) were not tested.

Graphical Analyses

The remaining issue addressed by Research Question 1 was investigated using graphical techniques. Specifically, box plots were used to allow the visualization of differences in the distributions of the proportions of χ^2/df ratios exceeding the cut point of 3, indicating misfit, for single items, pairs of items, and triplets of items. These results are presented in Figures 7, 8, and 9, respectively. As is evident from Figures 7-9, under the small and medium sample size condition, very high proportions of adjusted mean χ^2/df

ratios for single items, pairs of items, and triplets of items exceed 3, indicating misfit. However, the simulated data included no misfitting items. The unadjusted ratios are notably lower in the small and medium sample size conditions. Conversely, under the large sample size condition, the proportions of unadjusted mean χ^2/df ratios exceeding 3 are consistently higher than the proportions of adjusted ratios.

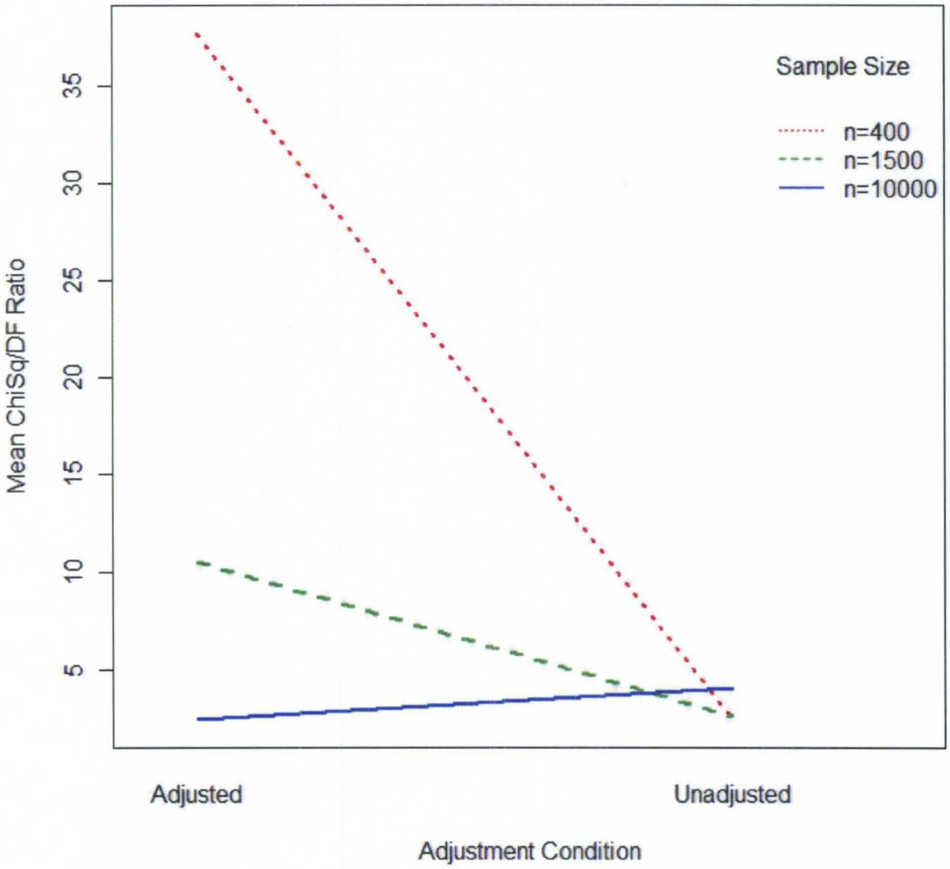

Figure 4. **Interaction of sample size by adjustment condition of mean χ^2/df ratios for single items, for simulated data with no misfitting items.**

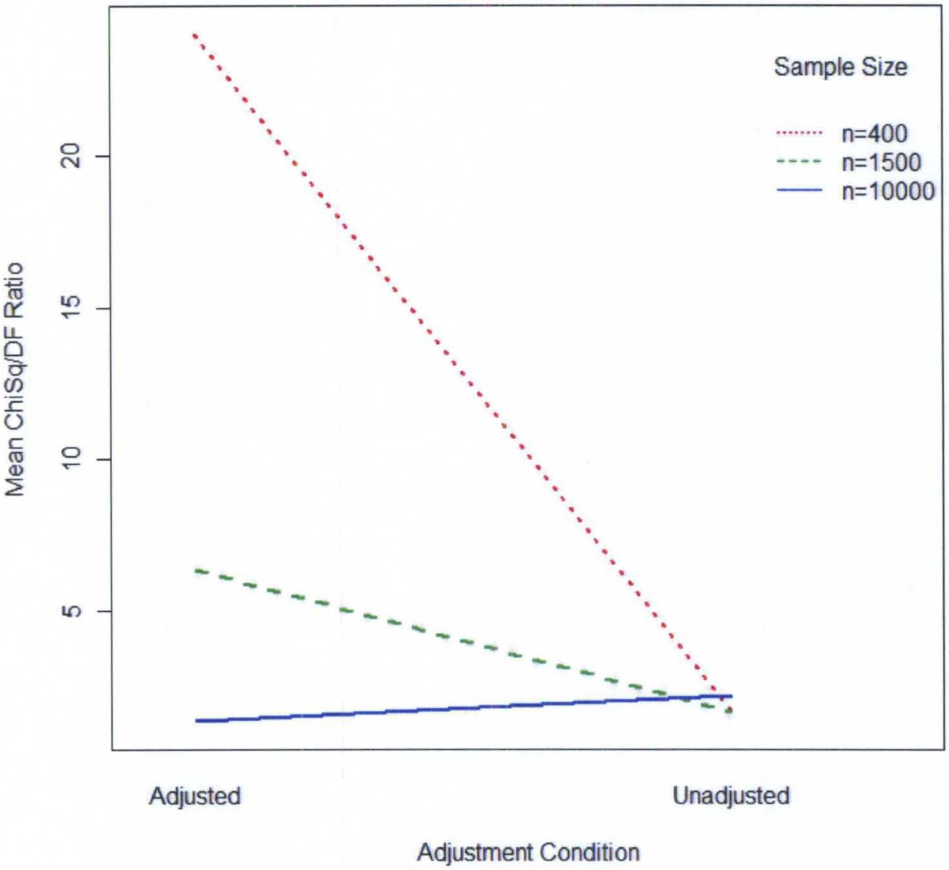

Figure 5. **Interaction of sample size by adjustment condition of mean χ^2/df ratios for pairs of items, for simulated data with no misfitting items.**

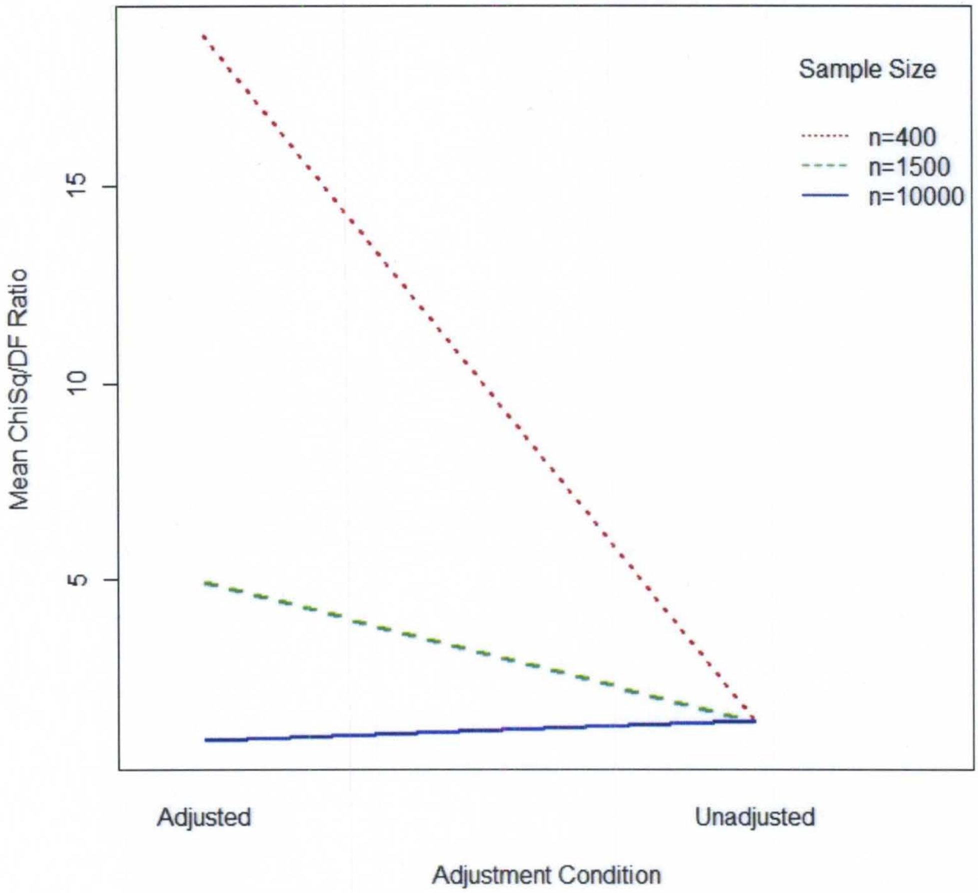

Figure 6. **Interaction of sample size by adjustment condition of mean χ^2/df ratios for triplets of items, for simulated data with no misfitting items.**

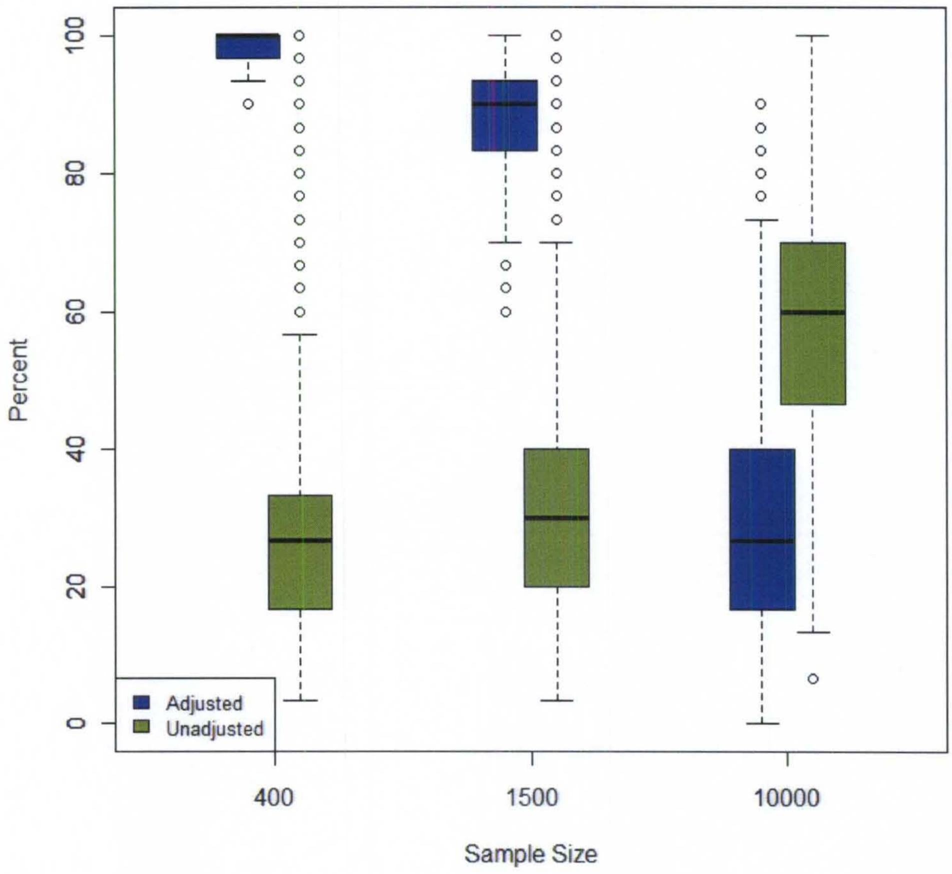

Figure 7. **Box plot of percentage of single items' χ^2/df ratios > 3 by sample size and adjustment condition, for simulated data with no misfitting items.**

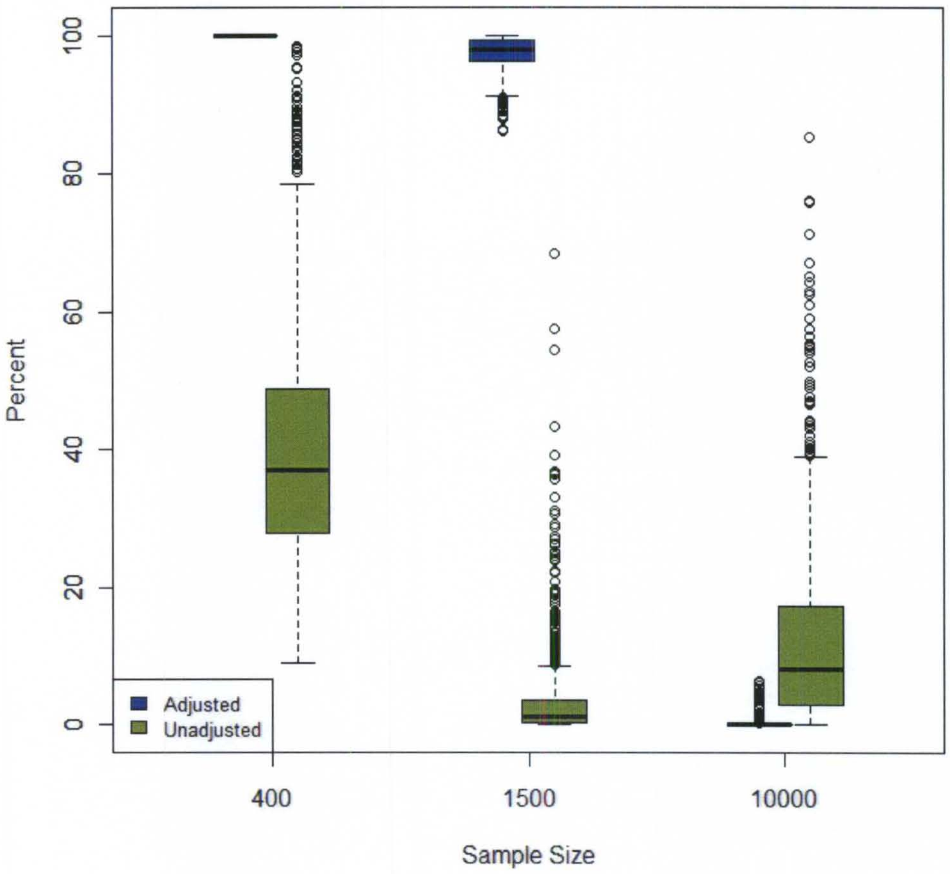

Figure 8. **Box plot of percentage of item pairs' χ^2/df ratios > 3 by sample size and adjustment condition, for simulated data with no misfitting items.**

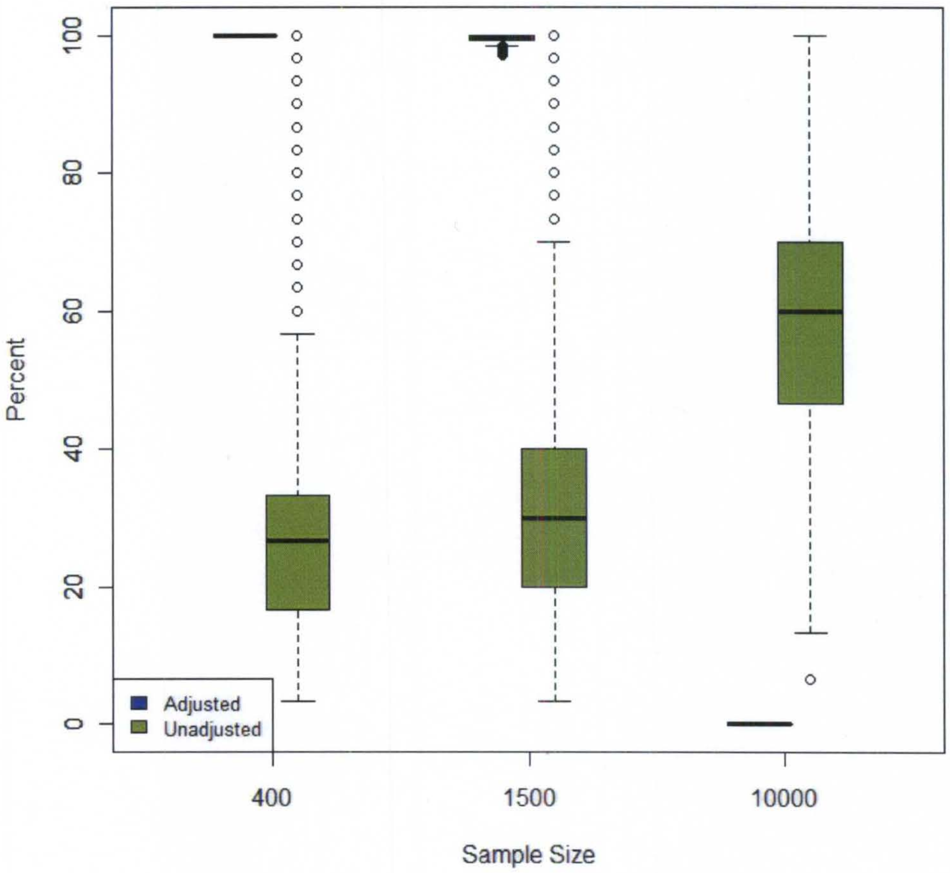

Figure 9. **Box plot of percentage of item triplets' χ^2/df ratios > 3 by sample size and adjustment condition, for simulated data with no misfitting items.**

Research Question 2: Sample Size, Type of Misfit, and Proportion of Misfitting Items

Data Characteristics

Mean χ^2/df ratios. For each level of sample size, type of misfit, and proportion of misfitting items, the means and standard deviations of χ^2/df ratios averaged across single items, pairs of items, and triplets of items are presented in Tables 5, 6, and 7, respectively. The distributions are illustrated in Figure 10. Recall that adjusted χ^2/df ratios were used for the largest sample size condition, while unadjusted ratios were used for the small and medium sample size conditions, based upon the results of Research Question 1. The simulated data included 1000 replications in each condition, for total N=18000.

Patterns observed in the magnitude and variation of χ^2/df ratios appeared to differ among those averaged across single items versus pairs and triplets of items. As reported in Table 5, for single items, the magnitude of mean ratios appeared more consistent within the 10% misfitting conditions (lowest M = 1.97, SD = 0.53; highest M = 2.69, SD = 3.65) than within the 33% misfitting conditions (lowest M = 1.33, SD = 0.28; highest M = 4.15, SD = 4.21). Across the 10% and 33% misfitting conditions, the highest mean χ^2/df ratios and largest standard deviations were observed in the conditions in which misfit was due to multidimensionality (e.g., for N = 10000 with 33% misfitting items due to multidimensionality, M = 3.81, SD = 16.77). Conversely, the lowest mean χ^2/df ratios across the 10% and 33% misfitting conditions were both observed in the largest sample size condition where misfit was due to generation from a competing model (the GPCM; for 10% misfitting, M = 1.97, SD = 0.53; for 33% misfitting, M = 1.33, SD = 0.28).

For pairs and triplets of items, several trends were noted. Across the 10% misfitting conditions, mean χ^2/df ratios appeared to decrease as sample size increased,

Table 5

Descriptive Statistics: Distribution of Mean χ^2/df Ratios[a] for Single Items by Sample Size, Type of Misfit, and Proportion of Misfitting Items

	Multidimensionality		DIF		Competing Model (GPCM)	
Condition	M	(SD)	M	(SD)	M	(SD)
10% items misfitting						
N = 400	2.62	(1.32)	2.43	(0.85)	2.33	(0.66)
N = 1500	2.46	(0.73)	2.62	(0.89)	2.46	(0.73)
N = 10000	2.69	(3.65)	2.41	(0.71)	1.97	(0.53)
33% items misfitting						
N = 400	4.15	(4.21)	2.40	(0.77)	2.16	(0.46)
N = 1500	2.13	(0.44)	2.57	(0.93)	2.14	(0.44)
N = 10000	3.81	(16.77)	2.29	(0.67)	1.33	(0.28)

[a]For N = 10000, χ^2/df ratios with N adjusted to 3000 are used; unadjusted χ^2/df ratios are used for the two smaller sample size conditions.

Table 6

Descriptive Statistics: Distribution of Mean χ^2/df Ratios[a] for Pairs of Items by Sample Size, Type of Misfit, and Proportion of Misfitting Items

	Multidimensionality		DIF		Competing Model (GPCM)	
Condition	M	(SD)	M	(SD)	M	(SD)
10% items misfitting						
N = 400	1.82	(0.75)	1.68	(0.36)	1.65	(0.29)
N = 1500	1.60	(0.25)	1.63	(0.29)	1.62	(0.25)
N = 10000	1.50	(1.31)	1.31	(0.24)	1.30	(0.17)
33% items misfitting						
N = 400	2.67	(2.22)	1.63	(0.31)	1.77	(0.21)
N = 1500	1.80	(0.16)	1.59	(0.28)	2.28	(0.19)
N = 10000	3.08	(5.13)	1.28	(0.22)	4.04	(0.18)

[a] For N = 10000, χ^2/df ratios with N adjusted to 3000 are used; unadjusted χ^2/df ratios are used for the two smaller sample size conditions.

Table 7

Descriptive Statistics: Distribution of Mean χ^2/df Ratios[a] for Triplets of Items by Sample Size, Type of Misfit, and Proportion of Misfitting Items

	Multidimensionality		DIF		Competing Model (GPCM)	
Condition	M	(SD)	M	(SD)	M	(SD)
10% items misfitting						
N = 400	1.55	(0.67)	1.42	(0.27)	1.38	(0.21)
N = 1500	1.34	(0.14)	1.35	(0.31)	1.34	(0.36)
N = 10000	1.00	(0.55)	0.88	(0.09)	0.93	(0.07)
33% items misfitting						
N = 400	2.34	(1.94)	1.42	(0.27)	1.49	(0.16)
N = 1500	1.42	(0.09)	1.32	(0.16)	1.84	(0.10)
N = 10000	1.91	(1.96)	0.85	(0.08)	2.87	(0.09)

Note. 1248 cases are missing for ratios calculated for triplets of items.

[a] For N = 10000, χ^2/df ratios with N adjusted to 3000 are used; unadjusted χ^2/df ratios are used for the two smaller sample size conditions.

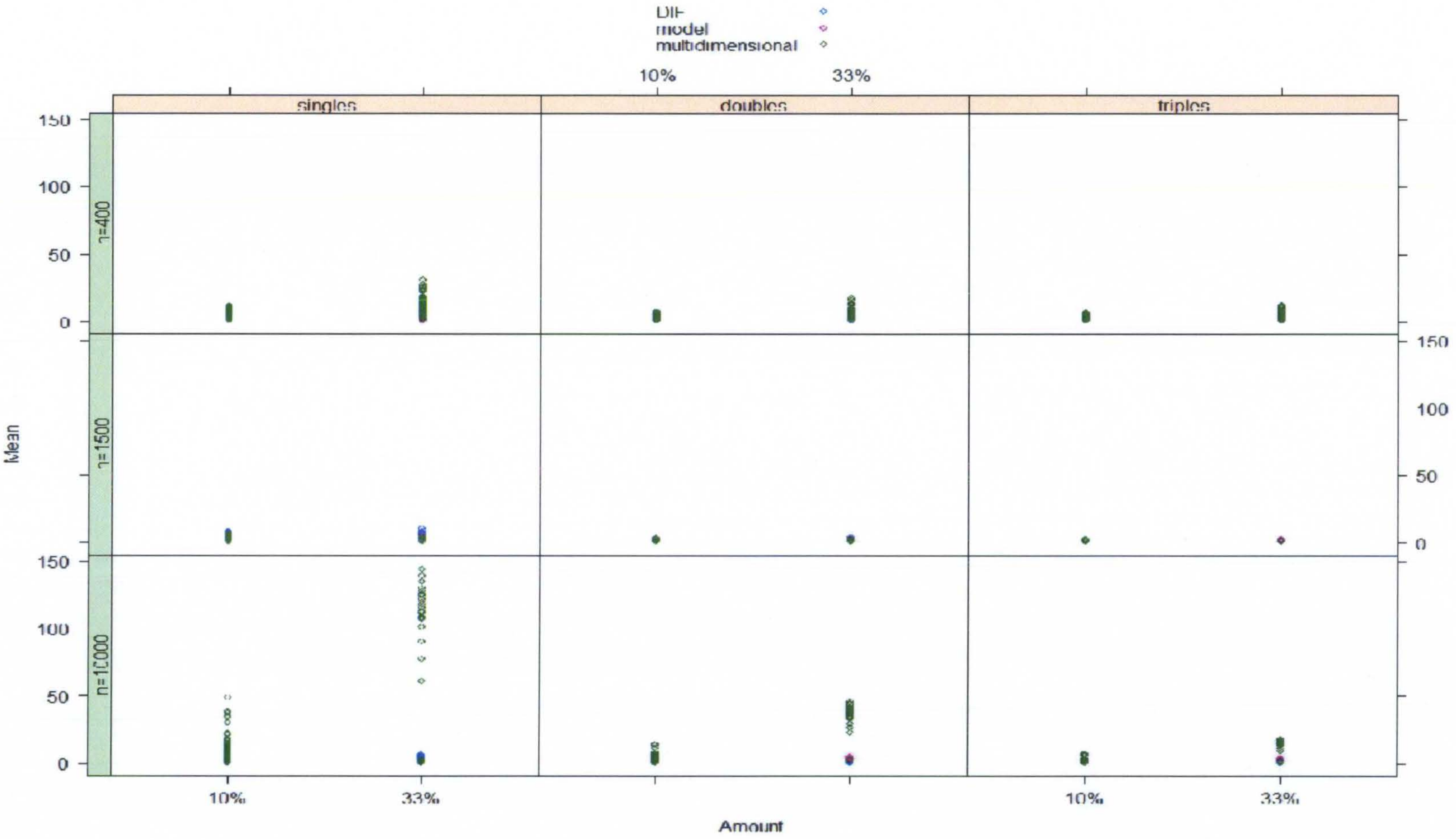

Figure 10. **Plot of distributions of mean χ^2/df ratios for single items, pairs of items, and triplets of items by sample size, type of misfit, and proportion of misfitting items.**

regardless of type of misfit. This pattern was not observed in the 33% misfitting conditions, however. For example, in the 10% misfitting conditions for triplets of items, the lowest mean χ^2/df ratios were observed for the largest sample size, ranging from M = 0.88 (SD = 0.09) to M = 1.00 (SD = 0.55). In contrast, in the 33% misfitting condition for triplets of items, some of the highest χ^2/df ratios were observed for the largest sample size (e.g., when misfit was due to multidimensionality, M = 1.91, SD = 1.96; when misfit was due to generation from a competing model, M = 2.87, SD = 0.09). Similar to the narrower range of mean ratios observed across the 10% misfitting conditions for single items, the mean ratios for pairs and triplets of items were more similar in magnitude within the 10% misfitting conditions (lowest pairs M = 1.30, SD = 0.17; highest pairs M = 1.82, SD = 0.75; lowest triplets M = 0.88, SD = 0.09; highest triplets M = 1.55, SD = 0.67) than within the 33% misfitting conditions (lowest pairs M = 1.28, SD = 0.22; highest pairs M = 4.04, SD = 0.18; lowest triplets M = 0.85, SD = 0.08; highest triplets M = 2.87, SD = 0.09).

Proportions of χ^2/df ratios > 3. The distributions of proportions of χ^2/df ratios exceeding the "rule of thumb" cut point of 3 for each level of sample size, type of misfit, and proportion of misfitting items are presented in Tables 8, 9, and 10 for single items, pairs of items, and triplets of items, respectively. Boxplots illustrate these distributions in Figure 11. These values represent the percentages of all single items, pairs of items, and triplets of items identified with misfit. In general, the proportion of ratios > 3 across all study conditions was largest among single items, smaller among pairs, and still smaller among triplets. Proportions of single item χ^2/df ratios > 3 across study conditions primarily ranged from 18.88% to 36.50%, except for two notably low values: the largest

Table 8

Descriptive Statistics: Distribution of Proportion[a] of Single Items with χ^2/df Ratios > 3 by Sample Size, Type of Misfit, and Proportion of Misfitting Items

	Multidimensionality		DIF		Competing Model (GPCM)	
Condition	M	(SD)	M	(SD)	M	(SD)
10% items misfitting						
N = 400	29.64	(18.52)	27.98	(16.52)	26.32	(13.58)
N = 1500	29.42	(15.15)	32.79	(17.85)	29.29	(14.96)
N = 10000	22.81	(20.45)	28.84	(16.45)	18.88	(12.91)
33% items misfitting						
N = 400	36.50	(27.47)	27.89	(16.35)	22.74	(10.52)
N = 1500	22.54	(10.16)	31.91	(18.12)	22.44	(10.57)
N = 10000	8.42	(14.60)	26.28	(15.94)	6.19	(5.68)

[a]Shown as *100%

Table 9

Descriptive Statistics: Distribution of Proportion[a] of Pairs of Items with χ^2/df Ratios > 3 by Sample Size, Type of Misfit, and Proportion of Misfitting Items

	Multidimensionality		DIF		Competing Model (GPCM)	
Condition	M	(SD)	M	(SD)	M	(SD)
10% items misfitting						
N = 400	9.65	(16.86)	6.06	(8.34)	5.71	(6.33)
N = 1500	2.94	(4.63)	2.96	(5.78)	2.98	(4.49)
N = 10000	5.01	(17.71)	0.24	(1.16)	0.71	(0.20)
33% items misfitting						
N = 400	22.99	(34.12)	4.72	(6.62)	10.14	(4.20)
N = 1500	9.77	(1.71)	2.16	(5.48)	11.16	(1.52)
N = 10000	12.32	(13.16)	0.25	(0.90)	10.34	(0.01)

[a] Shown as *100%

Table 10

Descriptive Statistics: Distribution of Proportion[a] of Triplets of Items with χ^2/df Ratios > 3 by Sample Size, Type of Misfit, and Proportion of Misfitting Items

	Multidimensionality		DIF		Competing Model (GPCM)	
Condition	M	(SD)	M	(SD)	M	(SD)
10% items misfitting						
N = 400	6.44	(15.45)	3.43	(4.64)	2.68	(2.96)
N = 1500	0.18	(0.05)	0.10	(0.55)	0.50	(0.55)
N = 10000	3.21	(8.83)	0.00	(0.00)	2.02	(0.00)
33% items misfitting						
N = 400	19.74	(32.18)	4.28	(4.73)	5.54	(2.58)
N = 1500	1.97	(0.69)	0.05	(0.49)	14.97	(2.21)
N = 10000	18.64	(12.25)	0.00	(0.00)	25.12	(0.00)

[a]Shown as *100%

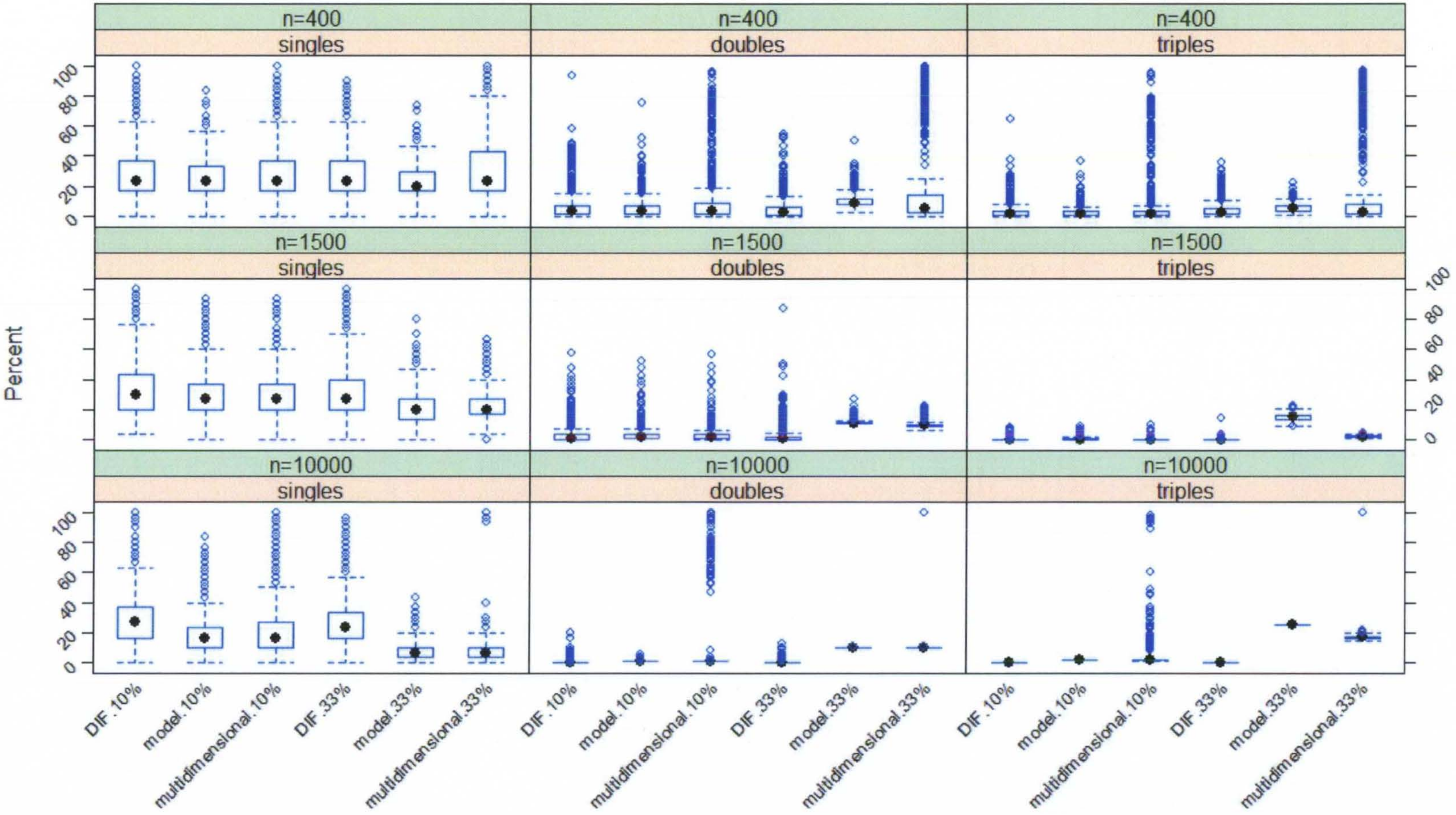

Figure 11. **Box plot of percentage of single items', pairs of items', and triplets of items' χ^2/df ratios > 3 by sample size, type of misfit, and proportion of misfitting items.**

sample size condition with 33% misfitting items due to multidimensionality (M = 8.42%, SD = 14.60%) and to generation from a competing model (M = 6.19%, SD = 5.68%). Among ratios for pairs of items, notably high proportions were observed in the conditions with 33% misfitting items due to multidimensionality for the smallest (M = 22.99%, SD = 34.12%) and largest (M = 12.32%, SD = 13.16%) sample sizes. Ratios for triplets of items were similarly high for those conditions (M = 19.74%, SD = 32.18%, and M = 18.64% and SD = 12.25%, respectively), as well as for the 33% misfitting items due to generation from a competing model with the largest sample size (M = 25.12%, SD = 0.00%). Other than these highlighted values, most proportions for pairs and triplets of items fell between 0% and 10% across conditions, though proportions generally appeared higher within the 33% misfitting conditions.

Sensitivity and specificity. Using the cut point of 3 to indicate item misfit, the mean sensitivity and specificity of the mean χ^2/df ratios for single items are presented in Table 11. Mean sensitivity was quite low (< 30%) across conditions, with particularly low values (< 10%) in the largest sample size condition when misfit was due to either multidimensionality or generation from a competing model, regardless of the proportion of misfitting items. Mean specificity was approximately 70% across all conditions, but highest for the largest sample size condition when misfit was due to either multidimensionality or generation from a competing model, particularly when 33% of items were misfitting (> 90%).

Hypothesis Testing

Mean χ^2/df ratios. Hypothesis 2.1 addressed the potential interaction between sample size, type of misfit, and amount of misfit on the magnitude of mean χ^2/df ratios

Table 11

Descriptive Statistics: Distribution of Sensitivity and Specificity of χ^2/df Ratios for Single Items by Sample Size, Type of Misfit, and Proportion of Misfitting Items

Condition	Multidimensionality		DIF		Competing Model (GPCM)	
	Sensitivity M (SD)	Specificity M (SD)	Sensitivity M (SD)	Specificity M (SD)	Sensitivity M (SD)	Specificity M (SD)
10% items misfitting						
N = 400	0.24 (0.26)	0.70 (0.20)	0.27 (0.29)	0.72 (0.16)	0.22 (0.24)	0.73 (0.15)
N = 1500	0.22 (0.24)	0.70 (0.16)	0.33 (0.31)	0.67 (0.18)	0.20 (0.24)	0.70 (0.16)
N = 10000	0.10 (0.24)	0.76 (0.21)	0.24 (0.27)	0.71 (0.17)	0.04 (0.12)	0.79 (0.14)
33% items misfitting						
N = 400	0.29 (0.22)	0.60 (0.33)	0.28 (0.20)	0.72 (0.17)	0.21 (0.14)	0.76 (0.13)
N = 1500	0.20 (0.14)	0.76 (0.14)	0.32 (0.22)	0.68 (0.19)	0.20 (0.15)	0.76 (0.14)
N = 10000	0.06 (0.15)	0.91 (0.16)	0.25 (0.20)	0.73 (0.17)	0.04 (0.07)	0.93 (0.08)

for single items, pairs of items, and triplets of items. This hypothesis was tested by fitting generalized least squares linear models with weighted least squares, with corrected alpha set at .003. Separate models were fit for three response variables: the inverse of the mean χ^2/df ratios for (a) single items, (b) pairs of items, and (c) triplets of items. For each model, residual plots were examined and were deemed appropriate. See Table 12 for detailed results. Overall, the three-way interaction of sample size, type of misfit, and amount of misfit was statistically significant ($p < .001$). Two reduced nested models (one including two-way interactions only and one including main effects only) for each response variable were also fit and compared to the full models; AIC values and results of likelihood ratio tests indicated that in each case, the full model demonstrated significantly better model fit. See Table 13 for model fit results. Interaction plots are presented in Figure 12, for mean χ^2/df ratios associated with single items, pairs of items, and triplets of items, respectively. Several observations can be made from Table 12 and Figure 12. First, in general, the magnitude of mean χ^2/df ratios decreases from single items to pairs to triplets. In the small and medium sample size conditions, as the proportion of misfitting items increases from 10% to 33%, the mean χ^2/df ratios increase when the misfit is due to multidimensionality (e.g., for pairs of items and N = 400, M = 1.82 when 10% of items are misfitting, and M = 2.67 when 33% of items are misfitting), but do not increase when misfit is due to DIF or generation from a competing model. When N = 10000, the effect of amount of misfit appears more pronounced, depending on the type of misfit; mean χ^2/df ratios appear similar whether there is 10% or 33% misfit present under the DIF condition, but they increase fairly dramatically when misfit is due to multidimensionality (e.g., from M = 1.50 to M = 3.08 for pairs of items). In the largest

Table 12

Linear Models (Weighted Least Squares): Effects of Sample Size, Type of Misfit, and Amount of Misfit on Mean χ^2/df Ratios[a]

	Single Items				Pairs of Items				Triplets of Items			
Effect	*Coef.*	*se*	*t*	*p*	*Coef.*	*se*	*t*	*p*	*Coef.*	*Se*	*t*	*p*
(Intercept)	0.45	0.00	119.25	<.001	0.78	0.00	187.35	<.001	1.14	0.00	321.72	<.001
Sample (n=1500)	-0.03	0.00	-5.79	<.001	-0.16	0.00	-30.78	<.001	-0.39	0.00	-90.16	<.001
Sample (n=400)	0.00	0.00	0.20	.84	-0.17	0.00	-31.79	<.001	-0.42	0.00	-82.47	<.001
Type (model)	0.09	0.01	16.51	<.001	-0.01	0.00	-1.06	.29	-0.06	0.00	-13.27	<.001
Type (multidimensional)	0.07	0.01	10.55	<.001	0.00	0.01	0.67	.50	-0.05	0.01	-7.11	<.001
Amount (33%)	0.02	0.00	4.04	<.001	0.01	0.01	2.29	.02	0.04	0.00	7.55	<.001
Sample (n=1500):Type (model)	-0.08	0.01	-10.08	<.001	0.01	0.01	0.91	.36	0.06	0.00	10.96	<.001
Sample (n=400): Type (model)	-0.09	0.01	-11.35	<.001	0.01	0.01	1.43	.15	0.07	0.01	10.55	<.001
Sample (n=1500): Type (multidimensional)	-0.05	0.01	-6.22	<.001	0.00	0.01	0.36	.71	0.06	0.01	7.26	<.001
Sample (n=400): Type (multidimensional)	-0.08	0.01	-9.44	<.001	-0.02	0.01	-2.73	.01	0.03	0.01	3.48	<.001
Sample (n=1500): Amount (33%)	-0.01	0.01	-1.77	.08	0.00	0.01	0.54	.59	-0.02	0.01	-2.92	<.001
Sample (n=400): Amount (33%)	-0.02	0.01	-2.59	.01	0.00	0.01	0.06	.95	-0.03	0.01	-4.45	<.001
Type (model): Amount (33%)	0.22	0.01	26.28	<.001	-0.55	0.01	-82.32	<.001	-0.77	0.01	-143.91	<.001
Type (multidimensional): Amount (33%)	0.22	0.01	23.05	<.001	-0.38	0.01	-45.05	<.001	-0.52	0.01	-63.74	<.001
Sample (n=1500): Type (model): Amount (33%)	-0.18	0.01	-16.44	<.001	0.34	0.01	40.59	<.001	0.55	0.01	79.63	<.001
Sample (n=400): Type (model): Amount (33%)	-0.20	0.01	-18.15	<.001	0.48	0.01	54.37	<.001	0.71	0.01	76.61	<.001

(table continues)

Table 12, continued

	Single Items				Pairs of Items				Triplets of Items			
Effect	*Coef.*	*se*	*t*	*p*	*Coef.*	*se*	*t*	*p*	*Coef.*	*Se*	*t*	*p*
Sample (n=1500): Type (multidimensional): Amount (33%)	-0.18	0.01	-15.22	<.001	0.29	0.01	28.89	<.001	0.45	0.01	48.71	<.001
Sample (n=400): Type (multidimensional): Amount (33%)	-0.27	0.01	-20.50	<.001	0.30	0.01	24.14	<.001	0.43	0.01	31.42	<.001

[a]Response variables are the inverse of the mean χ^2/df ratios.

Table 13

Comparing Nested Models: Full Model versus Reduced Models

		Single Items			Pairs of Items			Triplets of Items		
Model	*df*	AIC	BIC	LL	AIC	BIC	LL	AIC	BIC	LL
Sample:Type:Amount	36	-24459	-24178	12265*	-37839	-37559	18956*	-37055	-36776	18563*
Sample:Type + Sample:Amount + Type:Amount	32	-23890	-23640	11977*	-35417	-35168	17741*	-33273	-33026	16668*
Sample + Type + Amount	24	-21979	-21792	11013	-24066	-23878	12057	-21146	-20960	10597

Note. df = degrees of freedom. AIC = Akaike information criterion. BIC = Bayesian information criterion. LL = log likelihood.
*Model fits significantly better than next simpler model per likelihood ratio tests, $p < .001$.

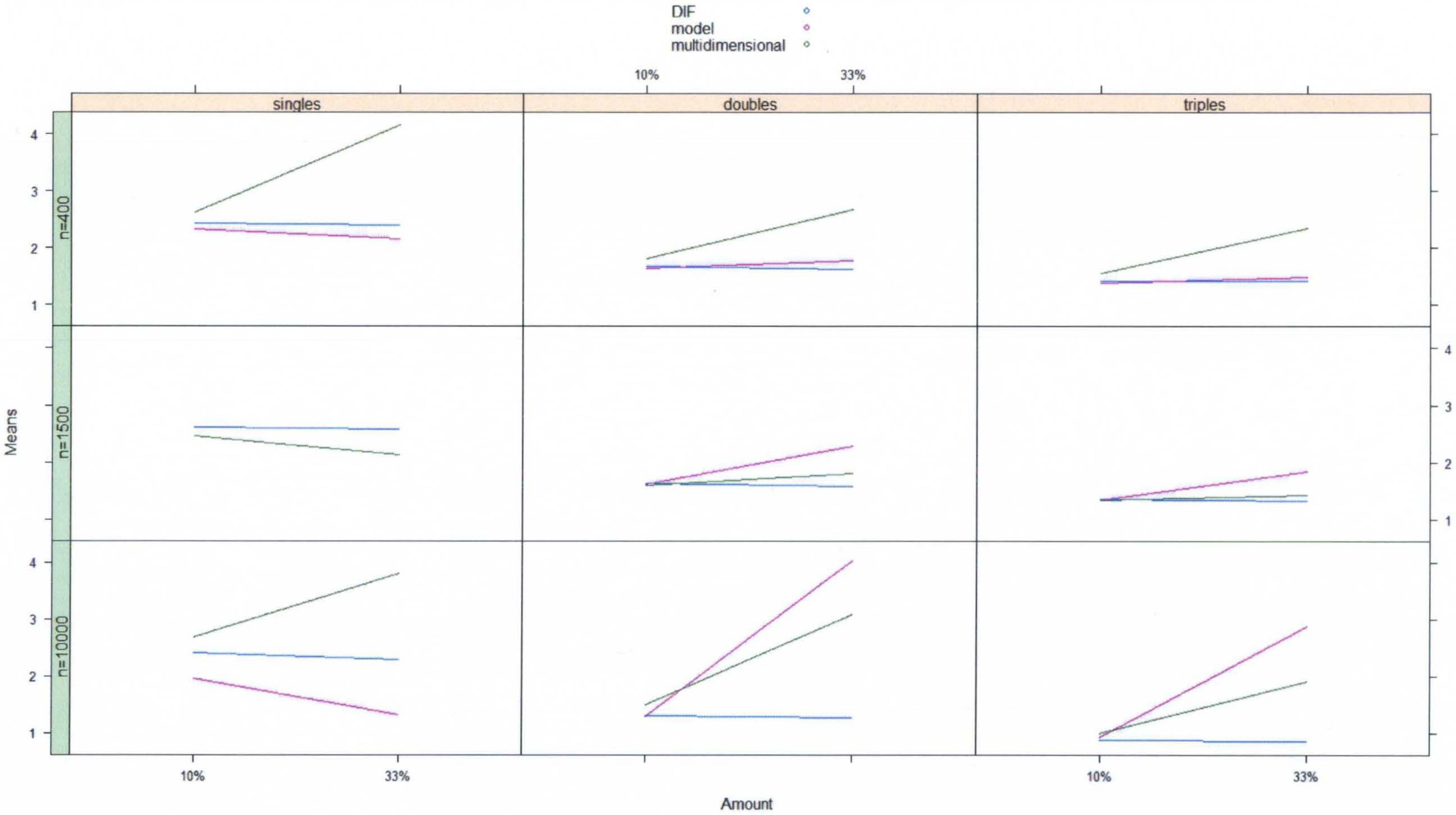

Figure 12. **Interaction plots of mean χ^2/df ratios for single items, pairs of items, and triplets of items, by sample size, type of misfit, and amount of misfit.** Note that mean χ^2/df ratios for single items in the N=1500 condition are equal for misfit due to model and multidimensionality at 10% and 33% misfitting items, so only one line appears.

sample size condition, when misfit is due to generation from a competing model, the direction of the effect of proportion of misfitting items is different among single items (from M = 1.97 to M = 1.33) versus pairs (from M = 1.30 to M = 4.04) and triplets (from M = 0.93 to M = 2.87) of items. Due to the detection of the significant interaction effect hypothesized in Hypothesis 2.1, Hypotheses 2.2 through 2.7 (regarding two-way interactions and main effects) were not tested.

Sensitivity and specificity of single items' χ^2/df ratios. Hypothesis 3.1 addressed the potential interaction of sample size, type of misfit, and amount of misfit on the sensitivity and specificity of using the χ^2/df ratios > 3 cut point for single items. This hypothesis was tested by fitting generalized linear models with a logit link, with corrected alpha set at .003 for significance. Separate models were fit for the response variables of sensitivity and sensitivity. See Table 14 for detailed results. The three-way interaction effect of sample size, type of misfit, and amount of misfit was statistically significant ($p <$ .001). Results of likelihood ratio tests and AIC values indicated that neither of two reduced nested models (one including two-way interactions only and one including main effects only) fit better than the full model including the three-way interaction term. See Table 15 for model fit results. Figure 13 presents box plots for the sensitivity and specificity of using the > 3 cut point for mean χ^2/df ratios to indicate misfit of single items. The primary observation from Table 14 and Figure 13 is that the lowest sensitivity levels are seen in the largest sample size condition when type of misfit is either due to multidimensionality or to generation from a competing model. When 10% of items are misfitting, sensitivity is 10% and 4% for these conditions, respectively; when 33% of items are misfitting, sensitivity is 6% and 4% for the same conditions. Concomitant

Table 14

Logistic Regression Models: Effects of Sample Size, Type of Misfit, and Amount of Misfit on Sensitivity and Specificity of $\chi^2/df > 3$ for Single Items

	Sensitivity				Specificity			
Effect	*Coef.*	*se*	*Z*	*p*	*Coef.*	*se*	*Z*	*p*
(Intercept)	-1.14	0.04	-26.84	<.001	0.89	0.01	65.70	<.001
Sample (n=1500)	0.44	0.06	7.71	<.001	-0.16	0.02	-8.52	<.001
Sample (n=400)	0.18	0.06	2.98	<.01	0.06	0.02	3.41	<.001
Type (model)	-2.04	0.10	-19.85	<.001	0.47	0.02	23.60	<.001
Type (multidimensional)	-1.10	0.07	-14.62	<.001	0.26	0.02	13.30	<.001
Amount (33%)	0.05	0.05	0.96	.33	0.12	0.02	5.79	<.001
Sample (n=1500):Type (model)	1.38	0.12	11.65	<.001	-0.36	0.03	-13.16	<.001
Sample (n=400): Type (model)	1.74	0.12	14.65	<.001	-0.41	0.03	-14.83	<.001
Sample (n=1500): Type (multidimensional)	0.53	0.09	5.55	<.001	-0.14	0.03	-5.34	<.001
Sample (n=400): Type (multidimensional)	0.93	0.10	9.76	<.001	-0.37	0.03	-13.47	<.001
Sample (n=1500): Amount (33%)	-0.10	0.07	-1.48	.14	-0.08	0.03	-2.74	<.01
Sample (n=400): Amount (33%)	-0.04	0.07	-0.62	.53	-0.12	0.03	-4.11	<.001
Type (model): Amount (33%)	-0.05	0.12	-0.39	.69	1.07	0.04	28.57	<.001
Type (multidimensional): Amount (33%)	-0.49	0.09	-5.52	<.001	1.00	0.03	28.72	<.001
Sample (n=1500): Type (model): Amount (33%)	0.09	0.13	0.68	.49	-0.76	0.05	-16.05	<.001
Sample (n=400): Type (model): Amount (33%)	-0.01	0.13	-0.06	.95	-0.89	0.05	-18.65	<.001
Sample (n=1500): Type (multidimensional): Amount (33%)	0.50	0.11	4.12	<.001	-0.70	0.05	-15.41	<..01
Sample (n=400): Type (multidimensional): Amount (33%)	0.73	0.11	6.54	<.001	-1.44	0.05	-31.99	<.001

Table 15

Comparing Nested Models: Full Model versus Reduced Models

		Sensitivity			Specificity		
Model	*df*	AIC	BIC	LL	AIC	BIC	LL
Sample:Type:Amount	18	50837	50977	-25400*	121666	121806	-60815*
Sample:Type + Sample:Amount + Type:Amount	14	50880	50989	-25426*	122863	122972	-61418*
Sample + Type + Amount	6	52514	52560	-26251	127648	127695	-63818

Note. df = degrees of freedom. AIC = Akaike information criterion. BIC = Bayesian information criterion. LL = log likelihood.

*Model fits significantly better than next simpler model per likelihood ratio tests, $p < .001$.

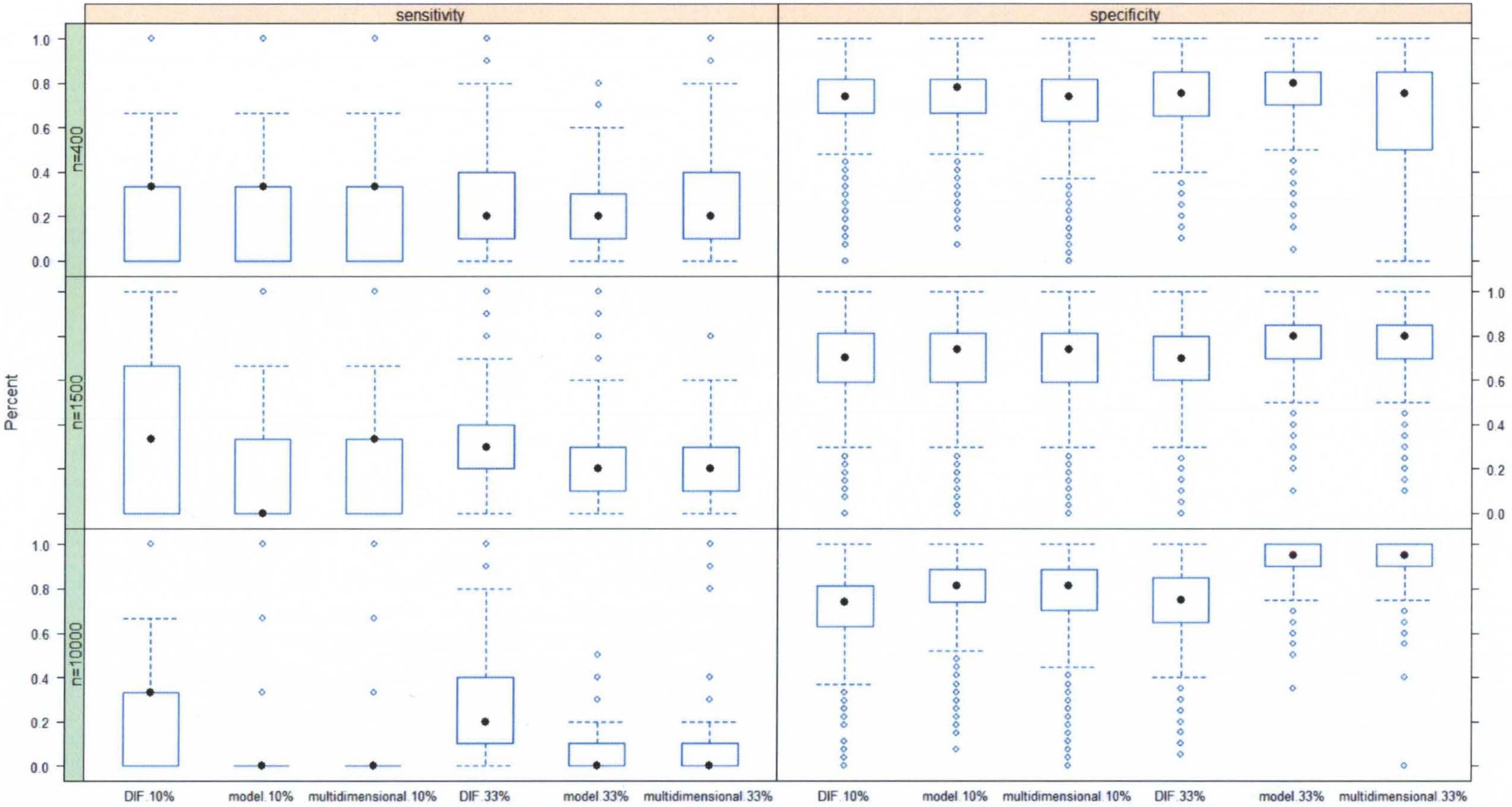

Figure 13. **Box plot of sensitivity and specificity of χ^2/df > 3 for single items.**

increases in specificity are also observed. Sensitivity is higher in the small and medium sample size conditions across all levels of type and proportion of misfitting items (see Table 11 for descriptive statistics per condition). Due to the detection of the significant interaction effect hypothesized in Hypothesis 3.1, Hypotheses 3.2 through 3.7 (regarding two-way interaction and main effects) were not tested.

CHAPTER V

DISCUSSION

Applied IRT researchers face several challenges, including traditional reliance on standalone IRT software with limited output and a lack of simple solutions for assessing how well a given item fits the selected IRT model. One solution to the latter difficulty, used by many applied IRT researchers investigating a variety of topics, has been to use the χ^2/df ratios method, developed by Drasgow et al. (1995) and easily implemented using the freely available MODFIT program (Stark, 2002). The developers of this approach are in the educational psychology field, in which many applications of IRT employ large datasets (N > 10000). However, *users* of the χ^2/df ratios method have investigated item fit with very small samples (N < 300), with no published guidance regarding its use outside of large sample research. In addition, item misfit can be caused by several issues and can be present in varying proportions within a given set of items, and the effects of these factors on detection of item misfit using χ^2/df ratios is unknown. Thus, this study aimed to systematically investigate the utility of χ^2/df ratios for detecting item misfit in applications of one frequently used IRT model—the graded response model (GRM)—as impacted by sample size, type of misfit, and proportion of misfitting items.

Summary of Findings

Research Question 1: Are adjusted (to N = 3000) or unadjusted χ^2/df ratios more appropriate for small-sample IRT research?

The use of adjusted χ^2/df ratios, in which the sample size is adjusted to 3000 to "standardize" findings and allow comparisons of item fit results across studies, is built into the MODFIT calculations and output. Applied IRT researchers routinely report adjusted χ^2/df ratios, regardless of their studies' sample sizes. Results of the current investigation suggest that adjusted χ^2/df ratios were appropriate for the largest sample size condition (N = 10000), but were extremely inflated for the small (N = 400) and medium (N = 1500) sample size conditions. Using the "rule of thumb" cut point of $\chi^2/df > 3$ to indicate item misfit, nearly all items in a 30-item set were identified as misfitting based on adjusted χ^2/df ratios in the small and medium sample size conditions, when in fact, no misfitting items were present. In contrast, use of unadjusted χ^2/df ratios in the small and medium sample size conditions resulted in far fewer (but still > 0) items being incorrectly flagged as misfitting, with lower proportions incorrectly flagged for χ^2/df ratios calculated (a) for pairs and triplets of items, compared to single items, and (b) for the medium sample size, compared to the small sample size. Uniformly lower percentages of items were incorrectly flagged as misfitting in the largest sample size condition when the adjusted χ^2/df ratios were used, as desired. Thus, under Hypothesis 1.1, the null hypothesis is rejected; there is a significant interaction effect of sample size and adjustment condition on the magnitude of χ^2/df ratios. For small-sample (N ≤ 1500) IRT research, the exclusive use of unadjusted χ^2/df ratios is recommended.

Research Question 2: As a means of assessing model fit for the GRM, how are the magnitude and utility of χ^2/df ratios affected by (a) sample size, (b) type of item misfit, and (c) proportion of misfitting items in a given set?

Results of this study suggest that χ^2/df ratios are differentially affected at different sample sizes by the type of misfit and proportion of misfitting items. For example, the mean χ^2/df ratios calculated for single items were highest when 33% of items were actually misfitting due to multidimensionality, and were lowest when 10% of items were misfitting due to generation from a competing IRT model. Effects of these three factors also differed depending on whether the χ^2/df ratios were averaged across single items, pairs of items, or triplets of items, complicating interpretation of the significant three-way interaction for the three response variables. Importantly, the distributions of mean χ^2/df ratios were quite skewed in several study conditions (see Figure 10). Thus, high proportions of χ^2/df ratios > 3 could be observed in conditions with relatively low mean ratios. This was especially true for χ^2/df ratios calculated for pairs and triplets of items, in which ratios in certain conditions tended to be very low for certain pairs and triplets but very high for others, resulting in low means, high standard deviations, and high proportions of ratios > 3 (e.g., N = 10000 with 33% of items misfitting due to multidimensionality or model). This finding was consistent with Drasgow and colleagues' (1995) and Stark's (2002) rationale that pairs and triplets of items with similar misfit should generate higher χ^2/df ratios than either (a) single items alone, or (b) pairs or triplets with dissimilar misfit.

To assess the utility of the χ^2/df ratios for identifying item misfit across the study conditions, the specificity and sensitivity of using the $\chi^2/df > 3$ cut point was investigated, for single items only. Results suggested that sensitivity was notably low across all conditions, ranging from a low of 4% (when N = 10000 and either 10% or 33% of items were misfitting due to generation from a competing model) to a high of 33% (when N = 1500 and 10% of items were misfitting due to DIF). Specificity was fairly high across all study conditions, ranging from a low of 60% (when N = 400 and 33% of items were misfitting due to multidimensionality) to a high of 93% (when N = 10000 and 33% of items were missing due to generation from a competing model). In general, sensitivity appeared to decrease as the sample size increased, particularly when misfit was due to multidimensionality or generation from a competing model. Under Research Question 2, for both Hypothesis 2.1 and 3.1, the null hypotheses were rejected, given the significant three-way interaction effects of sample size, type of misfit, and proportion of misfitting items on the magnitude, sensitivity, and specificity of χ^2/df ratios.

Limitations

Several limitations of this study should be highlighted. First, these results are specific to the GRM (and the GPCM, in one condition of type of misfit); many other IRT models may be employed. Findings may differ for other IRT models. Next, only three sample size conditions were tested. In practice, sample sizes in applied IRT research vary dramatically. The focus of the current study was only on comparing small ($N \leq 1500$) sample sizes to a large (N = 10000) sample size. Thus, results cannot necessarily be generalized to sample sizes not tested. Similarly, only three types of item misfit were tested: misfit due to multidimensionality, generation from a competing model, and DIF. Other types of item misfit may exist. Sensitivity and specificity were only calculated for single items' $\chi^2/df > 3$; however, since mean χ^2/df ratios computed for pairs and triplets of items tended to be lower but increase dramatically in certain conditions, this decision may have resulted in a "worst case scenario" picture of sensitivity and specificity. Further, it is important to note that no direct comparison was made between the χ^2/df ratios calculated in R and those generated using the MODFIT program, so conclusions regarding χ^2/df ratios should not be extended to MODFIT until the equivalence of these methods is established. On a related note, the developers of MODFIT (Stark, 2002) recommend the review of fit plots for each item, in addition to considering the distribution of χ^2/df ratios, in determining item fit. Fit plots were not generated in the current study, which focused solely on χ^2/df ratios. Finally, simulation functions in the R package *ltm* were used; replication of results would be beneficial.

Directions for Future Research

The current findings should be replicated and extended in several ways. For example, including a sample size condition between N = 1500 and N = 10000 would be helpful for applied IRT researchers seeking guidance regarding the use of χ^2/df ratios to assess item fit with the GRM. Different cut points and decision rules for the χ^2/df ratios could also be investigated (e.g., using *p*-values with alpha corrected for multiple comparisons instead of the $\chi^2/df > 3$ rule of thumb) to determine whether other approaches may improve sensitivity and specificity of the χ^2/df ratios index of misfit. In this study, sensitivity and specificity were calculated only for single items' $\chi^2/df > 3$. As the χ^2/df ratios appeared to behave differently when computed for pairs of items and triplets of items across study conditions, it would be informative to develop a way of assessing sensitivity and specificity using the ratios calculated for item pairs and triplets instead of single items only. Finally, comparison of χ^2/df ratios generated with R to those produced by MODFIT would be directly relevant to applied IRT researchers using this approach, as would investigation of the effects of sample size, type of misfit, and proportion of misfitting items on item fit plots, which may be used as an additional indicator of item fit.

Summary and Conclusions

In summary, applied IRT researchers using the χ^2/df ratio index for assessing item fit in the GRM should be aware of important considerations. First, χ^2/df ratios are affected by sample size in several ways. The use of unadjusted χ^2/df ratios is recommended for applications of the GRM with small sample sizes ($N \leq 1500$), as adjusted χ^2/df ratios are inappropriately inflated in these conditions; adjusted χ^2/df ratios, however, are recommended for large sample sizes (e.g., $N = 10000$). Sample size also interacts with type of item misfit and proportion of misfitting items to affect the magnitude, sensitivity, and specificity of χ^2/df ratios used to assess item fit.

Certain types of item misfit (e.g., generation from a competing model and multidimensionality) are associated with lower sensitivity of the $\chi^2/df > 3$ cut point for single items' ratios when the sample size is large. In addition, sensitivity of the $\chi^2/df > 3$ cut point for single items is quite low. This finding is consistent with Drasgow and colleagues' (1995) and Stark's (2002) rationale for examining ratios calculated for pairs and triplets of items, in addition to single items, to reveal misfit not detectable with single item χ^2/df ratios.

Finally, Drasgow and colleagues and Stark advocate the use of fit plots in addition to consideration of the distribution of χ^2/df ratios in determining item fit in IRT analyses. While easily accessible in MODFIT, fit plots are often not included in articles by applied IRT researchers, who frequently only provide tables summarizing the distributions of adjusted and unadjusted χ^2/df ratios. Given the current study's findings regarding effects of sample size, type of misfit, and proportion of misfit on χ^2/df ratios, the use and inclusion of item fit plots may be helpful, at least until further information is available regarding the performance of fit plots in conditions varying by the same three factors.

REFERENCES

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Akaike, H. (1974). A new look at statistical model identification. *Ieee Transactions on Automatic Control, 19*, 716-723.

Baker, F. B. (1992). Equating tests under the Graded Response Model. *Applied Psychological Measurement, 16*, 87-96.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bjorner, J. B., Smith, K. J., Stone, C. A., & Sun, X. (2007). *IRTFIT: A macro for item fit and local dependence tests under IRT models*. Lincoln, RI: Quality Metric Inc.

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

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*, 21-33.

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

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

Bolt, D. M., Hare, R. D., & Neumann, C. S. (2007). Score metric equivalence of the Psychopathy Checklist-Revised (PCL-R) across criminal offenders in North America and the United Kingdom - A critique of Cooke, Michie, Hart, and Clark (2005) and new analyses. *Assessment, 14*, 44-56. doi: 10.1177/1073191106293505

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*, 523-562.

Cooper, A., & Petrides, K. V. (2010). A psychometric analysis of the Trait Emotional Intelligence Questionnaire-Short Form (TEIQue-SF) using item response theory. *Journal of Personality Assessment, 92*, 449-457. doi: Pii 925549959 10.1080/00223891.2010.497426

De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: The Guilford Press.

Drasgow, F., Levine, M. V., & Mclaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indexes. *Applied Psychological Measurement, 11*, 59-79.

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

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

Estrada, A. X., Probst, T. M., Brown, J., & Graso, M. (2011). Evaluating the psychometric and measurement characteristics of a measure of sexual orientation harassment. *Military Psychology, 23*, 220-236.

Fryback, D. G., Palta, M., Cherepanov, D., Bolt, D., & Kim, J. S. (2010). Comparison of 5 health-related quality-of-life indexes using item response theory analysis. *Medical Decision Making, 30*, 5-15. doi: 10.1177/0272989x09347016

Gomez, R. (2008). Item response theory analyses of the parent and teacher ratings of the DSM-IV ADHD rating scale. *Journal of Abnormal Child Psychology, 36*, 865-885. doi: 10.1007/s10802-008-9218-8

Gomez, R., & Fisher, J. W. (2005). The spiritual well-being questionnaire: Testing for model applicability, measurement and structural equivalencies, and latent mean differences across gender. *Personality and Individual Differences, 39*, 1383-1393. doi: 10.1016/j.paid.2005.03.023

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Norwell, MA: Kluwer Academic Publishers.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Kim, G., Chiriboga, D. A., & Jang, Y. (2009). Cultural equivalence in depressive symptoms in older white, black, and Mexican-American adults. *Journal of the American Geriatrics Society, 57*, 790-796. doi: 10.1111/j.1532-5415.2009.02188.x

Lampenius, N., & Zickar, M. J. (2005). Development and validation of a model and measure of financial risk-taking. *Journal of Behavioral Finance, 6*, 129-143.

Lautenschlager, G. J., Meade, A. W., & Kim, S. H. (2006). *Cautions regarding sample characteristics when using the graded response model.* Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Lord, F. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-548.

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

Maydeu-Olivares, A. (2005). Further empirical results on parametric versus non-parametric IRT modeling of likert-type personality data. *Multivariate Behavioral Research, 40*, 261-279.

Mckinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*, 49-57.

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

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 48-62.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X-2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289-298. doi: Doi 10.1177/0146621603253011

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Reise, S. P. (1990). A Comparison of item-fit and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14*, 127-137.

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

Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance, 14*, 187-207.

Samejima, F. (1969). Calibration of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 17.*

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1*, 223-245.

Schmidt, M. J., Kihm, J. A., & Robie, C. (2000). Development of a global measure of personality. *Personnel Psychology, 53*, 153-193.

Stark, S. (2002). MODFIT [computer program]. Retrieved from http://io.psych.uiuc.edu/irt/mdf_modfit.asp

Stark, S., Chernyshenko, O. S., Lancaster, A. R., Drasgow, F., & Fitzgerald, L. F. (2002). Toward standardized measurement of sexual harassment: Shortening the SEQ-DoD using item response theory. *Military Psychology, 14*, 49-72.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40*, 331-352.

Tay-Lim, B. S.-H., & Harwell, M. (1997). *Effects of number of items and examinees on parameter estimation in item response theory: A research synthesis.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Teresi, J. A. (2001). Statistical methods of examination of differential item functioning with applications to cross-cultural measurement of functional, physical, and mental health. *Journal of Mental Health and Aging, 7*, 31-40.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567-577.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123-139.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, MJ: Lawrence Erlbaum Associates.

Ware, J. E. (2003). Conceptualization and measurement of health-related quality of life: Comments on an evolving field. *Archives of Physical Medicine and Rehabilitation, 84*, S43-S51. doi: 10.1053/apmr.2003.50246

Weiss, D. J. (1967). *Manual for the Minnesota Satisfaction Questionnaire.* Minneapolis, MN: University of Minnesota Industrial Relations Center.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

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*, 180-191.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20*, 71-87.

APPENDIX A

Study Design for Each Research Question

Research Question 1: Sample Size and Adjustment Condition

Number of simulated replications in each cell:

Sample Size	*Adjustment Condition*[a]		
		Adjusted χ^2/df Ratios	Unadjusted χ^2/df Ratios
	N = 400	1000	1000
	N = 1500	1000	1000
	N = 10000	1000	1000

[a] A repeated factor in analyses.

Research Question 2: Sample Size, Type of Misfit, and Proportion of Misfitting Items

Number of simulated replications in each cell:

Sample Size	Proportion of Misfitting Items	Type of Misfit		
		Multidimensional	DIF	Competing Model
N = 400				
	10%	1000	1000	1000
	33%	1000	1000	1000
N = 1500				
	10%	1000	1000	1000
	33%	1000	1000	1000
N = 10000				
	10%	1000	1000	1000
	33%	1000	1000	1000

APPENDIX B

R Code for Data Simulation and Preparation

```
###common code for all simulations in RQ1 and RQ2

##creating matrices to store the results ##
outBetas <- matrix(0, M, 150) # store the results
outsinglets<- matrix(0, M, 30) # store singlet chisq/dfs
poutsinglets<- matrix(0, M, 30) # store singlet p-values (NOT adj for multiple
comparisons)
adjoutsinglets<-matrix(0, M, 30) # store adjusted singlet chisq/dfs
padjoutsinglets<-matrix(0, M, 30) # store adjusted singlet p-values (NOT adj for multiple
comparisons)
outdoublets<- matrix(0, M, 435)# store doublets chisq/dfs
poutdoublets<- matrix(0, M, 435)# store doublets p-values (NOT adj for multiple
comparisons)
adjoutdoublets<-matrix(0, M, 435) # store adjusted doublet chisq/dfs
padjoutdoublets<-matrix(0, M, 435) # store adjusted doublets p-values (NOT adj for
multiple comparisons)
outtriplets<-matrix (0, M, 4060)#store triplets chisq/dfs
pouttriplets<-matrix (0, M, 4060)#store triplets p-values
adjouttriplets<-matrix (0, M, 4060)# store adjusted triplets chisq/dfs
padjouttriplets<-matrix (0, M, 4060) # store adjusted triplets p-values (NOT adj for
multiple comparisons)
cpoutsinglets.bon<- matrix(0, M, 30) # store singlet p-values (bonferroni)
cpadjoutsinglets.bon<-matrix(0, M, 30) # store adjusted singlet p-values (bonferroni)
cpoutsinglets.bh<- matrix(0, M, 30) # store singlet p-values (benjamini-hochman)
cpadjoutsinglets.bh<-matrix(0, M, 30) # store adjusted singlet p-values (benjamini-
hochman)
cpoutdoublets.bon<- matrix(0, M, 435)# store doublet p-values (bonferroni)
cpadjoutdoublets.bon<-matrix(0, M, 435) # store adjusted doublet p-values (bonferroni)
cpoutdoublets.bh<- matrix(0, M, 435)# store doublet p-values (benjamini-hochman)
cpadjoutdoublets.bh<-matrix(0, M, 435) # store adjusted doublet p-values (benjamini-
hochman)
cpouttriplets.bon<-matrix (0, M, 4060)# store triplet p-values (bonferroni)
cpadjouttriplets.bon<-matrix (0, M, 4060) # store adjusted triplet p-values (bonferroni)
cpouttriplets.bh<-matrix (0, M, 4060)# store triplet p-values (benjamini-hochman)
cpadjouttriplets.bh<-matrix (0, M, 4060) # store adjusted triplet p-values (benjamini-
hochman)
```

```
##SIMULATION FOR RQ 1: Sample size & adjustment condition

library(ltm)
library(multtest)

###the simulation function from ltm package

rmvordlogis <-  function (n, betas) {
    # function to simulate random responses
    # based on the Graded Response Model
    # using the additive parameterization
    ###deleted line setting p since it will always be 30
    ###deleted ncatg line since it will always be 5
    z <-  rnorm(n)
    gammas <-  lapply(betas, function (x) {
       nx <-  length(x)
       cbind(plogis(matrix(x[-nx], n, nx-1, TRUE)- x[nx] * z), 1)
    })
    prs <-  lapply(gammas, function (x) {
       nc <-  ncol(x)
       cbind(x[, 1], x[, 2:nc]-x[, 1:(nc-1)])
    })
    out <-  matrix(0, n, 30) ##replaced p with 30
    for (j in 1:30) { ## replaced p with 30
       for (i in 1:n) {
          out[i, j] <-  sample(5, 1, prob = prs[[j]][i, ])
          ##changed ncatg[j] to 5 since always 5 categories
       }
    }
    out
}

##the iprobs function
`iprobs` <-
function (betas, z) {
   n <- length(z)
   gammas <- lapply(betas, function (x) {
      nx <- length(x)
      cbind(plogis(matrix(x[-nx], n, nx - 1, TRUE) - x[nx] * z), 1)
   })
   lapply(gammas, function (x) {
      nc <- ncol(x)
      cbind(x[, 1], x[, 2:nc] - x[, 1:(nc - 1)])
   })
```

```
}

# take the betas from dataset A transformed
# as the true betas
true.betas<- read.csv('C:/Documents and Settings/crclar0/Desktop/DatasetAt.csv',
header=T, row.names=1)
n <- 400 #start with sample size N=400, change this line to 1500 and 10000 for other
sample sizes
M <- 1000 # number of simulations

### SIMULATING DATA, FITTING THE GRM, OBTAINING UNADJUSTED AND
ADJUSTED OUTCOMES ###
ind <- i <- 1
while(i <= M) {
    set.seed(100 + ind) # for reproducible results
        ind <- ind+1
        n<-400 ##again, change to 1500 and 10000 for other sample size conditions
        data <-  rmvordlogis(n, true.betas)
        indA<-sample(1:n, floor(n/2), replace=FALSE)
        dataA <- data[indA, ]
        dataB <- data[-indA, ]
        fit <-  try(grm(dataA)) ###if there is an error will not just stop, will go on to next i
        if(class(fit)=="try-error") next
        if(length(unlist(fit$coefficients))!=150) next
        outBetas[i, ] <-  unlist(fit$coefficients)  ## non standard param

        ##setting up for chisquare/df routines##
        X <- fit$X
        nams <- colnames(X)
        n <- nrow(X)
        betas <- fit$coef
        p <- length(betas)
        pr <- iprobs(betas, fit$GH$Z)  ##iprobs at top for reference
        GHw <- fit$GH$GHw
        X <- data.matrix(fit$X)[complete.cases(X), ]

        ###UNADJUSTED FOR  SINGLE ITEMS  ###
        #sindex <- ncol(X) ##don't need this, always 30
        #margins <- vector("list", sindex)
         for (j in 1:30) {
                p1 <- pr[[j]]
                        ncp1 <- ncol(p1)
                 #####for obs below - data is the cross-validation sample##
                        obs <- table(factor(dataB[,j], levels=1:5))
                        exp <- obs
                        exp <- n * colSums(GHw * p1)
        ##summing cells with EXPECTED counts <5
        sind <- which(exp<5)
        if (length(sind)>0) {
                obs <- c(obs[-sind], sum(obs[sind]))
                exp <- c(exp[-sind], sum(exp[sind]))
                }
        ##all cells w/ EXP < 5 are added to the cell w/the next smallest EXP > 5
        sind <- which(exp<5)
        if (length(sind)>0) {
        ind2 <- which.min(exp[-sind])
                obs <- c(obs[-sind][-ind2], sum(obs[sind], obs[-sind][ind2]))
                exp <- c(exp[-sind][-ind2], sum(exp[sind],exp[-sind][ind2]))
                }
        ##calculating df after collapsing cells for each routine above
        df <- length(exp)-1
        resid <- (obs - exp)^2/exp
                TotalResid <- sum(resid)
        schisqdf<-TotalResid/df
        outsinglets[i, j]<-schisqdf
        pvalue<-1-pchisq(TotalResid, df)
        poutsinglets[i, j]<-pvalue
        cpvalue<-pvalue*30
        cpoutsinglets.bon[i,j]<-cpvalue
        }

### ADJUSTED FOR SINGLE ITEMS (N=3000)###
#sindex <- ncol(X) ##don't need this, always 30
#margins <- vector("list", sindex)
adj<-3000/n
a<-3000
for (j in 1:30) {
                p1 <- pr[[j]]
                ncp1 <- ncol(p1)
        #####for obs below - data is the cross-validation sample##
                adjobs <- adj*table(factor(dataB[,j], levels=1:5))
                adjexp <- adjobs
                adjexp <- a * colSums(GHw * p1)
        ##summing cells with EXPECTED counts <5
        sind <- which(adjexp<5)
        if (length(sind)>0) {
                adjobs <- c(adjobs[-sind], sum(adjobs[sind]))
                adjexp <- c(adjexp[-sind], sum(adjexp[sind]))
                }
        ##all cells w/ EXP < 5 are added to the cell w/the next smallest EXP > 5
        sind <- which(adjexp<5)
        if (length(sind)>0) {
```

```
    ind2 <- which.min(adjexp[-sind])
    adjobs <- c(adjobs[-sind][-ind2], sum(adjobs[sind], adjobs[-sind][ind2]))
    adjexp <- c(adjexp[-sind][-ind2], sum(adjexp[sind],adjexp[-sind][ind2]))
        }
    ##calculating df after collapsing cells for each routine above
    df <- length(adjexp)-1
    resid <- (adjobs - adjexp)^2/adjexp
        TotalResid <- sum(resid)
    schisqdf<-TotalResid/df
    adjoutsinglets[i, j]<-schisqdf
    pvalue<-1-pchisq(TotalResid, df)
    padjoutsinglets[i, j]<-pvalue
    cpvalue<-pvalue*30
    cpadjoutsinglets.bon[i,j]<-cpvalue
    }

###UNADJUSTED FOR PAIRS OF ITEMS ###
index <- t(combn(p, 2))
dindex <- nrow(index)
 #margins <- vector("list", dindex)
 for (k in 1:dindex) {
        item1 <- index[k, 1]
        p1 <- pr[[item1]]
        ncp1 <- ncol(p1)
        item2 <- index[k, 2]
        p2 <- pr[[item2]]
        ncp2 <- ncol(p2)
    ######for obs below - data is the cross-validation sample##
    obs <- as.matrix(table(factor(dataB[, item1], levels=1:5), factor(dataB[,
    item2], levels=1:5)))
        pairs <- cbind(rep(1:ncp1, each = ncp2), rep(1:ncp2, ncp1))
        pp <- p1[, pairs[, 1]] * p2[, pairs[, 2]]
        exp <- obs
        exp[pairs] <- n * colSums(GHw * pp)
    ##summing cells with EXPECTED counts <5
    sind <- which(exp<5)
    if (length(sind)>0) {
        obs <- c(obs[-sind], sum(obs[sind]))
        exp <- c(exp[-sind], sum(exp[sind]))
        }
    ##all cells w/ EXP < 5 are added to the cell w/the next smallest EXP > 5
    sind <- which(exp<5)
    if (length(sind)>0) {
    ind2 <- which.min(exp[-sind])
        obs <- c(obs[-sind][-ind2], sum(obs[sind], obs[-sind][ind2]))
        exp <- c(exp[-sind][-ind2], sum(exp[sind],exp[-sind][ind2]))
```

```
                }
        ##calculating df after collapsing cells for each routine above
        df <- length(exp)-1
        resid <- (obs - exp)^2/exp
                TotalResid <- sum(resid)
        dchisqdf<-TotalResid/df
        outdoublets[i, k ]<-dchisqdf
        pvalue<-1-pchisq(TotalResid, df)
        poutdoublets[i, k]<-pvalue
        cpvalue<-pvalue*435
        cpoutdoublets.bon[i,k]<-cpvalue
        }

###FOR ADJUSTED PAIRS OF ITEMS (N=3000)  ###
adj<-3000/n
a<-3000
index <- t(combn(p, 2))
dindex <- nrow(index)
#margins <- vector("list", dindex)
 for (k in 1:dindex) {
                item1 <- index[k, 1]
                p1 <- pr[[item1]]
                ncp1 <- ncol(p1)
                item2 <- index[k, 2]
                p2 <- pr[[item2]]
                ncp2 <- ncol(p2)
        ######for obs below - data is the cross-validation sample##
        adjobs <- adj*as.matrix(table(factor(dataB[, item1], levels=1:5),
        factor(dataB[, item2], levels=1:5)))
                pairs <- cbind(rep(1:ncp1, each = ncp2), rep(1:ncp2, ncp1))
                pp <- p1[, pairs[, 1]] * p2[, pairs[, 2]]
                adjexp <- adjobs
                adjexp[pairs] <- a * colSums(GHw * pp)
        ##summing cells with EXPECTED counts <5
        sind <- which(adjexp<5)
        if (length(sind)>0) {
                adjobs <- c(adjobs[-sind], sum(adjobs[sind]))
                adjexp <- c(adjexp[-sind], sum(adjexp[sind]))
                }
        ##all cells w/ EXP < 5 are added to the cell w/the next smallest EXP > 5
        sind <- which(adjexp<5)
        if (length(sind)>0) {
        ind2 <- which.min(adjexp[-sind])
        adjobs <- c(adjobs[-sind][-ind2], sum(adjobs[sind], adjobs[-sind][ind2]))
        adjexp<- c(adjexp[-sind][-ind2], sum(adjexp[sind],adjexp[-sind][ind2]))
                }
```

```
    ##calculating df after collapsing cells for each routine above
    df <- length(adjexp)-1
    resid <- (adjobs- adjexp)^2/adjexp
        TotalResid <- sum(resid)
    dchisqdf<-TotalResid/df
    adjoutdoublets[i, k ]<-dchisqdf
    pvalue<-1-pchisq(TotalResid, df)
    padjoutdoublets[i, k]<-pvalue
    cpvalue<-pvalue*435
    cpadjoutdoublets.bon[i,k]<-cpvalue
    }

###UNADJUSTED  TRIPLETS OF ITEMS  ###
index <- t(combn(p, 3))
tindex <- nrow(index)
margins <- vector("list", tindex)
for (m in 1:tindex) {
        item1 <- index[m, 1]
        p1 <- pr[[item1]]
        ncp1 <- ncol(p1)
        item2 <- index[m, 2]
        p2 <- pr[[item2]]
        ncp2 <- ncol(p2)
        item3 <- index[m, 3]
        p3 <- pr[[item3]]
        ncp3 <- ncol(p3)
    obs <- as.array(table(factor(dataB[,item1], levels=1:5), factor(dataB[,
    item2], levels=1:5), factor(dataB[, item3], levels=1:5)))
    trips <- cbind(rep(1:ncp1, each = ncp2), rep(1:ncp2, ncp1))
    trips <- cbind(trips[rep(1:nrow(trips), ncp3), ], rep(1:ncp3, each
    nrow(trips)))
        pp <- p1[, trips[, 1]] * p2[, trips[, 2]] * p3[, trips[, 3]]
        exp <- obs
        exp[trips] <- n * colSums(GHw * pp)
    ##summing cells with EXPECTED counts <5
    sind <- which(exp<5)
    if (length(sind)>0) {
        obs <- c(obs[-sind], sum(obs[sind]))
        exp <- c(exp[-sind], sum(exp[sind]))
        }
    ##all cells w/ EXP < 5 are added to the cell w/the next smallest EXP > 5
    sind <- which(exp<5)
    if (length(sind)>0) {
    ind2 <- which.min(exp[-sind])
        obs <- c(obs[-sind][-ind2], sum(obs[sind], obs[-sind][ind2]))
        exp <- c(exp[-sind][-ind2], sum(exp[sind],exp[-sind][ind2]))
```

```
                }
        ##calculating df after collapsing cells for each routine above
        df <- length(exp)-1
        resid <- (obs - exp)^2/exp
                TotalResid <- sum(resid)
        tchisqdf<-TotalResid/df
                outtriplets[i, m]<-tchisqdf
        pvalue<-1-pchisq(TotalResid, df)
        pouttriplets[i, m]<-pvalue
        cpvalue<-pvalue*4060
        cpouttriplets.bon[i,m]<-cpvalue
        }

### ADJUSTED TRIPLETS OF ITEMS (N=3000)  ###
adj<-3000/n
a<-3000
index <- t(combn(p, 3))
 tindex <- nrow(index)
 margins <- vector("list", tindex)
 for (m in 1:tindex) {
                item1 <- index[m, 1]
                p1 <- pr[[item1]]
                ncp1 <- ncol(p1)
                item2 <- index[m, 2]
        p2 <- pr[[item2]]
                ncp2 <- ncol(p2)
        item3 <- index[m, 3]
        p3 <- pr[[item3]]
                ncp3 <- ncol(p3)
        adjobs <- adj*as.array(table(factor(dataB[,item1], levels=1:5),
        factor(dataB[, item2], levels=1:5), factor(dataB[, item3], levels=1:5)))
                trips <- cbind(rep(1:ncp1, each = ncp2), rep(1:ncp2, ncp1))
                trips <- cbind(trips[rep(1:nrow(trips), ncp3), ], rep(1:ncp3, each =
                nrow(trips)))
                pp <- p1[, trips[, 1]] * p2[, trips[, 2]] * p3[, trips[, 3]]
                adjexp <- adjobs
        adjexp[trips] <- a * colSums(GHw * pp)
        ##summing cells with EXPECTED counts <5
        sind <- which(adjexp<5)
        if (length(sind)>0) {
                adjobs <- c(adjobs[-sind], sum(adjobs[sind]))
                adjexp <- c(adjexp[-sind], sum(adjexp[sind]))
                }
        ##all cells w/ EXP < 5 are added to the cell w/the next smallest EXP > 5
        sind <- which(adjexp<5)
        if (length(sind)>0) {
                ind2 <- which.min(adjexp[-sind])
                adjobs <- c(adjobs[-sind][-ind2], sum(adjobs[sind], adjobs[-sind][ind2]))
                adjexp <- c(adjexp[-sind][-ind2], sum(adjexp[sind],adjexp[-sind][ind2]))
                        }
                ##calculating df after collapsing cells for each routine above
                df <- length(adjexp)-1
                resid <- (adjobs - adjexp)^2/adjexp
                        TotalResid <- sum(resid)
                tchisqdf<-TotalResid/df
                        adjouttriplets[i, m]<-tchisqdf
                pvalue<-1-pchisq(TotalResid, df)
                padjouttriplets[i, m]<-pvalue
                cpvalue<-pvalue*4060
                cpadjouttriplets.bon[i,m]<-cpvalue
                }
        i <- i+1
        }

## ADDITIONAL METHOD OF CALCULATING B-H CORRECTED P VALUES ###
### UNADJUSTED SINGLETS ###
for (j in 1:M) {
        bhvalue<-mt.rawp2adjp(poutsinglets[j,], proc="BH")
        adjp<-bhvalue$adjp[order(bhvalue$index),]
        cpoutsinglets.bh[j,]<-adjp[,2]
        }
### ADJUSTED SINGLETS ###
for (j in 1:M) {
        bhvalue<-mt.rawp2adjp(padjoutsinglets[j,], proc="BH")
        adjp<-bhvalue$adjp[order(bhvalue$index),]
        cpadjoutsinglets.bh[j,]<-adjp[,2]
        }
### UNADJUSTED DOUBLETS ###
for (j in 1:M) {
        bhvalue<-mt.rawp2adjp(poutdoublets[j,], proc="BH")
        adjp<-bhvalue$adjp[order(bhvalue$index),]
        cpoutdoublets.bh[j,]<-adjp[,2]
        }
### ADJUSTED DOUBLETS ###
for (j in 1:M) {
        bhvalue<-mt.rawp2adjp(padjoutdoublets[j,], proc="BH")
        adjp<-bhvalue$adjp[order(bhvalue$index),]
        cpadjoutdoublets.bh[j,]<-adjp[,2]
        }
### UNADJUSTED TRIPLETS ###
for (j in 1:M) {
        bhvalue<-mt.rawp2adjp(pouttriplets[j,], proc="BH")
        adjp<-bhvalue$adjp[order(bhvalue$index),]
        cpouttriplets.bh[j,]<-adjp[,2]
        }
### ADJUSTED TRIPLETS ###
for (j in 1:M) {
        bhvalue<-mt.rawp2adjp(padjouttriplets[j,], proc="BH")
        adjp<-bhvalue$adjp[order(bhvalue$index),]
        cpadjouttriplets.bh[j,]<-adjp[,2]
        }

## CALCULATING ALL OUTCOMES FOR USE IN ANALYSES - FOR
UNADJUSTED AND ADJUSTED CHISQ/DF ##
unadjusted<-matrix(0, M, 34) #storing outcomes for unadjusted chisq/df ratios
colnames(unadjusted)<-c("s mean", "s stdv", "d mean", "d stdv", "t mean", "t stdv", "s
percent", "d percent", "t percent", "s pvals", "d pvals", "t pvals", "s bon cpvals", "d bon
cpvals", "t bon cpvals", "s bh cpvals", "d bh cpvals", "t bh cpvals", "sens chisqdf", "spec
chisqdf", "PPV chisqdf", "NPV chisqdf","sens pval", "spec pval", "PPV pval", "NPV
pval", "sens bonpval", "spec bonpval", "PPV bonpval", "NPV bonpval", "sens bhpval",
"spec bhpval", "PPV bhpval", "NPV bhpval")
adjusted<-matrix(0, M, 34) #storing outcomes for adjusted chisq/df ratios
colnames(adjusted)<-c("s mean", "s stdv", "d mean", "d stdv", "t mean", "t stdv", "s
percent", "d percent", "t percent", "s pvals", "d pvals", "t pvals", "s bon cpvals", "d bon
cpvals", "t bon cpvals", "s bh cpvals", "d bh cpvals", "t bh cpvals", "sens chisqdf", "spec
chisqdf", "PPV chisqdf", "NPV chisqdf","sens pval", "spec pval", "PPV pval", "NPV
pval", "sens bonpval", "spec bonpval", "PPV bonpval", "NPV bonpval", "sens bhpval",
"spec bhpval", "PPV bhpval", "NPV bhpval")

## MEANS/STDEVS ##
## singlets ##
for (i in 1:M) {
        mean<-mean(outsinglets[i,])
        stdev<-sqrt(var(outsinglets[i,]))
        mean2<-mean(adjoutsinglets[i,])
        stdev2<-sqrt(var(adjoutsinglets[i,]))
        unadjusted[i,1]<-mean
        adjusted[i,1]<-mean2
        unadjusted[i,2]<-stdev
        adjusted[i,2]<-stdev2
        }
## doublets ##
for (i in 1:M) {
        mean<-mean(outdoublets[i,])
        stdev<-sqrt(var(outdoublets[i,]))
        mean2<-mean(adjoutdoublets[i,])
        stdev2<-sqrt(var(adjoutdoublets[i,]))
        unadjusted[i,3]<-mean
        adjusted[i,3]<-mean2
        unadjusted[i,4]<-stdev
        adjusted[i,4]<-stdev2
        }
## triplets ##
for (i in 1:M) {
        mean<-mean(outtriplets[i,])
        stdev<-sqrt(var(outtriplets[i,]))
        mean2<-mean(adjouttriplets[i,])
        stdev2<-sqrt(var(adjouttriplets[i,]))
        unadjusted[i,5]<-mean
        adjusted[i,5]<-mean2
        unadjusted[i,6]<-stdev
        adjusted[i,6]<-stdev2
        }

### PERCENTAGE OF CHSQ/DF RATIOS > 3 ###
## singlets ##
for (i in 1:M) {
        #freq <- length(which(outsinglets[i,]>3))
        #freq2 <- length(which(adjoutsinglets[i,]>3))
        #unadjusted[i,7]<-freq
        #adjusted[i,7]<-freq2
        percent <- length(which(outsinglets[i,]>3))/.30
        percent2 <- length(which(adjoutsinglets[i,]>3))/.30
        unadjusted[i,7]<-percent
        adjusted[i,7]<-percent2
        }
## doublets ##
for (i in 1:M) {
        #freq <- length(which(outdoublets[i,]>3))
        #freq2 <- length(which(adjoutdoublets[i,]>3))
        #unadjusted[i,8]<-freq
        #adjusted[i,8]<-freq2
        percent <- length(which(outdoublets[i,]>3))/4.35
        percent2 <- length(which(adjoutdoublets[i,]>3))/4.35
        unadjusted[i,8]<-percent
        adjusted[i,8]<-percent2
        }
## triplets ##
for (i in 1:M) {
        #freq <- length(which(outtriplets[i,]>3))
        #freq2<- length(which(adjouttriplets[i,]>3))
        #unadjusted[i,9]<-freq
        #adjusted[i,9]<-freq2
        percent <- length(which(outtriplets[i,]>3))/40.6
        percent2 <- length(which(adjouttriplets[i,]>3))/40.6
        unadjusted[i,9]<-percent
        adjusted[i,9]<-percent2
        }

### PERCENTAGE UNCORRECTED P-VALUES < .05 ###
## singlets ##
for (i in 1:M) {
        #freq <- length(which(poutsinglets[i,]<.05))
        #freq2 <- length(which(padjoutsinglets[i,]<.05))
        #unadjusted[i,10]<-freq
        #adjusted[i,10]<-freq2
        percent <- length(which(poutsinglets[i,]<.05))/.30
        percent2 <- length(which(padjoutsinglets[i,]<.05))/.30
        unadjusted[i,10]<-percent
        adjusted[i,10]<-percent2
        }
## doublets ##
for (i in 1:M) {
        #freq <- length(which(poutdoublets[i,]<.05))
        #freq2 <- length(which(padjoutdoublets[i,]<.05))
        #unadjusted[i,11]<-freq
        #adjusted[i,11]<-freq2
        percent <- length(which(poutdoublets[i,]<.05))/4.35
        percent2 <- length(which(padjoutdoublets[i,]<.05))/4.35
        unadjusted[i,11]<-percent
        adjusted[i,11]<-percent2
        }
## triplets ##
for (i in 1:M) {
        #freq <- length(which(pouttriplets[i,]<.05))
        #freq2<- length(which(padjouttriplets[i,]<.05))
        #unadjusted[i,12]<-freq
        #adjusted[i,12]<-freq2
        percent <- length(which(pouttriplets[i,]<.05))/40.6
        percent2 <- length(which(padjouttriplets[i,]<.05))/40.6
        unadjusted[i,12]<-percent
        adjusted[i,12]<-percent2
        }

### PERCENTAGE BONF CORRECTED P-VALUES < .05 ###
## singlets ##
for (i in 1:M) {
        #freq <- length(which(cpoutsinglets.bon[i,]<.05))
        #freq2 <- length(which(cpadjoutsinglets.bon[i,]<.05))
        #unadjusted[i,13]<-freq
```

```
          #adjusted[i,13]<-freq2
          percent <- length(which(cpoutsinglets.bon[i,]<.05))/.30
          percent2 <- length(which(cpadjoutsinglets.bon[i,]<.05))/.30
          unadjusted[i,13]<-percent
          adjusted[i,13]<-percent2
          }
## doublets ##
for (i in 1:M) {
          #freq <- length(which(cpoutdoublets.bon[i,]<.05))
          #freq2 <- length(which(cpadjoutdoublets.bon[i,]<.05))
          #unadjusted[i,14]<-freq
          #adjusted[i,14]<-freq2
          percent <- length(which(cpoutdoublets.bon[i,]<.05))/4.35
          percent2 <- length(which(cpadjoutdoublets.bon[i,]<.05))/4.35
          unadjusted[i,14]<-percent
          adjusted[i,14]<-percent2
          }
## triplets ##
for (i in 1:M) {
          #freq <- length(which(cpouttriplets.bon[i,]<.05))
          #freq2<- length(which(cpadjouttriplets.bon[i,]<.05))
          #unadjusted[i,15]<-freq
          #adjusted[i,15]<-freq2
          percent <- length(which(cpouttriplets.bon[i,]<.05))/40.6
          percent2 <- length(which(cpadjouttriplets.bon[i,]<.05))/40.6
          unadjusted[i,15]<-percent
          adjusted[i,15]<-percent2
          }

### PERCENTAGE B-H CORRECTED P-VALUES < .05 ###
## singlets ##
for (i in 1:M) {
          #freq <- length(which(cpoutsinglets.bh[i,]<.05))
          #freq2 <- length(which(cpadjoutsinglets.bh[i,]<.05))
          #unadjusted[i,16]<-freq
          #adjusted[i,16]<-freq2
          percent <- length(which(cpoutsinglets.bh[i,]<.05))/.30
          percent2 <- length(which(cpadjoutsinglets.bh[i,]<.05))/.30
          unadjusted[i,16]<-percent
          adjusted[i,16]<-percent2
          }
## doublets ##
for (i in 1:M) {
          #freq <- length(which(cpoutdoublets.bh[i,]<.05))
          #freq2 <- length(which(cpadjoutdoublets.bh[i,]<.05))
          #unadjusted[i,17]<-freq
```

```
        #adjusted[i,17]<-freq2
        percent <- length(which(cpoutdoublets.bh[i,]<.05))/4.35
        percent2 <- length(which(cpadjoutdoublets.bh[i,]<.05))/4.35
        unadjusted[i,17]<-percent
        adjusted[i,17]<-percent2
        }
## triplets ##
for (i in 1:M) {
        #freq <- length(which(cpouttriplets.bh[i,]<.05))
        #freq2<- length(which(cpadjouttriplets.bh[i,]<.05))
        #unadjusted[i,18]<-freq
        #adjusted[i,18]<-freq2
        percent <- length(which(cpouttriplets.bh[i,]<.05))/40.6
        percent2 <- length(which(cpadjouttriplets.bh[i,]<.05))/40.6
        unadjusted[i,18]<-percent
        adjusted[i,18]<-percent2
        }

##NAs for all sens/spec/etc outcomes for Dataset A only
for (i in 1:M) {
        for (j in 19:34) {
        unadjusted[i,j]<-NA
        adjusted[i,j]<-NA
        }
}

write.table(unadjusted, file="newUnadjustedA1n400.txt", sep="\t", quote=FALSE,
row.names=FALSE)
write.table(adjusted, file="newAdjustedA1n400.txt", sep="\t", quote=FALSE,
row.names=FALSE)


###########################################################


####SIMULATION FOR RQ2: Sample size, type of misfit, amount of misfit
#code is separated by TYPE###

####FOR TYPE = MULTIDIMENSIONAL#######
# uses different functions than RQ1 to generate the data, but then chisq routines and all
other calculations of outcomes are the same as above (but sensitivity/specificity will be
added at the end of the entire code document

rmvordlogisE1 <-  function (n, betas) { ###66% items fitting
   # function to simulate random responses
   # based on the Graded Response Model
   # using the additive parameterization
```

```
    ###deleted line setting p since it will always be 20
    ###deleted ncatg line since it will always be 5
    z <-  rnorm(n)
    gammas <-  lapply(betas, function (x) {
        nx <-  length(x)
        cbind(plogis(matrix(x[-nx], n, nx-1, TRUE)- x[nx] * z), 1)
    })
    prs <-  lapply(gammas, function (x) {
        nc <-  ncol(x)
        cbind(x[, 1], x[, 2:nc]-x[, 1:(nc-1)])
    })
    out <-  matrix(0, n, 20) ##replaced p with 20
    for (j in 1:20) { ##same here, replaced p with 20
        for (i in 1:n) {
            out[i, j] <-  sample(5, 1, prob = prs[[j]][i, ])
            ##changed ncatg[j] to 5 since always 5 categories
        }
    }
    out
}
rmvordlogisE2 <-  function (n, betas) { ###33% items misfitting
    # function to simulate random responses
    # based on the Graded Response Model
    # using the additive parameterization
    ###deleted line setting p since it will always be 10
    ###deleted ncatg line since it will always be 5
    z <-  rnorm(n)
    gammas <-  lapply(betas, function (x) {
        nx <-  length(x)
        cbind(plogis(matrix(x[-nx], n, nx-1, TRUE)- x[nx] * z), 1)
    })
    prs <-  lapply(gammas, function (x) {
        nc <-  ncol(x)
        cbind(x[, 1], x[, 2:nc]-x[, 1:(nc-1)])
    })
    out <-  matrix(0, n, 10) ##replaced p with 10
    for (j in 1:10) { ##same here, replaced p with 10
        for (i in 1:n) {
            out[i, j] <-  sample(5, 1, prob = prs[[j]][i, ])
            ##changed ncatg[j] to 5 since always 5 categories
        }
    }
    out
}

rmvordlogisB1 <-  function (n, betas) {  ###90% of items fitting
```

```
    # function to simulate random responses
    # based on the Graded Response Model
    # using the additive parameterization
    ###deleted line setting p since it will always be 27
    ###deleted ncatg line since it will always be 5
    z <-  rnorm(n)
    gammas <-  lapply(betas, function (x) {
        nx <-  length(x)
        cbind(plogis(matrix(x[-nx], n, nx-1, TRUE)- x[nx] * z), 1)
    })
    prs <-  lapply(gammas, function (x) {
        nc <-  ncol(x)
        cbind(x[, 1], x[, 2:nc]-x[, 1:(nc-1)])
    })
    out <-  matrix(0, n, 27) ##replaced p with 27
    for (j in 1:27) { ##same here, replaced p with 27
        for (i in 1:n) {
            out[i, j] <-  sample(5, 1, prob = prs[[j]][i, ])
            ##changed ncatg[j] to 5 since always 5 categories
        }
    }
    out
}

rmvordlogisB2 <-  function (n, betas) { ####10% items misfitting
    # function to simulate random responses
    # based on the Graded Response Model
    # using the additive parameterization
    ###deleted line setting p since it will always be 3
    ###deleted ncatg line since it will always be 5
    z <-  rnorm(n)
    gammas <-  lapply(betas, function (x) {
        nx <-  length(x)
        cbind(plogis(matrix(x[-nx], n, nx-1, TRUE)- x[nx] * z), 1)
    })
    prs <-  lapply(gammas, function (x) {
        nc <-  ncol(x)
        cbind(x[, 1], x[, 2:nc]-x[, 1:(nc-1)])
    })
    out <-  matrix(0, n, 3) ##replaced p with 3
    for (j in 1:3) { ##same here, replaced p with 3
        for (i in 1:n) {
            out[i, j] <-  sample(5, 1, prob = prs[[j]][i, ])
            ##changed ncatg[j] to 5 since always 5 categories
        }
    }
    out
}


##the iprobs function
`iprobs` <-
function (betas, z) {
    n <- length(z)
    gammas <- lapply(betas, function (x) {
        nx <- length(x)
        cbind(plogis(matrix(x[-nx], n, nx - 1, TRUE) - x[nx] * z), 1)
    })
    lapply(gammas, function (x) {
        nc <- ncol(x)
        cbind(x[, 1], x[, 2:nc] - x[, 1:(nc - 1)])
    })
}

# take the betas from dataset A transformed
# as the true betas
#true.betas1<- read.csv('C:/Users/crclar0/Desktop/DatasetBt1.csv', header=T,
row.names=1)
#true.betas2<- read.csv('C:/Users/crclar0/Desktop/DatasetBt2.csv', header=T,
row.names=1)
true.betas1<- read.csv('C:/Users/crclar0/Desktop/DatasetEt1.csv', header=T,
row.names=1)
true.betas2<- read.csv('C:/Users/crclar0/Desktop/DatasetEt2.csv', header=T,
row.names=1)

n <- 400 #start with sample size N=400
M <- 1000 # number of simulations


### SIMULATING DATA, FITTING THE GRM, OBTAINING UNADJUSTED AND
ADJUSTED OUTCOMES ###
ind <- i <- 1
while(i <= M) {
      set.seed(100 + ind) # for reproducible results
            ind <- ind+1
            n<-400  ##change to 1500 and 10000 when needed

            data1 <-  rmvordlogisE1(n, true.betas1)
            data2 <- rmvordlogisE2(n, true.betas2)
            data <- cbind(data1, data2)
### from here, code is the same as in RQ1.
```

```
##########FOR TYPE = DIF##############################

### uses different functions than RQ1 to generate the data, but then chisq routines and all
other calculations of outcomes are the same as above (but sensitivity/specificity will be
added at the end of the entire code document


##uses same rmvordlogis function as in RQ 1, differences don't appear till simulating
response data

true.betas.focus<- read.csv('C:/Users/crclar0/Desktop/DatasetCtfocus.csv', header=T,
row.names=1)
true.betas.reference<- read.csv('C:/Users/crclar0/Desktop/DatasetCtreference.csv',
header=T, row.names=1)

n <- 10000 #change to 1500 and 10000 when needed
M <- 1000 # number of simulations

ind <- i <- 1
while(i <= M) {
    set.seed(100 + ind) # for reproducible results
        ind <- ind+1
        n<-10000
        ## first have to simulate half the cases for each focus and reference group
        ## then have to combine those simulated datasets into one
        data.focus <-  rmvordlogis(n/2, true.betas.focus)
        data.reference <-  rmvordlogis(n/2, true.betas.reference)
        data<-rbind(data.focus, data.reference)
### from here, code is the same as in RQ1.



##########FOR TYPE = MODEL##############################

### uses different functions than RQ1 to generate the data, but then chisq routines and all
other calculations of outcomes are the same as above (but sensitivity/specificity will be
added at the end of the entire code document

###the simulation functions – one for GRM, one for GPCM

rmvordlogisgrm <-  function (n, betas) {
   # function to simulate random responses
   # based on the Graded Response Model
   # using the additive parameterization
```

```
    ###deleted ncatg line since it will always be 5
    z <-  rnorm(n)
    p<-20 ## 27 or 20 depending on amount of misfit################################
    gammas <-  lapply(betas, function (x) {
      nx <-  length(x)
      cbind(plogis(matrix(x[-nx], n, nx-1, TRUE)- x[nx] * z), 1)
    })
    prs <-  lapply(gammas, function (x) {
      nc <-  ncol(x)
      cbind(x[, 1], x[, 2:nc]-x[, 1:(nc-1)])
    })
    out <-  matrix(0, n, p)
    for (j in 1:p) {
      for (i in 1:n) {
        out[i, j] <-  sample(5, 1, prob = prs[[j]][i, ])
        ##changed ncatg[j] to 5 since always 5 categories
      }
    }
    out
}

rmvordlogisgpcm <-  function (n, betas) {
    # function to simulate random responses
    # based on the GPCM for the misfitting items generated from another model
    # NOT using the additive parameterization
    z <-  rnorm(n)
    p<-10 ##3 or 10 depending on D or G #############################
    prs <- lapply(crf.GPCM(betas, z, IRT=TRUE), t)
    out <-  matrix(0, n, p)
    for (j in 1:p) {
      for (i in 1:n) {
        out[i, j] <-  sample(5, 1, prob = prs[[j]][i, ])
        ##changed ncatg[j] to 5 since always 5 categories
      }
    }
    out
}

crf.GPCM<- function (betas, z, IRT.param = TRUE, log = FALSE, eps =
.Machine$double.eps^(1/2))
{
   lapply(linpred.GPCM(betas, z, IRT.param), function(x) {
     num <- exp(apply(x, 2, cumsum))
     if (!is.matrix(num))
       num <- t(num)
     den <- 1 + colSums(num)
```

```
        out <- rbind(1/den, num/rep(den, each = nrow(x)))
        if (any(ind <- out == 1))
            out[ind] <- 1 - eps
        if (any(ind <- out == 0))
            out[ind] <- eps
        if (log)
            out <- log(out)
        out
    })
}

linpred.GPCM<-function (betas, z, IRT.param = TRUE)
{
    lapply(betas, function(x) {
        nx <- length(x)
        if (IRT.param)
            t(x[nx] * outer(z, x[-nx], "-"))
        else outer(x[-nx], x[nx] * z, "+")
    })
}

##the iprobs function
`iprobs` <-
function (betas, z) {
    n <- length(z)
    gammas <- lapply(betas, function (x) {
        nx <- length(x)
        cbind(plogis(matrix(x[-nx], n, nx - 1, TRUE) - x[nx] * z), 1)
    })
    lapply(gammas, function (x) {
        nc <- ncol(x)
        cbind(x[, 1], x[, 2:nc] - x[, 1:(nc - 1)])
    })
}

grm1<-function (data, constrained = FALSE, IRT.param = TRUE, Hessian = FALSE,
    start.val = NULL, na.action = NULL, control = list())
{
    cl <- match.call()
    if ((!is.data.frame(data) & !is.matrix(data)) || ncol(data) ==
        1)
        stop("'data' must be either a numeric matrix or a data.frame, with at least two
columns.\n")
    X <- data.matrix(data)
    if (!is.null(na.action))
        X <- na.action(X)
    X <- apply(X, 2, function(x) {
        y <- x[!is.na(x)]
        if (any(y == 0))
            x + 1
        else x
    })
    colnamsX <- colnames(X)
    dimnames(X) <- NULL
    #ncatg <- apply(X, 2, function(x) if (any(is.na(x)))
        #length(unique(x)) - 1
    #else length(unique(x)))
    ncatg<-rep(5, 30)
    n <- nrow(X)
    p <- ncol(X)
    pats <- apply(X, 1, paste, collapse = "/")
    freqs <- table(pats)
    nfreqs <- length(freqs)
    obs <- as.vector(freqs)
    X <- unlist(strsplit(cbind(names(freqs)), "/"))
    X[X == "NA"] <- as.character(NA)
    X <- matrix(as.numeric(X), nfreqs, p, TRUE)
    con <- list(iter.qN = 150, GHk = 21, method = "BFGS", verbose =
getOption("verbose"),
        digits.abbrv = 6)
    con[names(control)] <- control
    GH <- GHpoints(data ~ z1, con$GHk)
    Z <- GH$x[, 2]
    GHw <- GH$w
    ind1 <- if (constrained)
        c(1, cumsum(ncatg[-p] - 1) + 1)
    else c(1, cumsum(ncatg[-p]) + 1)
    ind2 <- if (constrained)
        cumsum(ncatg - 1)
    else cumsum(ncatg)
    betas <- start.val.grm(start.val, X, obs, constrained, ncatg)
    environment(loglikgrm) <- environment(scoregrm) <- environment()
    old <- options(warn = (-1))
    on.exit(options(old))
    res.qN <- optim(unlist(betas), fn = loglikgrm, gr = scoregrm,
        method = con$method, hessian = Hessian, control = list(maxit = con$iter.qN,
            trace = as.numeric(con$verbose)), constrained = constrained)
    betas <- betas.grm(res.qN$par, constrained, ind1, ind2, p)
    names(betas) <- if (!is.null(colnamsX))
        colnamsX
    else paste("Item", 1:p)
    betas <- lapply(betas, function(x) {
```

```
    names(x) <- c(paste("beta.", seq(1, length(x) - 1), sep = ""),
      "beta")
    x
  })
  max.sc <- max(abs(scoregrm(res.qN$par, constrained)), na.rm = TRUE)
  fit <- list(coefficients = betas, log.Lik = -res.qN$value,
    convergence = res.qN$conv, hessian = res.qN$hessian,
    counts = res.qN$counts, patterns = list(X = X, obs = obs),
    GH = list(Z = Z, GHw = GHw), max.sc = max.sc, constrained = constrained,
    IRT.param = IRT.param, X = data, control = con, na.action = na.action,
    call = cl)
  class(fit) <- "grm"
  fit
}

##needed source code for using GPCM option
source("C:/Users/crclar0/Desktop/ltm/R/anova.gpcm.R")
source("C:/Users/crclar0/Desktop/ltm/R/anova.grm.R")
source("C:/Users/crclar0/Desktop/ltm/R/betas.gpcm.R")
source("C:/Users/crclar0/Desktop/ltm/R/betas.grm.R")
source("C:/Users/crclar0/Desktop/ltm/R/biserial.cor.R")
source("C:/Users/crclar0/Desktop/ltm/R/cd.tpm.R")
source("C:/Users/crclar0/Desktop/ltm/R/cd.vec.R")
source("C:/Users/crclar0/Desktop/ltm/R/chisq.irt.R")
source("C:/Users/crclar0/Desktop/ltm/R/coef.gpcm.R")
source("C:/Users/crclar0/Desktop/ltm/R/coef.grm.R")
source("C:/Users/crclar0/Desktop/ltm/R/coef.tpm.R")
source("C:/Users/crclar0/Desktop/ltm/R/cprobs.R")
source("C:/Users/crclar0/Desktop/ltm/R/crf.GPCM.R")
source("C:/Users/crclar0/Desktop/ltm/R/crf.GPCM2.R")
source("C:/Users/crclar0/Desktop/ltm/R/cumprobs.R")
source("C:/Users/crclar0/Desktop/ltm/R/descript.R")
source("C:/Users/crclar0/Desktop/ltm/R/EM.R")
source("C:/Users/crclar0/Desktop/ltm/R/fd.vec.R")
source("C:/Users/crclar0/Desktop/ltm/R/fitted.gpcm.R")
source("C:/Users/crclar0/Desktop/ltm/R/fitted.grm.R")
source("C:/Users/crclar0/Desktop/ltm/R/fscores.g.R")
source("C:/Users/crclar0/Desktop/ltm/R/fscores.gp.R")
source("C:/Users/crclar0/Desktop/ltm/R/fscores.l.R")
source("C:/Users/crclar0/Desktop/ltm/R/fscores.r.R")
source("C:/Users/crclar0/Desktop/ltm/R/fscores.t.R")
source("C:/Users/crclar0/Desktop/ltm/R/gauher.R")
source("C:/Users/crclar0/Desktop/ltm/R/GHpoints.R")
source("C:/Users/crclar0/Desktop/ltm/R/GoF.gpcm.R")
source("C:/Users/crclar0/Desktop/ltm/R/infoGPCM.R")
source("C:/Users/crclar0/Desktop/ltm/R/infoprobs.R")
source("C:/Users/crclar0/Desktop/ltm/R/information.R")
source("C:/Users/crclar0/Desktop/ltm/R/IRT.parm.grm.R")
source("C:/Users/crclar0/Desktop/ltm/R/IRT.parm.R")
source("C:/Users/crclar0/Desktop/ltm/R/item.fit.R")
source("C:/Users/crclar0/Desktop/ltm/R/jacobian.R")
source("C:/Users/crclar0/Desktop/ltm/R/linpred.GPCM.R")
source("C:/Users/crclar0/Desktop/ltm/R/logLik.gpcm.R")
source("C:/Users/crclar0/Desktop/ltm/R/logLik.grm.R")
source("C:/Users/crclar0/Desktop/ltm/R/loglikgpcm.R")
source("C:/Users/crclar0/Desktop/ltm/R/loglikgrm.R")
source("C:/Users/crclar0/Desktop/ltm/R/margins.gpcm.R")
source("C:/Users/crclar0/Desktop/ltm/R/margins.grm.R")
source("C:/Users/crclar0/Desktop/ltm/R/margins.R")
source("C:/Users/crclar0/Desktop/ltm/R/matArrays.R")
source("C:/Users/crclar0/Desktop/ltm/R/matches.R")
source("C:/Users/crclar0/Desktop/ltm/R/matMeans.R")
source("C:/Users/crclar0/Desktop/ltm/R/matSums.R")
source("C:/Users/crclar0/Desktop/ltm/R/observedFreqs.R")
source("C:/Users/crclar0/Desktop/ltm/R/start.val.grm.R")
source("C:/Users/crclar0/Desktop/ltm/R/scoregrm.R")

###new StartVals function code:
start.val.grm<-function(start.val, data, weight, constrained, ncatg) {
n <- nrow(data)
  p <- ncol(data)
  computeStartVals <- function(start.val) {
    ind <- if (!is.null(start.val)) {
      if (!is.list(start.val) && start.val == "random")
        return(list(compute = TRUE, random = TRUE))
      if (!is.list(start.val) && length(start.val) != p) {
        warning("'start.val' not of proper type; random starting values are used
instead.\n")
        TRUE
      }
      else if (!all(ncatg == sapply(start.val, length))) {
        warning("number of parameter in 'start.val' differ from the number of levels in
'data'; random starting values are used instead.\n")
        TRUE
      }
      else FALSE
    }
    else TRUE
    list(compute = ind, random = FALSE)
  }
  comp <- computeStartVals(start.val)
  if (comp$compute) {
    res <- vector("list", p)
    z <- if (comp$random)
      rnorm(n)
    else seq(-3, 3, length = n)[order(rowSums(data, na.rm = TRUE))]
    for (i in 1:p) {
      y <- data[, i]
      na.ind <- !is.na(y)
      y. <- y[na.ind]
      z. <- z[na.ind]
      weight. <- weight[na.ind]
      lev <- 5
      q <- lev - 1
      q1 <- lev%/%2
      y1 <- (y. > q1)
      fit <- glm.fit(cbind(1, z.), y1, weight., family = binomial())
      coefs <- fit$coefficients
      spacing <- qlogis((1:q)/(q + 1))
      thets <- -coefs[1] + spacing - spacing[q1]
      out <- c(thets[1], log(diff(thets)), coefs[-1])
      names(out) <- NULL
      res[[i]] <- out
    }
    if (constrained)
      res[seq(1, p - 1)] <- lapply(res[seq(1, p - 1)],
        function(x) x[-length(x)])
    res
  }
  else {
    lapply(start.val, function(x) {
      nx <- length(x)
      c(x[1], log(diff(x[-nx])), x[nx])
    })
  }
}


##now reading in true betas
true.betas1<- read.csv('C:/Users/crclar0/Desktop/DatasetGt1.csv', header=T,
row.names=1)
true.betas2<- read.csv('C:/Users/crclar0/Desktop/DatasetGt3.csv', header=T,
row.names=1) ### usual IRT parameters

n <- 400 #change to 1500 and 10000 when needed
M <- 1000 # number of simulations
```

```
###NOTE: MUST USE GRM1 FUNCTION INSTEAD OF GRM FOR CONDITIONS
WITH TYPE = MODEL

### SIMULATING DATA, FITTING THE GRM (GRM1), OBTAINING
UNADJUSTED AND ADJUSTED OUTCOMES ###
ind <- i <- 1
while(i <= M) {
    set.seed(100 + ind) # for reproducible results
        ind <- ind+1
        n<-400

        data1 <- rmvordlogisgrm(n, true.betas1)
        data2 <- rmvordlogisgpcm(n, true.betas2)
        #data2 <- rmvordlogis(n, true.betas2, IRT=FALSE, model="gpcm")
        data <- cbind(data1, data2)
##From here, same code as in RQ1, except fit GRM with grm1 instead of grm.

###########SENSITIVITY AND SPECIFICITY FOR ALL RQ2 CONDITIONS##

### SENSITIVITY, SPECIFICITY, PPV, and NPV for ALL 8 METHODS for
SINGLETS ###
## (1) ChiSq/df > 3; (2) adjChiSq/df > 3; (3) pval < .05; (4) adjpval< .05; (5) bonf pval <
.05; (6) adj bonf pval < .05; (7) bh pval < .05; (8) adj bh pval < .05 ##
for (i in 1:M) {
        true<-rep(c(0,1), c(27,3)) ##for datasets B,C,D
        #true<-rep(c(0,1), c(20,10)) ##for datasets E,F,G
        pred.misfit.1<-ifelse(outsinglets[i,] > 3, 1, 0)
        conf.mat.1<-table(factor(true, levels=0:1), factor(pred.misfit.1, levels=0:1))
        pred.misfit.2<-ifelse(adjoutsinglets[i,] > 3, 1, 0)
        conf.mat.2<-table(factor(true, levels=0:1), factor(pred.misfit.2, levels=0:1))

        sens1.1<-conf.mat.1[2,2]/(conf.mat.1[2,2] + conf.mat.1[2,1])
        spec1.1<-conf.mat.1[1,1]/(conf.mat.1[1,1] + conf.mat.1[1,2])
        ppv1.1<-conf.mat.1[2,2]/(conf.mat.1[2,2] + conf.mat.1[1,2])
        npv1.1<-conf.mat.1[1,1]/(conf.mat.1[1,1] + conf.mat.1[2,1])
        sens1.2<-conf.mat.2[2,2]/(conf.mat.2[2,2] + conf.mat.2[2,1])
        spec1.2<-conf.mat.2[1,1]/(conf.mat.2[1,1] + conf.mat.2[1,2])
        ppv1.2<-conf.mat.2[2,2]/(conf.mat.2[2,2] + conf.mat.2[1,2])
        npv1.2<-conf.mat.2[1,1]/(conf.mat.2[1,1] + conf.mat.2[2,1])

        unadjusted[i,19]<-sens1.1
        adjusted[i,19]<-sens1.2
        unadjusted[i,20]<-spec1.1
        adjusted[i,20]<-spec1.2
        unadjusted[i,21]<-ppv1.1
        adjusted[i,21]<-ppv1.2
```

```
unadjusted[i,22]<-npv1.1
adjusted[i,22]<-npv1.2

pred.misfit.1<-ifelse(poutsinglets[i,] < .05, 1, 0)
conf.mat.1<-table(factor(true, levels=0:1), factor(pred.misfit.1, levels=0:1))
pred.misfit.2<-ifelse(padjoutsinglets[i,] < .05, 1, 0)
conf.mat.2<-table(factor(true, levels=0:1), factor(pred.misfit.2, levels=0:1))

sens1.1<-conf.mat.1[2,2]/(conf.mat.1[2,2] + conf.mat.1[2,1])
spec1.1<-conf.mat.1[1,1]/(conf.mat.1[1,1] + conf.mat.1[1,2])
ppv1.1<-conf.mat.1[2,2]/(conf.mat.1[2,2] + conf.mat.1[1,2])
npv1.1<-conf.mat.1[1,1]/(conf.mat.1[1,1] + conf.mat.1[2,1])
sens1.2<-conf.mat.2[2,2]/(conf.mat.2[2,2] + conf.mat.2[2,1])
spec1.2<-conf.mat.2[1,1]/(conf.mat.2[1,1] + conf.mat.2[1,2])
ppv1.2<-conf.mat.2[2,2]/(conf.mat.2[2,2] + conf.mat.2[1,2])
npv1.2<-conf.mat.2[1,1]/(conf.mat.2[1,1] + conf.mat.2[2,1])

unadjusted[i,23]<-sens1.1
adjusted[i,23]<-sens1.2
unadjusted[i,24]<-spec1.1
adjusted[i,24]<-spec1.2
unadjusted[i,25]<-ppv1.1
adjusted[i,25]<-ppv1.2
unadjusted[i,26]<-npv1.1
adjusted[i,26]<-npv1.2

pred.misfit.1<-ifelse(cpoutsinglets.bon[i,] < .05, 1, 0)
conf.mat.1<-table(factor(true, levels=0:1), factor(pred.misfit.1, levels=0:1))
pred.misfit.2<-ifelse(cpadjoutsinglets.bon[i,] < .05, 1, 0)
conf.mat.2<-table(factor(true, levels=0:1), factor(pred.misfit.2, levels=0:1))

sens1.1<-conf.mat.1[2,2]/(conf.mat.1[2,2] + conf.mat.1[2,1])
spec1.1<-conf.mat.1[1,1]/(conf.mat.1[1,1] + conf.mat.1[1,2])
ppv1.1<-conf.mat.1[2,2]/(conf.mat.1[2,2] + conf.mat.1[1,2])
npv1.1<-conf.mat.1[1,1]/(conf.mat.1[1,1] + conf.mat.1[2,1])
sens1.2<-conf.mat.2[2,2]/(conf.mat.2[2,2] + conf.mat.2[2,1])
spec1.2<-conf.mat.2[1,1]/(conf.mat.2[1,1] + conf.mat.2[1,2])
ppv1.2<-conf.mat.2[2,2]/(conf.mat.2[2,2] + conf.mat.2[1,2])
npv1.2<-conf.mat.2[1,1]/(conf.mat.2[1,1] + conf.mat.2[2,1])

unadjusted[i,27]<-sens1.1
adjusted[i,27]<-sens1.2
unadjusted[i,28]<-spec1.1
adjusted[i,28]<-spec1.2
unadjusted[i,29]<-ppv1.1
adjusted[i,29]<-ppv1.2
```

```
unadjusted[i,30]<-npv1.1
adjusted[i,30]<-npv1.2

pred.misfit.1<-ifelse(cpoutsinglets.bh[i,] < .05, 1, 0)
conf.mat.1<-table(factor(true, levels=0:1), factor(pred.misfit.1, levels=0:1))
pred.misfit.2<-ifelse(cpadjoutsinglets.bh[i,] < .05, 1, 0)
conf.mat.2<-table(factor(true, levels=0:1), factor(pred.misfit.2, levels=0:1))

sens1.1<-conf.mat.1[2,2]/(conf.mat.1[2,2] + conf.mat.1[2,1])
spec1.1<-conf.mat.1[1,1]/(conf.mat.1[1,1] + conf.mat.1[1,2])
ppv1.1<-conf.mat.1[2,2]/(conf.mat.1[2,2] + conf.mat.1[1,2])
npv1.1<-conf.mat.1[1,1]/(conf.mat.1[1,1] + conf.mat.1[2,1])
sens1.2<-conf.mat.2[2,2]/(conf.mat.2[2,2] + conf.mat.2[2,1])
spec1.2<-conf.mat.2[1,1]/(conf.mat.2[1,1] + conf.mat.2[1,2])
ppv1.2<-conf.mat.2[2,2]/(conf.mat.2[2,2] + conf.mat.2[1,2])
npv1.2<-conf.mat.2[1,1]/(conf.mat.2[1,1] + conf.mat.2[2,1])

unadjusted[i,31]<-sens1.1
adjusted[i,31]<-sens1.2
unadjusted[i,32]<-spec1.1
adjusted[i,32]<-spec1.2
unadjusted[i,33]<-ppv1.1
adjusted[i,33]<-ppv1.2
unadjusted[i,34]<-npv1.1
adjusted[i,34]<-npv1.2
}
```

APPENDIX C

Parameters (“True Betas”) for 30 Unidimensional Items,
Free of Differential Item Functioning, from the Graded Response Model

Item	a	b_1	b_2	b_3	b_4
1	2.06	–0.78	–0.17	0.25	1.01
2	1.27	–1.80	–0.62	0.19	1.63
3	1.82	–1.32	0.04	0.53	1.10
4	1.66	–1.12	–0.81	1.00	1.54
5	1.73	–2.46	–0.50	0.76	1.94
6	1.78	–1.30	–0.72	0.90	1.46
7	1.67	–0.57	1.37	1.54	1.69
8	1.62	–1.77	–0.06	0.91	1.30
9	2.09	–1.83	0.22	0.83	1.22
10	1.31	–0.45	–0.08	1.14	1.60
11	1.56	–1.85	–0.80	–0.17	0.50
12	1.23	–1.45	–0.16	1.04	2.19
13	1.91	–1.51	0.24	0.46	0.71
14	1.55	–1.25	–0.70	–0.17	1.28
15	1.47	–0.95	–0.08	1.55	1.99
16	1.95	–1.68	–0.93	0.21	1.20
17	2.11	–1.96	–0.26	0.41	1.12
18	1.45	–2.18	–0.81	0.08	0.75
19	1.78	–1.68	0.10	0.87	1.53
20	1.74	–0.40	–0.18	0.19	1.57
21	1.54	–1.97	0.02	0.20	0.68
22	1.83	–0.60	0.95	1.07	2.34
23	2.10	–0.94	0.00	1.43	1.49
24	2.09	–1.51	0.65	0.84	1.96
25	1.91	–0.14	0.21	1.61	1.88
26	1.44	–1.86	0.25	1.26	1.40
27	1.88	–0.59	–0.27	0.27	1.84
28	1.94	–1.15	–0.01	1.30	2.72
29	1.81	–2.42	0.26	1.46	1.92
30	1.29	–1.00	0.11	0.81	2.09

Note. For Research Question 1, all 30 items were used. See Appendices D, E, and F for details regarding substitution of “true betas” (i.e., item parameters) for Research Question 2. Item parameters were taken from Bolt (2002).

APPENDIX D

Parameters ("True Betas") for 10 Multidimensional Items,
Free of Differential Item Functioning, from the Graded Response Model

Item	*a*	b_1	b_2	b_3	b_4
21	0.95	-4.26	-2.90	-1.25	2.01
22	1.48	-2.45	-1.44	-0.60	1.45
23	1.46	-2.07	-1.27	0.16	2.11
24	1.49	-1.75	-0.76	0.13	2.02
25	1.38	-2.19	-1.27	-0.35	1.52
26	1.35	-2.88	-1.97	-0.51	1.87
27	0.96	-3.77	-2.23	-1.27	1.34
28	1.32	-3.24	-2.29	-0.49	1.93
29	1.08	-3.28	-2.09	0.49	3.09
30	2.00	-1.57	-0.75	-0.13	1.68

Note. For Research Question 2, when type of item misfit was due to multidimensionality, "true betas" depicted in Appendix C were replaced with the above parameters as follows: In conditions with 10% misfitting items due to multidimensionality, items 28-30 from this table replaced items 28-30 from Appendix C. In conditions with 33% misfitting items due to multidimensionality, items 21-30 from this table replaced items 21-30 from Appendix C. Item parameters were taken from Lautenschlager, Meade, & Kim (2006).

APPENDIX E

Parameters (“True Betas”) for 10 Unidimensional Items Exhibiting Differential Item Functioning from the Graded Response Model

Item	*a*	b_1	b_2	b_3	b_4
Focus Group Items					
21	1.72	–0.54	0.36	0.80	1.51
22	1.72	–0.06	0.58	0.93	1.57
23	1.71	−1.02	0.22	0.97	1.60
24	1.70	−1.02	0.22	1.29	1.77
25	1.71	–0.48	0.37	0.63	1.44
26	1.70	–0.06	0.58	0.70	1.47
27	1.70	–0.53	0.71	1.21	1.99
28	1.69	–0.03	1.21	1.72	2.51
29	1.16	−1.01	0.45	1.14	2.00
30	2.23	−1.02	0.11	0.51	1.26
Reference Group Items					
21	1.71	−1.02	0.21	0.71	1.48
22	1.71	−1.02	0.21	0.71	1.48
23	1.71	−1.02	0.21	0.71	1.48
24	1.71	−1.02	0.21	0.71	1.48
25	1.71	−1.02	0.21	0.71	1.48
26	1.71	−1.02	0.21	0.71	1.48
27	1.71	−1.02	0.21	0.71	1.48
28	1.71	−1.02	0.21	0.71	1.48
29	1.71	−1.02	0.21	0.71	1.48
30	1.71	−1.02	0.21	0.71	1.48

***Note*.** For Research Question 2, when type of item misfit was due to differential item functioning (DIF), “true betas” depicted in Appendix C were replaced with the above parameters as follows: In conditions with 10% misfitting items due to DIF, items 28-30 from this table replaced items 28-30 from Appendix C. In conditions with 33% misfitting items due to DIF, items 21-30 from this table replaced items 21-30 from Appendix C. Simulated samples for these conditions were randomly split and assigned to focus and reference groups, then paired with the appropriate item parameters. Item parameters were taken from Bolt (2002).

APPENDIX F

Parameters for 10 Unidimensional Items, Free of Differential Item Functioning, Generated from Muraki's (1992) Generalized Partial Credit Model

Item	a	b_1	b_2	b_3	b_4
21	0.73	–2.12	2.62	–0.73	–1.02
22	1.13	–0.42	2.76	–0.58	2.28
23	1.30	–0.53	–0.26	3.51	–0.68
24	1.42	–1.52	1.85	–0.11	1.84
25	1.03	1.27	–0.88	2.95	0.44
26	0.81	–1.85	0.80	3.36	–1.41
27	1.00	0.84	–0.61	–0.38	1.78
28	1.45	–0.90	–0.04	1.30	2.69
29	1.30	–2.54	0.50	2.04	1.15
30	0.66	–0.05	0.74	0.20	1.56

Note. For Research Question 2, when type of item misfit was due to generation from a competing model, "true betas" depicted in Appendix C were replaced with the above parameters as follows: In conditions with 10% misfitting items due to generation from a different model, items 28-30 from this table replaced items 28-30 from Appendix C. In conditions with 33% misfitting items due to generation from a different model, items 21-30 from this table replaced items 21-30 from Appendix C. Item parameters were taken from Bolt (2002).

CURRICULUM VITAE

Christina R. Studts, Ph.D., L.C.S.W.

BIOGRAPHICAL INFORMATION

Work Address &
Contact Information:
101 Medical Behavioral Science Building
Lexington, Kentucky 40536-0086
Office: 859.323.1788
Cell: 859.523.6976
Fax: 859.523.5350
tina.studts@uky.edu

Home Address &
Contact Information:
1301 Cooper Drive
Lexington, Kentucky 40502
Cell: 859.523.6976

Birth Date:	April 16, 1971
Hometown:	Los Alamos, New Mexico
Citizenship:	U.S.A.
License:	L.C.S.W., Kentucky (#1424, since 5/2000)

EMPLOYMENT

07/08-present Assistant Professor, Research Title Series, Department of Behavioral Science, University of Kentucky College of Medicine

EDUCATION

04/08 Ph.D., Social Work, Kent School of Social Work, University of Louisville
Dissertation: *Improving screening for externalizing behavior problems in very young children: Applications of item response theory to evaluate instruments in pediatric primary care*

05/12 (exp.) M.S., Biostatistics, Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville (Anticipated completion: 05/12)
Thesis: *Utility of a goodness-of-fit index for the Graded Response Model with small sample sizes: A Monte Carlo investigation*

05/97 M.S.W., University of Kentucky

05/93 B.A., Psychology (With Honors), University of Notre Dame

GRANT SUPPORT

Current Support

1UL1RR033173-01 (Kern, PI) 06/1/2011 – 02/29/2016
NIH
Kentucky Center for Clinical and Translational Science
The University of Kentucky Center for Clinical and Translational Science (CCTS) is the "academic home" for the discipline of clinical and translational science, dedicated to growing the clinical and translational science research teams of the future, to providing the infrastructure needed to foster collaborations between basic and clinical scientists to facilitate research translation, and to enhancing outreach pathways to confront chronic health issues in rural Appalachia. The CCTS co-localizes clinical research facilities and core research support and infrastructure in areas such as pilot project funding, education and career development, biostatistics and research design, regulatory support and research integrity, biomedical informatics, development of novel therapeutics and medical devices, and translational technology development. Through focused community engagement efforts, the CCTS facilitates community-based and practice-based research.

U54 CA153604-02 (Dignan, PI) 09/1/2010 - 08/31/2015
NCI
Appalachian Community Cancer Network
The major goal of this project is to work towards a reduction in cancer health disparities in Appalachia by conducting community-based education, research and training activities in rural areas of six states. The project will also include a faith-based intervention to reduce obesity among rural Appalachian church members by improving diet and increasing physical activity.

SM059561-01 (Fulcher, PD; Studts, PI) 10/01/2009 – 09/30/2013
SAMHSA/Center for Mental Health Services
Pennyroyal Center Comprehensive Health Care
This longitudinal study will investigate the effectiveness of implementing primary health care services for seriously mentally ill adults within a community mental health care setting.
Role: *Principal Investigator/Evaluator*

R01 DK081324-01 (Schoenberg, PI) 09/01/2008 - 06/30/2013
NIH/National Institute of Diabetes and Digestive and Kidney Diseases
An Intergenerational Intervention to Reduce Appalachian Health Disparities
This community-based participatory research project is a group-randomized clinical trial investigating the effectiveness of a multifaceted faith-based lay health advisor intervention targeting energy balance outcomes in Appalachian families.
Role: *Co-Investigator*

R24 MD002757 (Schoenberg, PI) 05/22/2008 - 01/31/2013
NIH/National Center on Minority Health and Health Disparities
Faith Moves Mountains: A CBPR Appalachian Wellness & Cancer Prevention Program
This community-based participatory research project is a group-randomized clinical trial investigating the effectiveness of a multifaceted faith-based lay health advisor intervention targeting smoking cessation and cancer screening in Appalachian participants.
Role: *Co-Investigator*

Past Support

R03 CA139876-01 (Floyd, PI) 09/01/2009 - 07/31/2011
NIH/National Cancer Institute
A Population-based Controlled Investigation of QOL and Health Behaviors among Survivors of Cancer Diagnosed during Young Adulthood
This longitudinal, mixed-methods study will identify unique needs with regard to mental health, health behavior and quality of life outcomes in survivors of cancer diagnosed during young adulthood, an understudied group of cancer survivors. It will also set the stage for future targeted intervention efforts in this population.
Role: *Biostatistician*

R01 CA113932 (Dignan, PI) 05/01/2005 - 03/31/2011
NIH/National Cancer Institute
Increasing Colorectal Cancer Screening in Rural Kentucky
This community-based, participatory study is investigating the effects of a multicomponent intervention delivered to primary health care practices in rural Kentucky on health care providers' knowledge, attitudes, and behaviors re: screening for colorectal cancer.

R01 CA108696 (Schoenberg, PI) 09/01/2004 - 01/31/2011
NIH/National Cancer Institute and National Institute on Aging
An Appalachian Cervical Cancer Prevention Project
This community-based participatory research project is a randomized controlled trial focused on the development, administration, and evaluation of a cervical cancer prevention project involving faith communities and lay health advisers in rural Appalachia.
Role: *Co-Investigator*

1 R36 HS016940-01 (C.R. Studts, PI) Awarded 2008; declined
USDHHS/Agency for Healthcare Research and Quality Dissertation Grant
Improving Screening for Externalizing Behavior Problems in Very Young Children: Applications of Item Response Theory to Evaluate Instruments in Pediatric Primary Care
This study investigated the performance of items in two commonly used pediatric behavioral screening instruments, with special attention to differences between groups categorized by sex, race, and socioeconomic status.
Role: *Principal Investigator*

KLCRP403-JLS-03 (J.L. Studts, PI) 07/2003 - 06/2008
Kentucky Lung Cancer Research Program
Behavioral, Cognitive, and Affective Responses to Lung Cancer Screening
This study examined a range of important positive and negative sequelae of participation in a randomized controlled clinical trial of lung cancer screening.
Role: *Research Assistant*

KCHFS (Barbee, PI) 07/2003 - 08/2004
Kentucky Department of Health and Human Services
The Manualization and Replication of the Comprehensive Assessment and Training Services
Role: *Development team member, clinical social worker/supervisor*

PEER-REVIEWED PUBLICATIONS

Studts, C. R., Tarasenko, Y., Schoenberg, N. E., Shelton, B. J., Hatcher-Keller, J., & Dignan, M. B. (in press). A community-based faith-placed intervention to reduce cervical cancer in Appalachia. *Preventive Medicine.*

Knudsen, H. K., **Studts, C. R.**, & Studts, J. L. (2012). The implementation of smoking cessation counseling in substance abuse treatment. *Journal of Behavioral Health Services & Research, 39*(1), 28-41.

Swanson, M., **Studts, C. R.**, Bardach, S. H., Bersamin, A., & Schoenberg, N. E. (2011). Intergenerational energy balance interventions: A systematic literature review. *Health Education & Behavior, 38*(2), 171-197.

Hatcher, J., **Studts, C. R.**, Dignan, M. B., Turner, L. M., & Schoenberg, N. E. (2011). Predictors of cervical cancer screening for rarely or never screened rural Appalachian women. *Journal of Health Care for the Poor and Underserved, 22*(1), 176-193.

Studts, C. R. (2009). Improving screening for externalizing behavior problems among very young children in primary care: Applications of item response theory. *Proceedings of the Ohio State University 21st National Symposium on*

Doctoral Research in Social Work. Retrieved from **http://hdl.handle.net/1811/36781**.

Studts, C. R., Stone, R., & Barber, G. M. (2006). Predictors of access to health-care services among groups of TANF recipients in Kentucky. *Social Service Review, 80*(3), 527-548.

Studts, J. L., Ghate, S. R., Gill, J. L., **Studts, C. R.**, Barnes, C. N., LaJoie, A. J., Andrykowski, M. A., & LaRocca, R. V. (2006). Validity of self-reported smoking status among participants in a lung cancer screening trial. *Cancer Epidemiology, Biomarkers and Prevention, 15*(10), 1825-1828. ***Selected as a featured article among the 43 published in this issue.

MANUSCRIPTS UNDER REVIEW

Shelton, B. J., Schoenberg, N. E., Dignan, M. B., Dollarhide, K., Hatcher, J., **Studts, C. R.,** & van Meter, E. (2012). Trials and tribulations of sampling faith-based groups for a cancer screening intervention: Lessons learned from the Faith Moves Mountains project.

MANUSCRIPTS IN PREPARATION

Studts, C. R., & Brock, G. N. (2012). Utility of a goodness-of-fit index for the graded response model with small sample sizes: A Monte Carlo investigation.

Studts, C. R., & van Zyl, M. A. (2012). Item response theory calibration of items measuring externalizing behavior problems in very young children.

Studts, C. R., & van Zyl, M. A. (2012). Differential item functioning of items measuring externalizing behavior problems in very young children.

Studts, C. R., & van Zyl, M. A. (2012). Development of a brief screening instrument for externalizing behavior problems in very young children: An application of item response theory.

Studts, C. R., Tarasenko, Y., & Schoenberg, N. E. (2012). Barriers to cervical cancer prevention among rural Appalachian women.

Studts, C. R., Dignan, M. B., Havens, J. R., Duvall, J. L., Deskins, S., & Schoenberg, N. E. (2012). The development of a cultural identity questionnaire: The Appalachian Identity Project.

Studts, J. L., **Studts, C. R.**, Barnes, C. N., LaJoie, A. S., Andrykowski, M. A., & LaRocca, R. (2012). Predictors of adherence to screening protocol among participants in a lung cancer screening trial.

Studts, J. L., **Studts, C. R.**, & Matera, E. L (2012). Measuring numeracy: An item response theory analysis of the Numeracy Questionnaire.

Studts, J. L., **Studts, C. R.**, LaJoie, A. S., Barnes, C. N., Andrykowski, M. A., & LaRocca, R. (2012). Predictors of change in self-reported smoking status among participants in a lung cancer screening trial.

PRESENTATIONS

Howell, B. M., Schoenberg, N. E., Strath, S., & **Studts, C. R.** (2012). *Measurement of physical activity among rural Appalachian residents.* Paper to be presented at the 35th annual Appalachian Studies Association conference, Indiana, PA.

*Arrowood, A., Young, T. L., Quigley, J. M., Woods, S. H., & **Studts, C. R.** (2010). *Parent and physician readiness for the implementation of an integrated pediatric primary care clinic.* Poster presented at the 12th annual conference of the Collaborative Family Healthcare Association, Louisville, KY.
*Undergraduate psychology student mentored for independent study

Knudsen, H. K., **Studts, C. R.**, & Studts, J. L. (2010). *Implementation of smoking cessation counseling in substance abuse treatment.* Paper presented at the 105th annual meeting of the American Sociological Association, Atlanta, GA.

Studts, C. R., & Studts, J. L. (2010). *Development of the Smoking-Related Guilt Scale.* Poster presented at the 31st annual scientific sessions of the Society of Behavioral Medicine, Seattle, WA.

Schoenberg, N. S., **Studts, C. R.**, Tarasenko, Y., Hatcher, J., Wright, S., Dollarhide, K., & Dignan, M. B. (2009). *Barriers to cervical cancer prevention among middle-aged and older rural Appalachian women.* Paper presented at the 62nd annual scientific meeting of the Gerontological Society of America, Atlanta, GA.

Perry, J. B., Adams, Lora A., & **Studts, C. R.** (2009). *Creating a clean start: Avoiding common pitfalls and barriers to successful collaborative care programs using action plans.* Poster presented at the 11th annual conference of the Collaborative Family Healthcare Association, San Diego, CA.

Studts, C. R. (2009). *Improving screening for externalizing behavior problems among very young children in primary care: Applications of item response theory.* Paper presented at The Ohio State University College of Social Work 21st National Symposium on Doctoral Research in Social Work, Columbus, OH.

Studts, C. R. (2009). *Predisposing, enabling, and need factors associated with mental health services use among very young children.* Poster presented at the 30th annual scientific sessions of the Society of Behavioral Medicine, Montreal, Canada.

Studts, C. R. (2009). *Early identification of child behavior problems in primary care: Who initiates the discussion?* Poster presented at the 30th annual scientific sessions of the Society of Behavioral Medicine, Montreal, Canada.

Studts, C. R., van Zyl, M. A. (2009). *Differential item functioning in the measurement of child externalizing behavior problems: An application of item response theory.* Paper presented at the 13th annual conference of the Society for Social Work and Research, New Orleans, LA.

Hatcher, J., **Studts, C. R.,** Dignan, M., & Schoenberg, N. E. (2009). *Characteristics of rural Appalachian women who are rarely or never screened for cervical cancer.* Paper presented at the 23rd annual conference of the Southern Nurses Research Society, Baltimore, MD.

Studts, C. R., Studts, J. L., Andrykowski, M. A. (2008). *Psychometric properties of a new Subjective Numeracy Scale: Classical and IRT analyses.* Poster presented at the 30th annual meeting of the Society for Medical Decision Making, Philadelphia, PA.

Studts, C. R., Studts, J. L., Andrykowski, M. A. (2008). *Psychometric properties of a new Subjective Numeracy Scale: Classical and IRT analyses.* Poster presented at the 29th annual scientific sessions of the Society of Behavioral Medicine, San Diego, CA.

Studts, J. L., Barnes, C. N., **Studts, C. R.**, LaJoie, A. S., Andrykowski, M. A., & LaRocca, R. (2007). *Participant adherence in a RCT of lung cancer screening: Results from baseline to year 1.* Paper presented at the 28th annual scientific sessions of the Society of Behavioral Medicine, Washington, DC.

Ruberg, J. L., **Studts, C. R.,** Barnes, C. N., LaJoie, A. S., Cross, T., LaRocca, R. V., Andrykowski, M. A., Studts, J. L. (2007). *Smoking cessation among participants in a RCT of lung cancer screening: Baseline to year one.* Poster presented at the 28th annual scientific sessions of the Society of Behavioral Medicine, Washington, DC.

Studts, C. R., Stone, R., & Barber, G. M. (2007). *A multilevel model of health status change among welfare recipients following welfare reform.* Paper presented at the 11th annual conference of the Society for Social Work and Research, San Francisco, CA.

Studts, C. R., Stone, R., & Barber, G. M. (2006). *Predictors of health care access among welfare recipients.* Poster presented at the 27th annual scientific sessions of the Society of Behavioral Medicine, San Francisco, CA.

Studts, C. R., Matera, E. L., & Studts, J. L. (2006). *An item response theory analysis of the Numeracy Questionnaire.* Poster presented at the 28th annual meeting of the Society of Medical Decision Making, Boston, MA.

Studts, C. R., Sephton, S., Helm, C. W., Studts, J. L. (2006). *Perceived cervical cancer risk among women undergoing colposcopy.* Poster presented at the 27th annual scientific sessions of the Society of Behavioral Medicine, San Francisco, CA.

Barnes, C. N., **Studts, C. R.**, LaJoie, A. S., Ruberg, J. L., Cross, T., Andrykowski, M. A., LaRocca, R. V., Studts, J. L. (2006) *Participant adherence in a RCT of lung cancer screening: Baseline to year 1.* Poster presented at Research Louisville, University of Louisville, Louisville, KY.
***2nd Place Award for Student Research

Studts, J. L., Ghate, S., Marmarato, J., Barnes, C., **Studts, C. R.**, LaJoie, A. S., LaRocca, R. (2006). *Validity of self-reported smoking status among lung cancer screening participants.* Poster presented at the 30th annual meeting of the American Society of Preventive Oncology, Bethesda, MD.

Studts, C. R., Elliott, A., Faith, T., Royer, B., & Young, S. (2004). *Positive behavioral supports in a Head Start classroom.* Paper presented at the annual meeting of the Midwest School Social Workers, Louisville, KY.

INVITED PRESENTATIONS

Studts, C. R. (2011). *Item response theory: A brief overview.* Invited lecture to course SW 773, Advanced Measurement in Social Work Research, Doctoral Program, Kent School of Social Work, University of Louisville.

Quigley, J. M., Young, T. L., Woods, S. H., & **Studts, C. R.** (2011). *Implementation of an integrated pediatric primary care clinic in General Pediatrics at the University of Kentucky.* Paper presented at the Contemporary Pediatrics Conference, sponsored by UK Children's Hospital and the Department of Pediatrics.

Studts, C. R. (2010). *Basic biostatistics I and II: An overview for scientist-practitioners.* Invited two-part lecture in the Spring Lecture Series of the Department of Psychiatry, College of Medicine, University of Kentucky.

Studts, C. R. (2008). *Item response theory: A brief overview.* Invited lecture to course SW 773, Advanced Measurement in Social Work Research, Doctoral Program, Kent School of Social Work, University of Louisville.

Studts, C. R. (2008). *Improving screening instruments for early identification of externalizing behavior problems: Applications of item response theory.* Invited presentation to the Department of Behavioral Science, University of Kentucky College of Medicine.

Studts, C. R. (2007). *Basic biostatistics I and II: An overview for clinicians.* Invited two-part lecture in the Spring Lecture Series of the Department of Ophthalmology, School of Medicine, University of Louisville.

Studts, C. R. (2006). *Introduction to item response theory.* Invited lecture to the Behavioral Oncology Lab, Cancer Prevention & Control Program, James Graham Brown Cancer Center, University of Louisville.

Studts, C. R., Matera, E. L., & Studts, J. L. (2006). *An item response theory analysis of the Numeracy Questionnaire.* Invited lecture in the Fall Lecture Series of the Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville.

Studts, C. R. (2005, 2006). *Item response theory: Introduction and application to measurement of health outcomes.* Invited lectures to course PHCI 602, Health Services and Outcomes Research, CREST program, School of Public Health and Information Sciences, University of Louisville.

CONSULTING EXPERIENCE

2/04 - 3/04 **Kent School of Social Work**, Louisville, Kentucky
Consulted with Masters program research faculty and students to expedite the preparation and submission of MSSW research project proposals to the University of Louisville Institutional Review Board (IRB). Assisted students in preparing and revising protocols to meet ethical and scientific research standards. Collaborated with IRB staff as needed.

PRE-DOCTORAL RESEARCH EXPERIENCE

7/05 - 5/08 **Graduate Research Assistant**, James Graham Brown Cancer Center, U. of L.
Performed data collection, database maintenance, statistical analyses, literature reviews, and manuscript writing for three ongoing research projects:
Lung Cancer Screening Study
Colposcopy Study
Lung Cancer Decision Making Study
Supervisor: David Hein, Ph.D., Director of Cancer Prevention and Control Program, James Graham Brown Cancer Center, University of Louisville School of Medicine

8/03 - 4/08 **University Fellow,** Kent School of Social Work, University of Louisville

Engaged in research activities with several faculty members during studies in the doctoral program, including:

Gerard Barber, Ph.D. and Ramona Stone, Ph.D., Kent School of Social Work: *Health and Welfare Reform*

Andy Frey, Ph.D., Kent School of Social Work: *Positive Behavioral Supports in Head Start*

Jamie L. Studts, Ph.D., James Graham Brown Cancer Center: *Colposcopy Study, Lung Cancer Decision Making Study, Lung Cancer Screening Study*

9/92 - 12/92 **Women and AIDS Coalition**, South Bend, Indiana

Performed data entry and conducted preliminary statistical analyses.

TEACHING EXPERIENCE

8/09 - 12/10 **Instructor,** PSY 395 (Independent Study in Psychology), University of Kentucky

Mentored three undergraduate psychology students (Andrea Arrowood, Elyse Hoxby, and Gina Sabato) in aspects of conducting research, including human subjects protection, participant recruitment, data collection, data entry, and introductory analyses in SPSS. Trained and engaged students in pilot study responsibilities.

1/07 - 3/07 **Research Rotation Supervisor**, School of Medicine Med-Peds Program, University of Louisville

Supervised the research rotations of Demeka Y. Campbell, M.D., and Cynthia Bowman-Stroud, M.D., med-peds residents. Aspects of the rotations included guidance and training on literature reviews, formulating research questions and hypotheses, study design, survey design, data collection, data analysis, and scientific writing.

11/05 **Guest Lecturer**, CREST, School of Public Health & Information Sciences, U. of L.

PHCI 602, Health Services and Outcomes Research

Topic: "Measuring Depression and Anxiety"

12/04 Developed a masters-level social work elective course, including syllabus, readings, assignments, and examinations: *Social Work Practice in Health Care Settings*

9/04 - 12/04 **Tutor**
Provided private tutoring in graduate-level statistics for social work students.

5/04 - 7/04 **Co-Instructor,** Kent School of Social Work, Louisville, Kentucky
SW 766, Doctoral Preparation
Co-taught the summer course for incoming doctoral students, with Ruth Huber, Ph.D. (director of the doctoral program). Assisted with development of course outline and syllabus; prepared and administered the online Blackboard course site; taught portions of class sessions; provided individual and group tutoring to students; and provided feedback on homework and in-class assignments. Course content included a review of the basics of social work research, statistics, use of SPSS, and significant past and present themes in the social work literature.

3/99 - 3/01 **Teacher,** Kaplan Educational Center, Lexington, KY and Durham, NC
Prepared and taught preparation courses for all portions of the GRE, LSAT, ACT, and SAT standardized tests to classes of up to 20 students as well as to individual students. Also taught the Verbal sections of the PCAT (Pharmacy) and DAT (Dental) standardized tests.

ADVISING ACTIVITY

Completed Advising

Doctoral Committees:
- Stephan Buckingham – Social Work (University of Louisville – def. 7/09)
- Shaena Y. Gardner - Clinical Psychology (Spalding University – def. 3/06)

CLINICAL & ADMINISTRATIVE EXPERIENCE

5/04 - 9/05 **FORECAST, Kent School of Social Work,** Louisville, Kentucky

Member of Development Team, Clinician, and Clinical Supervisor
Assisted with development of a clinic designed to provide comprehensive assessments of families involved with the Jefferson County Department of Community Based Services. Contributed to development of protocols, clinic resources, and relevant literature reviews. Provided comprehensive assessments of potential foster/adoptive parents, foster/adoptive families facing possible disruption, and biological parents involved with Child Protective Services, in a clinic funded by the Kentucky Cabinet for Health and Family Services. Consulted with Cabinet staff and supervisors to

provide assessments and recommendations. Provided clinical supervision to masters-level certified social workers pursuing independent licensure.

7/01 - 7/03 **Seven Counties Services, Inc.**

Principal Social Worker, School Based Services, Louisville, KY
Provided mental health assessments, treatment planning, and services for elementary school children in the school setting. Consulted with school staff (teachers, guidance counselors, principals) in three Jefferson County elementary schools to provide education and recommendations regarding clients and school populations in general. Advocated for special needs of clients, such as psychoeducational assessments, classroom accommodations, and placements. Coordinated with community agencies (courts, social services) to provide appropriate and effective services. Collaborated with multidisciplinary mental health professionals as part of a treatment team.

Senior Social Worker, Bullitt County Child and Family,
Shepherdsville, KY
Provided mental health assessments, treatment planning, and services for children, adolescents, and families in a rural outpatient mental health agency. Coordinated with community agencies (schools, courts, social services) to provide appropriate and effective services for clients. Collaborated with multidisciplinary mental health professionals as part of a treatment team.

7/00 - 7/01 **Duke University Medical Center**

Clinical Social Worker, Duke Children's Primary Care, Durham, NC
Provided clinical and case management social work services for a pediatric primary care clinic. Performed crisis assessment and intervention, in addition to ongoing support services. Collaborated with physicians, nurses, psychologists, and other health professionals to optimize family access to treatment and resources. Educated medical residents in the clinic setting on psychosocial/mental health issues and community resources. Coordinated efforts with multiple local agencies to improve patient and family care. Participated in on-call and coverage teams with pediatric clinical social workers throughout the medical center.

5/97 - 6/00 **Bluegrass Regional Mental Health – Mental Retardation Board**

Program Director, R.I.S.E. Program, Harrodsburg, KY
Directed a mental health and educational program for 60 children. Hired, trained, and supervised 13 mental health specialists and teachers. Coordinated efforts with local schools and agencies. Maintained administrative and direct service records. Assisted in the development and promotion of a new R.I.S.E. program in Lawrenceburg, Kentucky.

Outpatient Therapist, Harrodsburg and Stanton, KY
Provided mental health assessments, treatment planning, and services for children, adults, and families in the outpatient Comprehensive Care Centers.

Mental Health Specialist, R.I.S.E. Program, Harrodsburg, KY
Provided mental health and educational services to 15 children as part of a multidisciplinary team. Maintained appropriate clinical documentation of services. Assisted program director with administrative tasks and program preparation.

SERVICE EXPERIENCE

Academic Service

2011 - 2012 **Co-Chair, Rapid Communications Track, 33rd annual meeting of the Society of Behavioral Medicine**

2011 **Ad hoc reviewer for:**
Supportive Care in Cancer (4)
Journal of Behavioral Health Services & Research (1)
Head & Neck (1)

2010 **Ad hoc reviewer for:**
Supportive Care in Cancer (5)
Journal of Behavioral Health Services & Research (2)

2007 - 2011 **Abstract reviewer for annual meetings of the Society of Behavioral Medicine**
Medical Settings track (2010 - 2011)
Psychological and Person Factors track (2007 – 2011)

2009 - 2010 **Planning Committee Member, 12th Annual Conference of the Collaborative Family Healthcare Association**
Served on the planning committee for the annual CFHA conference. Contributing member of a subcommittee planning the Kentucky

Health Policy Summit, *Integrating Mental Health and Primary Care Services in Kentucky*, a regional summit on collaborative family healthcare.

Institutional Service

10/11 - **UK Center for Clinical and Translational Science**
Serving on a committee focused on social network analysis and bioinformatics.

8/04 - 5/05 **Kent School Doctoral Faculty, Kent School Faculty, and Kent Assembly**
Doctoral student representative to meetings; provided information to doctoral students and solicited input to present to faculty and staff.

8/03 - 5/05 **Kent School of Social Work Outcomes Committee**
Served on committee with focus on improving and monitoring outcome measures of the Kent School as required by accrediting bodies.

Community Service

7/04 - 4/06 **Metro United Way Success by 6, Nurturing Young Children Action Team**
Louisville, Kentucky
Served on the Nurturing Young Children Action Team to promote collaboration of community programs and services targeting school readiness. Helped initiate a subgroup focusing on child health and safety issues.

7/00 - 7/01 **Durham Interagency Council for Young Children with Special Needs**, Durham, North Carolina
Collaborated with community leaders toward improving local efforts to identify infants and toddlers with special needs and promoting services for this population. Served as Council Secretary.

PRACTICUM EXPERIENCE

1/97 - 5/97 **Domestic Violence Prevention Board**, Lexington, Kentucky
Participated in multidisciplinary and interagency strategic planning groups on state and local levels.

8/96 - 12/96 **Bluegrass Regional Mental Health – Mental Retardation Board, Inc.**, Winchester, Kentucky

Performed intake psychosocial assessments, determined preliminary diagnoses, and triaged client assignments to therapists under clinical supervision.

8/95 - 5/96 **Jessamine County School District**, Nicholasville, Kentucky
Developed and facilitated treatment groups in a middle school and an alternative high school under the supervision of an at-risk counselor.

VOLUNTEER EXPERIENCE

8/94 - 5/95 **Family Life Head Start Child Development Center, CAP Volunteer Program**, Mt. Vernon, Kentucky
Supervised and guided the developmental play time of 25 Head Start preschoolers. Created lesson plans and planned activities after conducting assessments of individual children's needs. Full-time volunteer.

8/93 - 8/94 **Family Life Services, CAP Volunteer Program**, Mt. Vernon, Kentucky
Provided extensive follow-up services (parenting, budgeting, problem solving, emotional support) to 30 families who completed a residential program. Assisted with day-to-day operations in the shelter through a wide variety of tasks. Full-time volunteer.

9/92 - 12/92 **University of Notre Dame Crisis Line**, Notre Dame, Indiana
Volunteer Crisis Telephone Peer Counselor

9/91 - 12/91 **St. Mary of the Angels Youth Program**, London, England
Volunteer Staff Member

9/90 - 5/91 **St. Mary's Native American Tutoring Program**, South Bend, Indiana
Volunteer Tutor for elementary school students

TRAINING AND WORKSHOP EXPERIENCE

2012 Society of Behavioral Medicine, *New Orleans, LA*
Multiphase Optimization Strategy (Linda Collins)

2011 NIH Summer Institute on Behavioral Randomized Clinical Trials, *Airlie, VA*
Introduction to Social Network Analysis (Methods Work), *Silver Spring, MD*

2009 Society of Behavioral Medicine, *Montreal, Quebec, Canada*
Using the Statistical Language R to Analyze Item Response Data for Measurement Development

2008 Introduction to Latent Class/Latent Transition Analysis (Linda Collins), *Lexington, KY*

2006 Evidence-Based Practice (Eileen Gambrill & Leonard Gibbs), *Louisville, KY*
Society of Behavioral Medicine, *San Francisco, CA*
Latent Class and Latent Profile Analysis: Creating Typologies via Categorical Latent Variables
Communication Skills in Statistical Consulting (Janice Derr), *Louisville, KY*

2005 Society of Behavioral Medicine, *Boston, MA*
Modern Psychometrics and Health Outcomes Assessment
Introduction to Item Response Theory: Methods and Applications
Measurement: Theory and Applications in Social Work Research (William Nugent), *Lexington, KY*

2003 Ethics in Social Work Practice, *Louisville, KY*
Clinical Supervision Training for Kentucky Board of Social Work, *Louisville, KY*

2002 Kentucky Play Therapy Association Conference, *Louisville, KY*
SCERTS Interventions for Autistic Spectrum Disorders, *Indianapolis, IN*

2001 Safe Crisis Management, *Louisville, KY*
HIV/AIDS Awareness Training, *Louisville, KY*

2000 Explosive and Inflexible Children, *Lexington, KY*
Expressive Therapies with Sexually Abused Children, *Lexington, KY*

1999 V.I.S.I.O.N. Training (Multicultural Issues in Mental Health), *Lexington, KY*
The Canadian Play Therapy Institute, *Lexington, KY*

1997 Domestic Violence Training, *Lexington, KY*
Victims Advocacy Training, *Frankfort, KY*

1996 The Canadian Play Therapy Institute, *Lexington, KY*
Kentucky School Social Work Conference, *Louisville, KY*
ADHD Workshop, *Lexington, KY*

1995 The Fall Institute: Children and Families First, *Louisville, KY*
Family Literacy: Creating a Community of Learners, *Lexington, KY*

HONORS & AWARDS

2011 Accepted as a Fellow to the 11th Annual Summer Institute on the Design and Conduct of Randomized Clinical Trials Involving Behavioral Interventions, sponsored by the NIH Office of Behavioral and Social Sciences Research (July 10 – July 22, 2011)

2009 Selected to present at The Ohio State University College of Social Work 21st National Symposium on Doctoral Research in Social Work

2008 Nominated for Society of Behavioral Medicine 2009 Outstanding Dissertation Award

2008 Nominated for Society for Social Work and Research 2009 Outstanding Social Work Doctoral Dissertation Award

2008 University of Louisville Graduate School Dean's Citation

2007 Travel Awards: University of Louisville Graduate Student Council, Kent School of Social Work Alumni Fund, and Kent School Student Association

2003 - 2007 University of Louisville Graduate School Fellowship

1997 Alpha Delta Mu Honorary Society

1996 - 1997 University of Kentucky Graduate School Presidential Fellowship

1995 - 1996 University of Kentucky College of Social Work Scholarship

1989 - 1993 University of Notre Dame Orchestra

1989 - 1993 University of Notre Dame Dean's List

MEMBERSHIP IN PROFESSIONAL ORGANIZATIONS

04/06 - American Statistical Association, Student Member

12/03 - Society for Social Work and Research, Member

12/03 - Society of Behavioral Medicine, Member