TYPE I ERROR RATES OF THE DIF MIMIC APPROACH USING JÖRESKOG'S COVARIANCE MATRIX WITH ML AND WLS ESTIMATION

by

MICHAELA NICOLE GELIN

M.A., The University of British Columbia (2001)
Diploma in Counselling Psychology, The University of British Columbia (1999)
B.A. (Psychology), The University of British Columbia (1998)

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Measurement, Evaluation & Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA

February 2005

ABSTRACT

This dissertation presents new research that examines the Type I error rate of a structural equation modeling (SEM) approach for investigating differential item functioning (DIF) in short scales. Specifically Muthén's SEM model for DIF is examined using a new estimation method (Jöreskog, 2002). In general, this method conditions on the latent variable while simultaneously testing the effect of the grouping variable over-and-above the underlying latent variable of interest. Thus, it is a multiple-indicators, multiple-causes (MIMIC) model for DIF. The Type I error rates of this DIF MIMIC approach are explored using data that are reflective of short scales with ordinal item response formats typically found in the social and behavioral sciences. The variables included in this Monte Carlo simulation are 7 sample size combinations (3 equal and 4 unequal group combinations), 2 item response distributions (symmetric and positively skewed), 2 scale lengths (10 and 20 items per scale), and 2 estimation methods (maximum likelihood and weighted least-squares). The results indicate that the Type I error rates for the DIF MIMIC model are inflated for both estimation methods under all of the design conditions. These results are discussed in the context of validity including the implications of inflated Type I error rates for items identified as displaying DIF.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I first wish to gratefully acknowledge my mentor and supervisor, Dr. Bruno Zumbo, for his enduring support, guidance, statistical expertise, and enthusiasm throughout all stages of this dissertation. His assistance has been invaluable and working with him has been a rich and rewarding experience. Without his support I would not be where I am today. I am sure I could never express the depth of my appreciation as he has had a tremendous influence on my life.

I would also like to acknowledge my committee members, Dr. Anita Hubley and Dr. Kimberly Schonert-Reichl, for providing insightful and thoughtful suggestions throughout this dissertation. I would also like to thank Dr. Beth Haverkamp and Dr. Jim Frankish for being the University examiners and Dr. Susan Maller for being the external examiner. On a personal note, I would like to especially thank my parents and grandmother for their unwavering support and confidence as I pursued my undergraduate and graduate degrees. Additional thanks are due to my friends, in particular Sophie Aerts, who have kept me laughing and enjoying life.

Finally, I wish to acknowledge the Isaak Walton Killam Memorial Fund as well as the Social Sciences and Humanities Research Council (SSHRC) of Canada for providing financial support throughout my degree. This support allowed me to direct my time and energy toward my studies and to the completion of this dissertation.

CHAPTER I: INTRODUCTION TO THE PROBLEM

Historical background to DIF

The purpose of this introduction is to give a brief historical background through which matters of test bias and differential item functioning (DIF) methods developed. This brief history is shaped by the work of Gould (1996). DIF arose within the context of test bias and high-stakes decision making involving ability, achievement, certification and licensure measures[1] used in large-scale testing programs. Concerns about test bias appear to have begun in 1905 when Alfred Binet and Theophile Simon developed the first intelligence scale (cited in Binet & Simon, 1973). In 1916 the Binet-Simon Intelligence Scale was expanded and reworked by Lewis Terman at Stanford University (Gould, 1996). Terman's work resulted in the *Stanford Revision of the Binet-Simon Scale* (also known as the Stanford-Binet). Given Terman's belief that intelligence was hereditary he investigated whether IQ test items functioned differently for children who were from different social classes. In particular, he posed the question "Is the place of so-called lower classes in the social and industrial scale the result of their inferior native endowment, or is their apparent inferiority a result of their inferior home and school environment?" (Terman, 1916, p. 19). Based on his research, Terman concluded that children of higher social-class parents would be better endowed than children of lower social-class parents (Gould, 1996; Minton, 1998). Implicit, but empirically untested, is the assumption that the items and test function the same for the various social classes.

With the publication of the Stanford-Binet scale, Terman was called to serve on a committee, led by Robert Yerkes, to devise mental tests for the United States army; test

[1] The terms test, measure, and scale are used synonymously in this dissertation.

scores which were used during World War I to determine job placements or discharge from the army. The United States army testing program played an important role in fueling the mental testing movement during which time advocates and test developers such as Henry H. Goddard, Lewis M. Terman and Robert M. Yerkes believed that intelligence was mostly hereditary (Mackintosh, 1998; Minton, 1984). After the war, Yerkes helped revise the army tests for school use and by 1920 the "National Intelligence Tests" for grades three to eight were ready for use with the purpose of classifying students into homogeneous ability groups (Mackintosh, 1998). According to Minton (1998), it was during the early 1920s that critiques of the testing movement began to raise questions about bias. Specifically, they pointed to the cultural bias of tests that placed individuals with little education (e.g., African Americans) and recent immigrants (e.g., Southern & Eastern Europeans) at a distinct disadvantage. These criticisms led to two major public debates in 1928 and 1940 around the issue of whether enriched environmental experiences (e.g., preschool experience) could significantly raise IQ scores (see Minton, 1984). These debates sparked further interest in whether group differences in ability were due to genetics or to the environment (e.g., Eells, David, Havighurst, Herrick, & Tyler, 1951) and whether tests are biased for sub-groups of examinees.

This nature versus nurture debate (also called the hereditary-environment controversy) continued over the next 20 years during which time matters of *fairness and equity* were paramount. That is, there should be an equal playing field where, for example, male and female students have equal opportunities to do well in large scale assessments, and hence being treated equitably in terms of test score performance. During the late 1960s and 1970s, the civil rights movement in the United States began focusing its attention on the role

of tests in denying access to employment opportunities and equal education (Camilli & Shepard, 1994). These concerns of test bias were further heightened by the publication of Jensen's (1969) work on the heritability of intelligence (i.e., IQ differences between blacks and whites). As result of the significant concern for equitable tests, a considerable amount of literature on statistical methods for analyzing test bias appeared during the mid-1970s (Camilli & Shepard).

The statistical methods introduced in the 1970s for examining test bias were based on early classical test theory (CTT) techniques that often focused on the interaction between item performance and group membership by comparing mean item scores across sub-groups of interest. CTT statistics, such as the mean, *p* values (item difficulty) or item-total correlations (i.e., item discrimination), summarize the sample as a whole and do not take into account that the measurement properties of the test or measure may vary as a function of variation within the sample. In other words, CTT indices depend on group differences in the underlying latent variable measured by the test. For example, an item's *p* value from a population exhibiting high levels of depression will be higher than the *p* value for the same item from a population exhibiting low levels of depression. Accordingly, the key problem with these early CTT techniques is that they never conditioned (i.e., matched) examinees on a variable that represented the construct under investigation. In order to "match" individuals, each individual in one group (e.g., females) must be matched with an individual in the other group (e.g., males). The matching is done so that the two individuals are equivalent (or nearly equivalent) with respect to a specific variable (i.e., matching criterion) that the researcher would like to control. The matching criterion can be the observed total test score, latent variable score or an external criterion (e.g., diagnosis) that represent the specific skill,

ability, or behavior that the test has been developed to measure. For example, two individuals who were from different groups (e.g., aboriginal and non-aboriginal) but who had the same total score on an overall test (e.g., IQ test) would be considered "matched." Because this matching is not part of CTT methods, what appear to be group differences in observed variables (i.e., item mean differences between groups) from CTT techniques may be distorted because one's measure may not be invariant across the groups being compared. Another way of saying this is that the groups need to be matched before one can compare item performance (Clauser & Mazor, 1998). As a result, a class of statistical methods under the name item bias procedures was introduced, which were later (1980s) termed DIF (e.g., Holland & Thayer, 1988).

This shift in terminology from item bias to DIF arose out of the conflict between the public and technical community's use of the term bias. As Cole (1993) discusses, the public viewed the word bias as bad, unfair, and working against equal opportunity (e.g., prejudice). In terms of testing, the public viewed an item as biased when it unfairly favors one group over another. Hence, there was an inherent value judgment in the term bias. However, the technical psychometric community was using the term to denote a type of statistical characteristic. As stated by Cole "to us [in the technical community], bias was certainly not good but it was a less than optimal technical characteristic, not a social evil" (p. 27). Thus, it was necessary to differentiate between the social and technical terminologies more clearly. Accordingly, "the term *differential item functioning* (DIF) rather than bias is used commonly to describe the empirical evidence obtained in investigations of bias" (Hambleton, Swaminathan & Rogers, 1991, p. 109). "Empirical evidence of differential performance is necessary, but not sufficient, to draw the conclusion that bias is present; this conclusion

involves an inference that goes beyond the data" (Hambleton et al., p. 109). The presence of bias requires the further condition that the differential performance is attributed to some characteristic of the test item or testing situation that is irrelevant to the testing purpose (i.e., unrelated to the construct measured by the test).

What is DIF?

The statistical methodology called DIF was introduced as a method for evaluating differential item response patterns among different groups (e.g., gender, ethnic, social class, age). Based on Clauser and Mazor's (1998) definition, "differential item functioning is present when examinees from different groups have differing probabilities or likelihoods of success on an item, after they have been matched on the ability of interest" (p. 21). The defining feature of DIF detection is that one is investigating differences in item responses after *statistically matching*[2] the groups on the variable of interest (Angoff, 1993; Camilli & Shepard, 1994). It was reasoned that if subgroups have (statistically) the same amount or level of the latent trait (or ability) on the construct measured by the test they should have the same probability of responding the same way to items on the test. Thus, it was determined that matching groups on the latent variable measured by the test is important for determining whether item responses are equally valid for different groups.

The fact that the item response is conditional on the level of the latent trait is fundamental for distinguishing between differences in the functioning of the item and differences in the underlying trait (e.g., ability) of the groups. If, after statistically controlling for the variable of interest (i.e., the conditioning or matching variable), the groups differ in

[2] As opposed to experimental matching wherein the experimenter physically matches participants, DIF uses statistically matching such as that used in the analysis of covariance (ANCOVA).

terms of their item responses, this would be an indication of either *item bias* (i.e., systematic unfairness in how the test measures) or *item impact* (i.e., differences in test results reflect genuine differences on the construct or measure of interest). Throughout this dissertation the terms matching and conditioning variable will be used interchangeably.

Item bias

Item bias occurs if the source of the differential functioning of the item is *irrelevant* to the purpose and interpretation of the scale (Camilli & Shepard, 1994; Clauser & Mazor, 1998). In other words, item bias presents itself when the differences in item responses between the groups are erroneously attributed to the construct of interest, hence being spurious. For example, the differences could be due to an irrelevant factor such as the characteristic of the test item or testing situation. In terms of a depression inventory, item bias would be present if an item is measuring social desirability as opposed to depression. In essence, item bias is an artefact of the testing procedure. For item bias to be present DIF must be apparent; however, as Zumbo (1999) reminds us, "DIF is a necessary, but not sufficient, condition for item bias" (p. 12).

Item impact

Alternatively, item impact occurs if the source of the differential functioning of the item is a *relevant* characteristic of the scale. In other words, the differences in item responses between the groups are correctly attributable to the construct of interest, thereby inferring that there are 'real' differences between the groups of interest. For example, item impact would occur if an item from a depression inventory is measuring depression and males and females truly differ on depression. That is, item impact is evident when one group of

examinees is found to endorse an item more than another group of examinees because the two groups truly differ on the underlying factor being measured by the scale item.

Uses of DIF

According to its definition, an item exhibits DIF if individuals of comparable ability, but from different groups, do not have the same probability of endorsing an item or getting the item correct (Camilli, 1993; Hambleton, Swaminathan & Rogers, 1991). Based on the definition, Zumbo and Gelin (in press) describe three general uses for DIF:

1. *Fairness and equity in testing.* This purpose of DIF is often because of policy and legislation in which the groups (e.g., visible minorities or language groups) are defined ahead of time.
2. *Dealing with a possible "threat to internal validity."* In this case, DIF is often investigated so that one can make group comparisons and rule-out measurement artifact as an explanation for the group difference. The groups are identified ahead of time and are often driven by an investigator's research questions (e.g., gender differences in depression).
3. *Trying to understand the (cognitive and/or psychosocial) processes of item responding and test performance, and investigating whether these processes are the same for different groups of individuals.* In this context, the groups are not identified ahead of time; instead, latent class or other such methods are used to "identify" or "create" groups and then these new "groups" are studied to see if one can learn about the process of responding to the items. An example of this type of DIF can be found in a study by Li, Cohen and Ibarra (2004) who investigated the characteristics of mathematics items associated with gender DIF.

Purpose and structure of the dissertation

A variety of statistical methods have been developed over the years to aid the researcher in identifying DIF items for the purposes described above. The purpose of this dissertation is to investigate the statistical properties of a relatively new DIF detection method that is becoming widely used in psychological research – the structural equation modeling MIMIC DIF detection method. A common characteristic of many psychological scales is that they are relatively shorter (i.e., have fewer items) than the scales found in large-scale educational testing wherein DIF detection methods evolved (Zumbo & Hubley, 2003). Large scale tests are often tests that have well over 50 items. In Chapter II I will briefly review, compare, and contrast the most commonly used statistical DIF detection methods with an eye toward conducting a computer simulation study to investigate the Type I error rate of the SEM DIF method. This Chapter will close with a detailed description of the MIMIC DIF method.

Chapter III includes the simulation methodology for investigating the Type I error rates of the proposed DIF MIMIC approach for short scales. This chapter is divided into three subsections. The first subsection discusses the general method, followed by a section describing the simulation study design. This latter section includes a discussion of the variables that are manipulated in the simulation, including the scale length, item response distribution, and sample size combinations. The third section describes the nine steps required to generate the data and run the simulation. Chapter IV discusses the results of the simulation study, including the reliability of the scales, confirmatory factor analysis results, model fits, and rejection rates for Type I error rates.

The closing chapter is a summary and discussion of the study results in the context of educational and psychological measurement. This chapter revisits the research purpose and reminds the reader of the novel contribution made to educational and psychological measurement by this dissertation. Next, a review of the study results is provided along with a description of the limitations of the research study. This is followed by a discussion of the implications for research and practice, which includes a discussion of the consequences of inflated Type I error rates for items displaying DIF. Lastly, the future directions for research are considered.

In terms of format, this dissertation is prepared in a manuscript-based format following the 2004 University of British Columbia Faculty of Graduate Studies Doctoral Theses submission guidelines. Accordingly, this dissertation contains elements of the traditional thesis format (e.g., literature review, objectives, discussion, and conclusion) and these are integrated and presented as part of a unified document. This dissertation format is chosen in order to help the author "gain scientific writing experience in a format used by researchers in their field of study" (The University of British Columbia, 2004, p. 1).

CHAPTER II: FRAMEWORKS FOR DIF: REVIEW OF THE LITERATURE

Frameworks for considering DIF

Given the general definition of DIF described in Chapter I, there are a number of different approaches for detecting DIF that can be classified into two frameworks: the observed score framework and the latent variable modeling framework. The essential difference between these approaches is whether the matching variable is an observed or latent variable. Using these frameworks, the following section will briefly review different methodological approaches for detecting DIF. In doing so, this review will highlight the advantages of a latent variable SEM approach for detecting DIF in short-scales as compared to other DIF methods.

Observed score framework

The key characteristic of observed score DIF approaches is that they condition (i.e., match) on the observed or manifest variable. This observed variable is often an aggregate or composite score such as the total scale score or it can be an item purified score such as the rest score (e.g., the total scale score minus the studied item which is found in the corrected item-total correlation). This observed variable could also be a different criterion measure such as a medical diagnosis of depression from a clinician or from a 'gold standard' (e.g., diagnostic interview for depression). This matching alternative would be useful if, for example, the aggregate score was found to be misleading or inappropriate, such as the case if DIF were found for many of the scale items.

Within the observed score framework, there are two broad classes of DIF detection methods: multi-way contingency tables (e.g., Mantel-Haenszel; MH) and generalized linear models (GLIM) (e.g., logistic regression and ordinal logistic regression).

Multi-way Contingency Tables

The most widely used multi-way contingency table method is the Mantel-Haenszel (MH). The MH class of methods (Holland & Thayer, 1988) make use of three-way contingency tables for detecting DIF. "The three dimensions of the contingency table involve (a) whether one gets an item correct or incorrect [or whether one endorses the item or not], (b) group membership, while conditioning on (c) the total score discretized into a number of category score bins" (Zumbo & Hubley, 2003, p. 506). Consequently, the conditioning variable is categorized into levels that are often arbitrary and may therefore affect the statistical decision regarding DIF (Zumbo, 2003). Moreover, MH methods do not test for interactions (i.e., non-uniform DIF) and do not enable one to control for covariates other than the group membership being tested (i.e., no multiple conditioning variables).

Generalized Linear Models (GLIM)

Alternatively, the second class of methods within the observed score framework use regression analysis to statistically test the effect of the grouping variable (i.e., uniform DIF – the item is more difficult at all ability levels for one group than for the other group) and interaction of group by conditioning variable (i.e., non-uniform DIF) over-and-above the variable of interest – the matching variable. As Zumbo and Hubley (2003) remind us, these methods are akin to ANCOVA or attribute by treatment interaction (ATI) designs. When the item format is binary, a logistic regression approach (LogR; Swaminathan & Rogers, 1990)

is used wherein the conditioning variable is either the total score or rest score on the scale. This LogR equation can be expressed as

$$Y = b_0 + b_1(Conditioning) + b_2(group) + b_3(Conditioning * group) + \varepsilon \quad (1)$$

where Y is a natural log of the odds ratio and ε is error. As stated by Zumbo (1999), the advantages of LogR compared to the MH method are that (i) a continuous conditioning variable does not need to be discretized, (ii) uniform and non-uniform DIF can be modeled, and (iii) the LogR methodology can be extended for use with ordinal item formats (i.e., an ordinal Log R approach can be used; Zumbo, 1999).

When using the ordinal LogR approach for rating scale data, one can re-express Equation (1) as a linear regression of predictor variables on a latent continuously distributed random variable, Y*,

$$Y^* = b_0 + b_1(Conditioning) + b_2(group) + b_3(Conditioning * group) + \varepsilon .$$

One advantage of Zumbo's (1999) ordinal LogR approach is that it allows for a direct comparison of the results from binary and ordinal scored items and hence, both item formats can be used (Gelin & Zumbo, 2003). In addition, this method has a corresponding effect size estimator that can be used with binary and ordinal items to help determine the magnitude of DIF.

Both the LogR and ordinal LogR approaches for detecting DIF have a number of general assumptions that must be kept in mind. One assumption is that of local independence. That is, for persons with roughly the same amount of the latent trait on the construct measured by the scale (e.g., same amount of 'depression' as measured by an individual's true score), the item responses are independent across items. A related second assumption is that the measure is essentially unidimensional. In the statistical context, this means that the latent

variable accounts for the correlation among the items; psychometrically, it suggests that each individual possesses some amount of the latent variable that the scale measures and hence there is only one 'thing' driving individuals' responses.

Possible disadvantages of this approach are that the item scoring may impact DIF detection and low item variability may result in incorrect DIF detection (Gelin & Zumbo, 2003). In addition, the use of the total test score as a conditioning variable is not optimal (see Millsap & Everson, 1993). This latter disadvantage also applies to multi-way contingency table methods and will be further discussed later in this chapter.

Latent variable framework

In addition to observed score DIF methods, one can also approach DIF from a latent variable framework. Before describing the four latent variable DIF detection methods, the concept of a latent variable needs some clarification because this term has a number of different meanings, each of which can lead to quite different variables (Zumbo & Rupp, 2004). The first definition, and that which is the closest description to a latent variable in classical test theory, is that latent variables are real variables that could, in principle, be measured (e.g., proficiency or knowledge in a domain such as math). A second form of a latent variable is when observed scores arise by recording whether or not underlying variable(s) have values above or below fixed thresholds (e.g., a response to a Likert-type question). This form of a latent variable is found in probit models, ordinal logistic regression and polychoric and tetrachoric correlations. The former definition can be conceptualized within a framework of the latter definition, although it does not necessarily have to be so. The third definition, and most commonly used meaning in the social sciences, describes latent variables as constructed variables that come prior to the items (or indicators) of which

we measure. This meaning is used in factor analysis, latent variable modeling, and covariance structure models; it is therefore used in all four latent variable methods of DIF described in this chapter. It is worth noting that a latent variable is different from a construct. A latent variable is a statistical entity in which individuals can have an amount of a latent variable (e.g., depression) – it is model dependent. A construct is, in essence, higher-order compared to a latent variable, and sits within a nomological network of ideas (e.g., theories of depression). In other words, the items are indicators of the latent variable, and constructs are inferred from latent variables.

The defining feature of latent variable methods is that a latent variable is constructed from the data. Moreover, these methods use the joint distribution among items to model a latent variable. Within this latent variable framework, there are four broad classes of DIF detection methods: (1) conditional, (2) unconditional, (3) multidimensional, and (4) structural equation modeling (SEM).

Conditional methods of DIF

Conditional methods use a similar modeling strategy as LogR, except a latent variable or trait score is used as the conditioning variable. This is, therefore, a two-step approach in which the factor score (i.e., latent variable score) is first computed and saved in the database, then either the LogR or ordinal LogR methodology is applied with the factor score as the conditioning variable. This can be expressed as

$$Y = b_0 + b_1(\theta) + b_2(group) + b_3(\theta * group) + \varepsilon$$

where θ is the latent variable score. Because this approach uses regression analysis methodology, the same assumptions as LogR methods apply. However, unlike the observed score methods, the major limitation of this two-step approach is that a large number of items

are required to accurately estimate the latent variable, θ, via the saved factor score. Later on in the discussion, I address what is meant by a large number of items. As the number of items decreases, the measurement error of theta increases thereby reducing the reliability. A second limitation is that the factor score can be biased if the model from which it was estimated is incorrect. For example, if method effects were not modeled before the factor score was saved, the factor score would be biased. Although one can use this 2-step method to incorporate complex modeling (i.e., create a confirmatory factor analysis model with method effects and then save the latent variable score), it is very time consuming and statistically inefficient. Moreover, this two-stage approach leads to biased R-squared values and estimates of model parameters (e.g., Shevlin, Miles & Bunting, 1997) and in the case wherein the latent variable scales are based on too few items (i.e., short scales), the "bias can be severe" (Lu, Thomas & Zumbo, in press, p. 4). Specifically, in a Monte Carlo study for binary items, Lu, Thomas and Zumbo found that the regression approach is insensitive to sample size but has appreciable attenuation in regression parameter estimates and R^2 values (approximately 34% bias for sample sizes 300 to 2000 with 10 items). In addition, they found that bias in regression approach decreases as the number of items increases from 10 to 30. Thus, as Skrondal and Laake (2001) conclude, "conventional factor score regression performs badly and should definitely be abandoned."

Unconditional methods of DIF

Alternatively, one can conduct DIF without conditioning on the variable of interest. This unconditional method is, in essence, the item response theory (IRT) approach for DIF where item response models are based on mathematical functions that relate the probability of a particular response to an item to the level of the underlying trait or factor, θ, measured

by the scale. DIF occurs whenever the conditional probability, P(θ), of an item response differs between groups. This is identified by comparing item characteristic curves (ICC) created from the same item but computed from different groups. "Comparing the IRT parameter estimates or ICCs is an unconditional analysis because it implicitly assumes that the ability [latent trait] distribution has been 'integrated out'" (Zumbo & Hubley, 2003, p. 507). The focus is, therefore, on the area between the ICCs of the groups as opposed to the effect of the grouping variable over-and-above the conditioning variable. However, in order to accurately estimate the parameters (i.e., ability or difficulty parameters) required in IRT methods, it is well known that a substantial number of items are required because the unobserved continuous variable named theta, θ, is constructed and predicted from the joint relationship from other items. Although the measurement literature does not explicitly define how many items are required, examples in the literature suggest that more than 30 items are required. For example, Seong (1990) notes that 45 items are needed to yield stable parameter estimations. Similarly, Stone (1992) found that increasing test length from 10 to 40 items significantly reduced estimation errors and variability of the estimates. One exception to this limitation is nonparametric IRT (Ramsay, 2001; Zumbo & Witarsa, 2004) that does not impose a predefined parametric function, but rather lets the data speak for themselves, and thus, fewer items are needed. A second limitation of IRT models used to detect DIF is that they assume unidimensionality. Thus it is inappropriate to use IRT methods if the scale is multidimensional.

Multidimensional methods of DIF

The third latent variable DIF detection method is that of multidimensional models, which, as the name implies, assumes that all scales are to some extent multidimensional –

strict unidimensionality does not occur in practice. As stated in Zumbo and Hubley (2003), "the multidimensional approach to DIF, as implemented in SIBTEST [simultaneous item bias test], allows for a variety of scenarios that comprise differential dimensionality as the source for DIF" (p. 507). In general, the focus of this method is on item sets or item bundles, rather than on individual items as in the previously discussed DIF approaches. Again, however, this approach requires a large number of items to accurately estimate the latent variable, theta.

Structural Equation Modeling method of DIF (MIMIC model)

The fourth latent variable method for DIF is the structural equation modeling (SEM) approach first proposed by Muthén (1989). Originally, Muthén introduced this approach as a method for estimating the effects of student background variables on students' performance on each item of a test (what was referred to as item specific effects of background variables). This method was presented in the context of an example wherein Muthén modeled item-specific opportunity-to-learn (OTL) information (e.g., instructional coverage) from a sample of eight dichotomously scored algebra items from the Second International Mathematics Study (SIMS). OTL variables are assumed to be item-specific because different amounts or types of instructional coverage among students may affect how they respond to certain items whereas background variables such as gender and ethnicity are constant over the items. In other words, it is possible that having OTL improves the specific skills needed to correctly solve the corresponding item.

This method was illustrated using a SEM approach which allows the difficulty parameter for each item to vary with the level of the background variable (e.g., OTL) and, as a result, "item bias" (more accurately termed DIF) detection is obtained as a by-product. The "item bias" in the OTL context can be thought of as instructional sensitivity for each item. In

his 1989 paper, Muthén also demonstrated the weakness of traditional IRT methodology to assess measurement differences compared to this new SEM DIF methodology. Muthén (1989) reported that the newly proposed method "does not necessitate the traditional creation of groups to assess item bias and avoids the problem of the standard item bias detection approach" (p. 386). This is the case because the combination of covariates may, indirectly, represent group membership. Of course, this method can be used if the groups are known ahead of time, for example, gender DIF. It should be noted, however, that the SEM DIF method assumes the measurement model is the same in both groups (this is an implicit assumption in GLIM models, as well as conditional and unconditional DIF methods). One limitation of the SEM DIF method is that it does not test for interactions (i.e., non-uniform DIF); it only investigates uniform DIF. That is, the SEM DIF method only examines DIF that is attributable to differences in item difficulty (i.e., differences in thresholds).

Later, Muthén, Kao and Burstein (1991) applied this SEM technique using real data for the purpose of studying instructional sensitivity of achievement items. Specifically, data on 40 core items with a five category multiple-choice item format from the SIMS, which covered a variety of numeracy components (e.g., algebra, arithmetic, geometry), with 3724 eighth grade students in the United States was used. Multiple background variables were included such as variables that were assumed to influence all the items (e.g., math pretest scores, attitudes, family background, demographics, class type) and item-specific variables (e.g., OTL). To analyze the data, the LISCOMP computer program with unweighted least-squares estimation was used.

In a second example, Muthén and colleagues (Muthén, Tam, Muthén, Stolzenberg & Hollis, 1993) illustrated how the SEM DIF approach can be useful in studying invariance of

attitude measurements, specifically related to college students' career choice preferences. Unlike the application by Muthén, Kao and Burstein (1991) who used the SEM DIF approach with multiple-choice data (i.e., binary item format), this application applied the SEM DIF approach to categorical responses. The data for this application was based on a large-scale, longitudinal survey consisting of 2645 students. Eleven background variables (eight dichotomous and three continuous) were used as predictors to identify respondents' socioeconomic status, gender, race, and educational background and ten dependent variables were used that concerned respondents' career choice preferences. The LISCOMP software with the weighted least squares estimator was used to analyze the data (see Appendix A).

More recently, this SEM method of DIF has been used to investigate DIF in measures of functional disability (e.g., Fleishman, Spector & Altman, 2002; Mast & Lichtenberg, 2000), depression (e.g., Christensen et al., 1999; Gallo, Anthony & Muthén, 1994; Gallo, Rabins, & Anthony, 1999; Grayson, Mackinnon, Jorm, Creasey & Broe, 2000), cognitive functioning (e.g., Jones, 2003; Jones & Gallo, 2002), and general health status (e.g., Fleishman, 2004). A table detailing the construct under investigation, the number of items, the rating scale format, sample size, and the estimation methods used for all of these studies is included in Appendix A. As can be seen in Appendix A, given the increasing number of studies since 1989 that have used a SEM DIF method, this method is growing in popularity for short scales in the psychological, health, and sociological domains.

In addition, using the latent variable approach to DIF is in line with recommendations by Zwick (1990), Meredith (1993), Meredith and Millsap (1992), and Millsap and Meredith (1992), who argue that observed variable matching DIF methods like the Mantel-Haenszel and logistic regression are not generally diagnostic of item bias. That is, these observed

score matching variable DIF methods use the manifest matching variable as a proxy for the latent matching variable and hence will only be appropriate when the two (i.e., manifest and latent) correspond. This correspondence holds when the observed item responses are consistent with a Rasch (i.e., one-parameter logistic) item response theory model. This is the case because, under the Rasch model, the observed total score is a sufficient statistic for the latent variable score – hence assuring the correspondence between the observed and latent matching variables. Another situation where the observed and latent matching variables correspond is with long scales in which all of the items are strong indicators (i.e., high factor loadings) of one underlying latent variable (assuming a one-dimensional scale).

In general, the SEM DIF method conditions on the latent variable while simultaneously testing the effect of group membership (e.g., gender) over-and-above the underlying latent variable of interest. In essence, this is a multiple-indicators, multiple-causes (MIMIC) model which is akin to a latent variable ANCOVA[3]. The MIMIC model was first introduced by Jöreskog and Goldberger (1975) and contains one or more latent variables that, as the name implies, are simultaneously identified by both multiple endogenous item indicators (i.e., the items that comprise the scale under consideration) and by multiple exogenous causal variables (e.g., background variables such as gender or ethnicity). Accordingly, the MIMIC model allows the regression of latent variables on the background variables. Several uses of the MIMIC approach are described by Muthén (1989) and colleagues (e.g., Muthén, Tam, Muthén, Stolzenberg & Hollis, 1993). One advantage of this approach is that it involves the inclusion of multiple relevant background variables which allows one to study the relative importance of the predictors. Moreover, including multiple

[3] One can also think of this method as a simultaneous estimation method for the two-step conditional DIF method described above.

exogenous variables provides one with extra information about the measurement, which is particularly useful in detecting population heterogeneity (see Mast & Lichtenberg, 2000) and providing information to help validate scales (i.e., it is able to test the factor structure of a measure). In addition, the MIMIC approach allows for the detection of measurement non-invariance (i.e., DIF).

Muthén's (1989) modeling approach (i.e., the MIMIC model) can be graphically depicted using a path diagram similar to that shown in Figure 2.1. For ease of interpretation, Figure 2.1 can be thought of in the context of an example using a 10-item depression inventory. The model specification and symbolic language adopted in this dissertation follows that of Jöreskog and Sörbom (1996). As Figure 2.1 illustrates, the MIMIC model consists of three components: (1) a measurement model, (2) a regression model, and (3) a direct effects estimate. In general, the measurement component refers to the hypothesized relationship between a latent variable and its indicators. As shown on the right hand side of Figure 2.1, the *measurement model* relates the observed indicators represented by *y*s to the continuous latent variable, ETA-1 (η_1), representing 'depression'. Thus, the latent variable, η_1, is defined for this analysis by the 10 items that form the 10-item scale measuring depression. The relationship between the latent variable, η_1, and its indicators (i.e., factor loadings) are represented by LAMBDA (λ). Accordingly, these factor loadings[4] are associated with the endogenous measurement model and in Jöreskog's notation the matrix Lambda *Y*. In Figure 2.1, the factor loadings are represented by directional arrows that point from the latent conditioning variable to the 10 individual items. The measurement errors for the *indicators* of the endogenous variables (i.e., residuals) are denoted by EPSILON (ε) and

[4] In this model, the factor loadings are set free (i.e., they are unknown parameters to be estimated by the program).

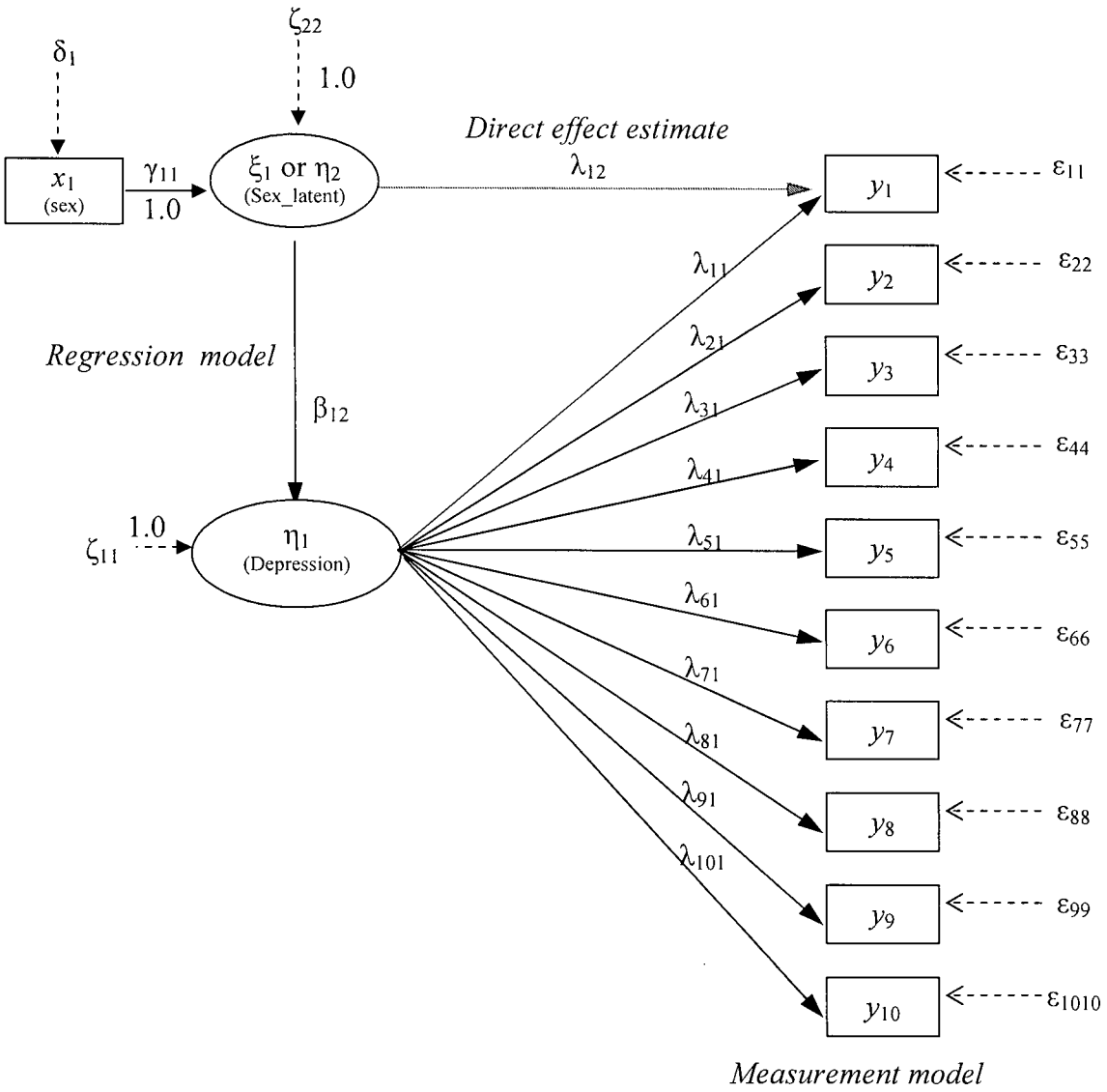
are set free in this model. Similarly, the measurement errors for the endogenous *latent* factors are denoted by ZETA (ζ_{11} and ζ_{22}) and in Jöreskog's notation are contained in the Psi matrix[5]. In general, the measurement model can be expressed as

$$y = \nu + \Lambda\eta + \varepsilon,$$

where y represents a vector of endogenous items (e.g., depression items), ν are intercepts associated with the endogenous items, Λ represents a matrix of factor loadings and therefore relates the observed indicators to the latent trait, η_1, and ε contains the error terms associated with each item.

[5] In this model, the diagonal of the Psi matrix is free and the other elements are fixed to 1.0.

Figure 2.1. Path diagram for the DIF MIMIC model for a 10-item scale.



On the left hand side, the *regression model* relates the latent variable 'depression' (η_1) to the latent covariate 'sex' (denoted by the Greek letter KSI, ξ)[6]. The effect of the grouping variable (which is assumed to influence the latent factor) on the underlying latent construct is represented by an arrow from the latent grouping variable (i.e., the covariate) to the latent variable 'depression'. This single directional relationship is identified by matrix BETA (β) and is set free in this model. This is analogous to regression of a continuous outcome variable onto one or more covariates (e.g., sex, gender, marital status, education level). For statistical modeling in the software LISREL, the grouping variable (x_1) is converted to a latent variable (ξ_1) by making a direct relational path from the observed grouping variable, denoted by the rectangle labeled 'Sex,' to its constructed latent variable, denoted by the oval labeled 'Sex Latent'. This path is identified by GAMMA (γ), of which matrix Gamma is fixed to the assigned value of 1.0 because these two variables are theoretically equivalent in this model; thus, there is a perfect directional relationship between them. The measurement error for the indicator of the exogenous variable is denoted by DELTA (δ), and thus the unique factors (residuals) for each exogenous variable are in matrix Theta-delta. Using Jöreskog's notation, the error covariance of the exogenous factor (ξ_1) is contained in the Phi matrix.

The interpretation of the regression coefficient for the grouping variable will depend on the coding. If, for example, the grouping variable denotes gender such that males are 0 and females are 1, a negative coefficient for the regression of the latent variable (e.g., depression) on gender would indicate that females have lower underlying depression than males.

---

[6] KSI-1 (ξ_1), representing the variable 'Sex latent', is also part of the endogenous measurement model, and thus is also called ETA-2 (η_2).

The regression model can be expressed as

$$\eta = \alpha + \beta\eta + \Gamma x + \zeta$$

where α are intercepts, β represent the effect of the latent variable, Γ contains the coefficients for the regression of η on x, and ζ are error terms. The error terms ε and ζ are assumed to be uncorrelated with each other and with η (Gallo et al., 1994).

The third component is a *direct effect estimate* that detects measurement invariance in an item response associated with group membership. In other words, DIF is incorporated by adding direct effects from the covariate(s) to the observed indicators, unmediated by the latent factor. This is shown in Figure 2.1 by a directional arrow pointing from the latent grouping variable 'sex latent' to the individual item being analyzed. This analysis is repeated for each individual item on the scale that one wishes to investigate DIF. Thus, in accordance with the definition of DIF, this path represents a systematic difference in responses, controlling for the latent variable.

To correctly read the subscripts associated with directional paths (i.e., those represented by the β, γ, and λ parameters), the first subscript references the target variable (i.e., the 'effect') and the second subscript references the source variable (i.e., the 'cause'). For example, λ_{41} tells us that y_4 is an indicator of the first endogenous latent variable, η_1. Likewise, β_{12} implies that the first endogenous variable, η_1, is directly influenced by the second endogenous variable, η_2. On the contrary, the subscript order is not important for non-directional linkages (e.g., double-headed arrows). The MIMIC model assumes that the endogenous items constitute an adequate model for a unidimensional construct of depression across gender.

Jöreskog's (2002) ML and WLS estimation methods for SEM

Using Muthén's (1989) model, Muthén, Kao, and Burstein (1991) used an unweighted least-squares (UWLS) estimation to "simplify the computations" (p. 11). Recently, however, Jöreskog (2002) proposed that either a weighted least squares (WLS) or a maximum likelihood (ML) estimation method be used for this sort of SEM. These estimation methods take into consideration that one or more ordinal variables are observed jointly with a covariate(s) (i.e., possible explanatory variables). Thus, Jöreskog's method of estimation for SEM is more appropriate for use with the MIMIC model wherein the explanatory variables (i.e., covariates) are assumed to affect the latent variable(s) that are indicated by other observed variables. The essence of the estimation problem comes down to constructing and estimating the correct covariance matrix of the grouping variable and item response variables for input into the structural equation model.

In order to understand the advantage of Jöreskog's (2002) estimation method, I will clarify the estimation problem. For ordered discrete response data (ordinal data) the proper correlation measure is a polychoric (tetrachoric if ordered binary) correlation. For metric data (interval or ratio) the proper correlation is a Pearson correlation. We also know from regression and correlation theory that for truly binary variables (e.g., gender) we can use the Pearson correlation and this models a difference in means for the continuous variables. That is, the binary vector or a design matrix represents the contrast among means. The method of estimation becomes a problem when one has a mix of ordinal and continuous data which are used to create the correlation matrix. Figure 2.2 illustrates this problem, in which items 1 through 3 are 4-point ordered discrete response categories, and the variables 'age' and 'height' are continuous (truly discrete binary variables such as gender are also treated as

continuous). The correct correlation between the items in Figure 2.2 (e.g., item1 and item2) is a polychoric correlation (ordinal:ordinal). Similarly, the correct correlation between the continuous variables 'age' and 'height' is a Pearson correlation (continuous: continuous). However, the correlation between an ordinal variable (e.g., item1) and a continuous variable (e.g., age) is problematic because of their different variable formats.

Figure 2.2. Example of a correlation matrix with mixed variable formats.

	Item 1	Item 2	Item 3	Age	Height
Item 1		ordinal: ordinal	ordinal: ordinal	ordinal: continuous	ordinal: continuous
Item 2			ordinal: ordinal	ordinal: continuous	ordinal: continuous
Item 3				ordinal: continuous	ordinal: continuous
Age					continuous: continuous
Height					

If the data contains mixed variable formats, as is the case shown in Figure 2.2 between the ordinal and continuous variables, and a Pearson correlation matrix is used, it will treat the ordinal item responses as interval or ratio, resulting in incorrect attenuated correlation values as illustrated in the shaded area of Figure 2.3. This type of measurement error caused by using Pearson's correlation with ordinal data (e.g., Likert-type response formats) has long been debated in the literature (e.g., O'Brian, 1979; Bollen & Barb, 1981). As cited by Byrne (1998), Jöreskog and Sörbom (1993) noted that when the observed variables in SEM analyses are either all ordinal or a combination of ordinal and metric scales,

the analyses should be *not* be based on Pearson product-moment correlation, but rather be based on either polychoric or polyserial correlations.

Figure 2.3. A Pearson correlation matrix would incorrectly treat the ordinal items as interval or ratio.

	Item 1	Item 2	Item 3	Age	Height
Item 1					
Item 2					
Item 3					
Gender					
Height					

On the other hand, if a polychoric (or tetrachoric for ordered binary) correlation matrix is used when data are of mixed formats, the continuous variables will be treated as ordinal, which they are not. The resulting correlation values will be incorrect (illustrated in the shaded area of Figure 2.4).

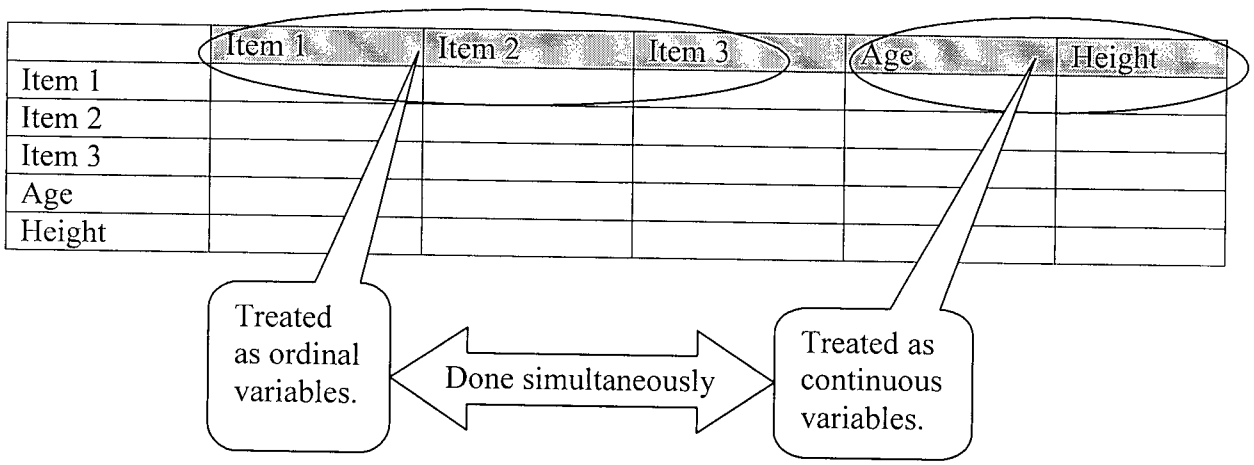
Figure 2.4. A polychoric correlation matrix would incorrectly treat the continuous variables as ordinal.

	Item 1	Item 2	Item 3	Age	Height
Item 1					
Item 2					
Item 3					
Age					
Height					

Jöreskog's (2002) new method, however, *correctly* treats the variables according to their variable type (see Figure 2.5). The ordinal item responses (items 1 through 3 in Figure 2.5) are correctly treated as ordinal variables, and the 'age' and 'height' variables are correctly treated as continuous. This estimation method allows one to compute the joint covariance matrix of the predictor and the variables underlying each of the ordinal variables (this is done simultaneously). That is, given that one or more ordinal item response variables

are jointly observed with one or more manifest (i.e., observed) variables (e.g., gender) that can be treated as covariates or predictor variables, one can estimate the effect of the predictor variables on the probability of responding to the ordered categorical (ordinal) variables using either a logistic or probit model. In addition, one can compute the joint covariance matrix of the predictor and the variables underlying each of the ordinal variables. This covariance matrix can then be used as input for any structural equation modeling and ML or WLS estimation can be correctly applied.

Figure 2.5. Jöreskog's estimation method correctly treats the variables according to their variable type.



The statistical test of DIF is examined via (a) the *t*-statistic, or (b) a Chi-squared difference test of the two models, wherein the nominal alpha of .05 is used in the test for DIF. Having described Muthén's DIF model and Jöreskog's estimation strategy, let us now turn to a major advantage of the SEM DIF approach for short psychological scales.

DIF for short scales

The statistical methods developed for analyzing DIF have primarily focused on educational ability and achievement tests that are typically quite long (i.e., tests containing many items). As a result, most DIF methods require tests that contain many items (e.g., greater than 30) for the results to be reliable (e.g., Fidalgo, Mellenbergh, & Muñiz, 2000) and hence to meet the Meredith and Millsap (e.g., Meredith, 1993; Meredith & Millsap, 1992; Millsap & Meredith, 1992) conditions described above. Psychological scales (e.g., Rosenberg's Self-Esteem Scale (RSE; Rosenberg, 1965), Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977)), however, can have very few items, ranging from 3 to 30 items. Unfortunately there is very little documentation in the educational, psychological, and measurement literature that attempts to define the number of items in a scale for it to be considered short, moderate, or long.

For example, Fidalog, Mellenbergh and Muñiz (2000) suggest that a test length of 60 items is long, 40 items is moderate, and 20 items is short. Uttaro and Millsap (1994) also consider 40 items and 20 items as moderate and short test lengths, respectively. Likewise, Scholte and De Bruyn (2001) consider the 89-item Revised Junior Eysenck Personality Questionnaire (JEPQ-R: Corulla, 1990) a relatively long test, the 48-item version of the JEPQR-S (Corulla, 1990) "too long in specific research situations," and thus prefer the 24-item abbreviated version (JEPQR-A: Francis, 1996) thereby implying that 24 items is considered short. Similarly, Fossum (2002) considers the 30-item simple rating scale for personality a "brief measure" of the five-factor model of personality and Swaminathan and Gifford (1983) consider 10, 15, and 20 item tests short. Lu, Thomas, and Zumbo (in press) refer to "large numbers of test items" (p. 18) as greater than 30. Other psychological

researchers such as Burisch (1997) suggest that extremely short scales have two to four items each and he considers eight and nine item depression scales as "relatively short" compared to 28 and 50 item depression scales; and scales with 43-88 items as "very long." Moreover, in simulation study on the robustness of item parameter estimation, Kirisci, Hsu and Yu (2001) state that 40 items is a long test. In an article addressing strategies for reducing the length of self-report scales, Stanton, Sinar, Balzer and Smith use a 72-item job satisfaction measure as an example of a large scale that could be reduced in length. Lastly, in a systematic review based on two educational measurement journals and two educational psychology journals from 1998 to 2002, Witarsa (2003) found that achievement testing research, in general, used larger numbers of items (ranging from 20 to 150) than psychological (ranging from 10 to 56 items) and survey (ranging from 9 to 80 items) research. Based on the above studies and from a further literature review of educational and psychological articles in PsycINFO, this author has concluded that short scales range from 3 to 30 items.

Unfortunately, reliability decreases with shorter scales and hence measurement error increases. Therefore, methods like LogR that match on the observed score, which has measurement error, are of particular concern in short scales because of the lower reliability. Thus, a latent variable approach for investigating DIF with short scales is more appropriate compared to an observed score approach because one can condition on the "measurement error free" latent variable. Specifically, the SEM MIMIC method with Jöreskog's (2002) estimation methods is most appropriate because it models the covariance structure thereby avoiding the problems (e.g., mixed response data formats, large numbers of items) associated with estimating a theta score. Accordingly, fewer items are required compared to the

conditional[7], parametric IRT, unconditional parametric IRT, and multidimensional modeling approaches that make the DIF MIMIC method ideal for short scales.

Example to motivate the problem

Recently, Gelin and Zumbo (2003) investigated gender DIF in the Center for Epidemiologic Studies Depression scale (CES-D; Radloff, 1977). The CES-D is a widely used 20 item self-report measure developed for use in studies exploring the epidemiology of depressive symptomology in the general population. Each item is rated on a four-point (0 - 3) Likert-type scale of which a total scale score is computed from the sum of the 20 items (ranging from 0 - 60). Six hundred community-dwelling adults living in northern British Columbia (290 females; 310 males) were included in the study[8]. The mean age of female participants was 42 years (SD = 13.4, range = 18 to 87 years), and the mean age of male participants was 46 years (SD = 12.1, range = 17 to 82 years). Using Zumbo's (1999) ordinal logistic regression method with corresponding logistic regression effect size estimator, the CES-D Item 17 '*crying*' showed large uniform gender DIF (DIF R^2 = .218) in which females were over nine times likely to score higher on this item than males based on the odds ratio (Gelin & Zumbo, 2003).

Using the same data as Gelin and Zumbo (2003), gender DIF in the CES-D was investigated using the SEM DIF methodology with Jöreskog's ML estimation method. Item-by-item DIF results were examined and seven items were found to exhibit DIF. Similar to regression, the 1 degree of freedom omnibus test can be characterized as a *t* statistic in

---

[7] As Lu, Thomas and Zumbo (in press) show, two-step IRT conditional methods are not very efficient because they require one to estimate the factor score separately.

[8] The item response data came from the Health and Health Care Survey carried out by the Institute for Social Research and Evaluation in the fall of 1998.

LISREL. The seven items displaying DIF were: Item 2 (t = -4.33), Item 7 (t = 3.38), Item 8 (t = 3.18), Item 12 (t = 3.13), Item 16 (t = 2.99), Item 17 (t = -9.54), and Item 18 (t = -3.05). The sign of the coefficient indicates the direction of DIF, wherein a negative t-value indicates that females endorse the item more often than males. See Appendix B for a list of the items in the CES-D.

Comparing the results from Gelin and Zumbo (2003) which used the ordinal logistic regression approach to the results from the SEM DIF methodology, only Item 17 (*crying*) showed DIF using both methodologies. However, the SEM DIF method identified six additional items that showed DIF compared to the ordinal logistic regression approach. Given that the SEM DIF method found more items displaying DIF than the ordinal logistic regression approach the question arose as to whether or not the items flagged as DIF using the SEM DIF method were due to an inflated Type I error rate, or is the Type I error rate within the nominal range and the SEM DIF is more statistically powerful. Accordingly, a study investigating the Type I error rates (and if the Type I error rates are valid, an investigation of the statistical power) of the SEM DIF method is needed.

Research purpose

Given that (a) short scales are typically found in the educational and psychological disciplines, (b) the MIMIC method is the most appropriate method for investigating DIF in short scales, and (c) the increasing number of published articles using the MIMIC method suggests this approach is growing in popularity, the purpose of this dissertation is to present new research that examines the Type I error rate of a DIF MIMIC method. Type I error rates are often referred to as operating characteristics of a test. The proposed MIMIC methodology uses Muthén's (1989) SEM model computed via Jöreskog's (2002) ML and WLS estimation

methods. To this author's knowledge, the Type I error rate of this DIF approach hs not been investigated. Thus, the primary focus of this dissertation is to examine the Type I error rate of the proposed MIMIC approach under a variety of study conditions.

It is important that a statistical test maintain its Type I error rate to be a valid test of the hypothesis. A Type I error rate (i.e., probability of rejecting H_0 when in fact it is true) in detecting DIF refers to declaring an item as DIF when it is not a DIF item. Once the statistical null hypothesis is rejected and one concludes that an item functions differentially for different groups, further evaluation of the item is necessary in order to determine whether the DIF is attributable to item bias or item impact. In the context of high stakes testing, making a Type I error may be of great concern because of the matter of test fairness. That is, the Type I error rate is not only important for statistical reasons but it is also important in terms of the decisions being made about items flagged as showing DIF. As a result, the empirical Type I error rate of the DIF MIMIC model must be explored. If the Type I error rate is found to be within reason (e.g., 0.05; Bradley, 1978), the power of the DIF MIMIC model needs to be examined (i.e., power is not formally defined unless the statistical test protects the Type I error rate).

The Type I error rates of this proposed MIMIC methodology will be explored by means of a simulation study using data that are reflective of short scales with ordinal item response formats typically found in psychological measures. The simulation study will investigate the Type I error rates of Muthén's (1989) SEM model for a single-factor model under a variety of conditions including use of Jöreskog's (2002) ML and WLS estimation methods. A single-factor model was chosen because the majority of educational and psychological measures assess unidimensional latent constructs or unidimensional sub-scale

scores (e.g., Minnesota Multiphasic Personality Inventory-2 [MMPI-2] clinical scales; Butcher, Dahlstrom, Graham, Tellegen & Kaemmer, 1989). Thus, this study design represents a commonly found structural model (i.e., single-factor model) in short scales.

Jöreskog's (2002) method will be explored using the software program LISREL because it is the most widely used software for SEM in the social sciences, and it is also the oldest widely available software for conducting SEM. It should be noted, however, that the DIF MIMIC model is mostly being used with Muthén's software program MPlus (previously called LISCOMP). The reason the literature (to date) applies the DIF MIMIC approach with Muthén's software is because Jöreskog's approach is relatively new (2002). It is expected, however, that the DIF MIMIC model applied with LISREL will start becoming more prevalent in the literature as Jöreskog's approach becomes recognized and as articles using his method become published.

CHAPTER III: METHODOLOGY

This chapter will discuss the methodology used for the DIF MIMIC simulation study. Monte Carlo methods are used to examine the Type I error rates of the proposed DIF MIMIC methodology for a single-factor model computed via Jöreskog's (2002) ML and WLS estimation methods. The Type I error rates are investigated under various conditions designed to reflect real responses to short scales in the social and behavioral sciences. The factors included in this study are chosen based on simulation designs seen in the literature (as described below) as well as on real response data using the 10 and 20 item versions of the Center for Epidemiologic Studies Depression (CES-D) scale. Data from the CES-D scale is chosen because it is a commonly used measure and hence is reflective of measures used in the social and behavioral sciences. As a demonstration of the widespread use of the CES-D, PsycInfo, a computerized database produced by the American Psychological Association, located 291 records in which the search criteria were set to publications between the years 2000 and 2004 with keywords "CESD" or "CES-D" in the default fields. Moreover, the item characteristics are representative of data typically found in psychological measures. These item characteristics, such as the scale length and item format, are described below. In addition, empirical findings suggest that the 10 item short form (CESD-10: Andresen, Malmgren, Carter, Patrick, 1994) and the original 20 item (CESD-20: Radloff, 1977) CES-D scales are essentially unidimensional (e.g., Clark, Aneshensel, Frerichs & Morgan, 1981; Hertzog, Van Alstine, Usala, Hultsch & Dixon, 1990; Sheehan, Fifield, Reisine & Tennen, 1995; Zumbo, Gelin & Hubley, 2002), supporting the use of a single-factor model with both test lengths for this simulation.

In brief, the variables in this simulation study are seven sample size combinations (three equal and four unequal group combinations), two item response distributions (normal and positively skewed), two scale lengths (10 and 20 items per scale), and two estimation methods (Jöreskog's 2002 ML and WLS). In addition, all of the items are ordinal (i.e., Likert or rating scale format) with four response categories (i.e., four point response scale). For ease of interpretation, this simulation study is divided into two sub-studies. The first sub-study (Part A) investigates the Type I error rates in which two groups have *equal* sample sizes (e.g., 200 simulees per group). The second sub-study (Part B) investigates the Type I error rates in which two groups have *unequal* sample sizes (e.g., 200 simulees in one group and 800 simulees in the second group). As a result, the first sub-study (Part A) has a 2 x 2 x 2 x 3 factorial design: two estimation methods by two item response distributions by two scale lengths by three sample size combinations. Similarly, the second sub-study has a 2 x 2 x 2 x 4 factorial design, of which the variables are the same as in Part A except there are four sample size combinations instead of three. Given that the simulation methodology is the same for both sub-studies, only the results section of this simulation study will be divided into the sub-studies. See Appendix C for a visual representation of the study design.

Method

The following section describes the general methodology used for the simulation study. To provide a realistic set of values within the various study design variables described below in the simulation study, real item response data from the CESD-20 was used. This data was the same data as that presented in the example used to motivate the problem, and was therefore discussed above. This same item response data was also used to represent the short 10 item CES-D scale. Specifically, as described below, 10 items were dropped from the

CESD-20 scale, and the remaining 10 items were used as data that represents the CESD-10 (see Appendix B).

Study design

Scale length and item format

Consistent with the CESD-10 and CESD-20 scales, data are simulated to represent 10 and 20 item scales, respectively. These two scale lengths are also chosen because they are representative of numerous short scales typically found in the social and behavioral sciences. Conducting a PsycINFO search for the keyword "10 item" in the abstract and publication year 2000-2004 produced 231 records, of which the keyword "10 item" referred to 10 items in a scale, measure, quiz, inventory, index, questionnaire, test, and instrument. Examples of scales with 10 items include the Rosenberg Self-Esteem Scale (RSE: Rosenberg, 1965), Alcohol Use Disorders Identification Test (AUDIT: Babor, De la Fuente, Saunders & Grant, 1992), and the Multidimensional Anxiety Scale for Children (MASC-10: March, 1997). In support of the common use of 20 items in a scale, a PsycINFO search was conducted for the keyword "20 item" in the abstract and publication year 2000-2004; this search produced 287 records. Examples of scales that use 20 items include the Beck Hopelessness Scale (BHS: Beck, Weissman, Lester & Trexler, 1974), Child Anxiety Scale (CAS: Gillis, 1980), and the Toronto Alexithymia[9] Scale (TAS-20: Bagby, Parker & Taylor, 1994; Bagby, Taylor & Parker, 1994).

---

[9] Alexithymia is a personality construct that is characterized by the inability to identify, express, and discriminate among emotions.

As found in the CES-D scales, all items are simulated to represent ordered categorical data with four categories. This number of rating scale points is also representative of item response formats typically encountered in psychological measures. Using PsycINFO, a search for publications between the years 2000 to 2004 in which the abstract contained a keyword relating to a 4-point item format produced a number of records. Specifically, seven records were located when the keyword was "4 point rating," thirteen records were located when the keyword was "4 point Likert," and 36 records were located when the keyword was "4 point scale." Such scales include the RSE scale (Rosenberg, 1965) in which all 10 items are answered on a 4-point Likert response scale ranging from 1 (strongly disagree) to 4 (strongly agree). Likewise, the 12 item General Health Questionnaire (GHQ12; Goldberg & Williams, 1988) asks respondents about changes in normal functioning such as changes in their general level of happiness, anxiety, self-confidence, depression, and stress on a 4-point Likert-type scale (i.e., 1 = not at all, 2 = no more than usual, 3 = rather more than usual, and 4 = much more than usual).

Ordinal variables are commonly referred to as "rating scale," or "Likert" variables, and thus these terms will be used interchangeably. As in numerous psychological, educational, and behavioural sciences, the ordinal variables used in this dissertation are conceptualized as observed ordered-categorical variables, y, wherein the underlying variable, y^*, is completely unobserved (i.e., latent) and continuous. As the normally distributed latent variable increases beyond certain threshold values, the observed variable takes on higher scores, referred to as scale points. Thus, a person endorsing one category has more of a characteristic than if he/she had chosen a lower category, but we do not know how much more.

Item response distribution

Following the simulation study by DiStefano (2002), two distributions are investigated: approximately normally distributed and non-normally distributed. To approximate Likert-type data with four ordered response categories, the generated continuous data are divided using three threshold values.

For the normal (symmetric) distribution, the three equal interval cut points (i.e., thresholds) used to categorize the continuous data into four ordered categories are chosen in accordance with the area under the normal curve. For the ordered categories 1 through 4, the item response thresholds (-1.67, 0, and 1.67) corresponded to approximately 5%, 45%, 45%, and 5% of the area under the normal curve. A check on the generated item-level characteristics revealed that the population data (i.e., all items for both the 10 and 20 item scales) are approximately normally distributed for both groups (Skewness $\approx$ 0; Kurtosis $\approx$ -0.2).

To determine the effect of skewness of the item response distribution on the DIF MIMIC method, the generated continuous data are divided into non-normally distributed four-category ordered categorical data with a targeted skewness of 1.7. This skewness level is chosen based on data from the CESD-20 in which skewness values ranged from 0.64 to 3.1, with an average positive skew of 1.7. This type (i.e., positive) and magnitude of skewness is also consistent with item characteristics of other psychological measures (e.g., Golding, 1988; Micceri, 1989; Olsson, 1979) and with other simulation studies (e.g., Babakus, Ferguson & Jöreskog, 1987). To create skewed ordered categorical data, the percentage of responses in each category is approximately 66, 22, 7, and 5 under the normal curve (as determined from real data using the CESD-20) for response categories 1 though 4,

respectively (thresholds = 0.4, 1.16, 1.65). A check on the generated item-level data for both the 10 and 20 item scales show skewness and kurtosis values close to the target levels for both groups in the population data (skewness ≈ 1.6, kurtosis ≈ 1.8).

Sample size combinations

To inform the choice of sample size, a review of the literature was conducted. Building on simulation designs seen in the literature (e.g., De Champlain & Gessaroli, 1998; Curran, Bollen, Paxton, Kirby, & Chen, 2002; Muñiz, Hambleton, & Xing, 2001; Muthén & Kaplan, 1992), as well as from published literature using the CES-D between 2000 and 2004 (PsycINFO search), seven combinations of equal and unequal sample sizes are considered.

As previously mentioned, the first sub-study investigates the Type I error rates of the DIF MIMIC model when two groups have *equal* sample sizes. The equal sample size combinations included 1000, 500, and 200 simulees per group. The second sub-study investigates the Type I error rates in when the two groups have *unequal* sample sizes. For this sub-study, a total sample size of 1000 is used to avoid the problem of confounding the sample size with the per group size. By controlling the total sample size to be 1000 allows for the investigation of whether the Type I error rates are affected by differences in group sizes; if the total sample size was not held constant it would be difficult to distinguish whether or not the Type I error rate was affected by the difference in group sizes or the total sample size. Using a sample size of 1000, four different ratios are considered: 1:9, 2:8, 3:7, and 4:6. These ratios represent the size of Group 1 compared to the size of Group 2. For example, the ratio 1:9 indicates that there are 100 simulees in Group 1 and 900 simulees in Group 2. Overall, these sample size combinations reflect the range of sample sizes used in psychological and educational research (i.e., moderate-to-small-scale testing).

Estimation methods

Given that (i) the primary focus of this dissertation is on short scales that are typically found in the educational and psychological disciplines of which often contain ordinal item formats (e.g., 4-point scale) and (ii) DIF often involves a truly binary variables (e.g., gender), Jöreskog's (2002) ML[10] and WLS estimation methods will be used. As previously described, Jöreskog's estimation methods were chosen because the LISREL software is widely used and it *correctly* treats the variables according to their variable type thereby allowing one to compute the joint covariance matrix of the predictor and the variables underlying each of the ordinal variables. In turn, this covariance matrix can then be used as input for any structural equation modeling and ML or WLS estimation can be correctly applied.

Procedure / data generation

The study is carried out in nine steps:

1. *Create a population covariance matrix that will be used to generate continuous item response variables.* A population covariance matrix, Σ, as $\Sigma(y^*)_g = \Lambda_g \Phi_g \Lambda_g' + \Theta_g$ for two subgroups is created from pre-specified factor loadings. Unlike some simulation studies in which researchers choose factor loadings arbitrarily, the factor loadings (i.e., lambdas) from real data were used to reflect the range of item loadings commonly encountered in practice. These factor loadings were then used to specify the population covariance matrices[11]. These factor loadings were obtained by first reading real data describe earlier into the software program PRELIS 2.51 program

[10] The ML reported in this dissertation involves the asymptotic covariance matrix. Jöreskog and colleagues refer to this as Robust ML, however, I will refer to it as ML throughout.
[11] All data are analyzed using the MINRES factor analysis procedure with the polychoric correlation matrix with one extracted factor in LISREL.

(Jöreskog & Sörbom, 2003a), and then extracting one factor using the MINRES estimation with a tetrachoric correlation matrix (See Appendix B for factor loadings as they correspond to the scale items).

I. For simulating the 10 item scale, the real item response data from the CESD-10 is used and the resulting factor loadings are obtained: 0.669, 0.744, 0.857, 0.743, 0.532, 0.653, 0.597, 0.680, 0.658, and 0.775.

II. For simulating the 20 item scale, the real item response data from the CESD-20 is used and the resulting factor loadings are: 0.698, 0.533, 0.918, 0.462, 0.692, 0.856, 0.697, 0.554, 0.751, 0.658, 0.584, 0.708, 0.671, 0.713, 0.505, 0.749, 0.729, 0.853, 0.605, and 0.734.

In general, the factor loadings range from 0.53 to 0.86 and 0.46 to 0.92 for the 10 and 20 item scales, respectively. This range of loadings, from moderate to high, are representative of those found in the majority of published unidimensional CFA simulation studies (see Koh, 2003, for details). In addition, simulating data to represent a unidimensional scale is reflective of the majority of psychological measurement instruments that are typically single-factor models, as demonstrated by the fact that composite (i.e., total) scale scores are commonly computed from item responses (similar to the CES-D scales).

2. *Generate continuous item response data for two groups.* Continuous item response data, y^*, for the simulation is generated as described in the steps below:

I. Using the factor loadings, the population correlation matrix among the variables involved in the factor loadings is created for one group (i.e., *Group 1*).

II. Next, the item response data, of a specified population size, with normally distributed but independent (i.e., uncorrelated) continuous scores are generated for Group 1[12]. If the data is to be used for an equal sample size condition, the specified sample size is 50 000 for Group 1. However, if the data is for one of the unequal sample size conditions, the specified sample sizes for Group 1 are either 10 000, 20 000, 30 000, or 40 000 which correspond to the data with sample size ratios of 1:9, 2:8, 3:7, and 4:6, respectively. These normally distributed scores represent the (typically unobserved) latent scores from which ordered responses are generated.

III. As is customary in simulation studies, the generated continuous data are divided into four ordered categories by using three thresholds. Thus, the ordered responses are computed by recoding the continuous item response data into the appropriate thresholds for a 4-point scale: the thresholds for the symmetric data (i.e., equal latent thresholds) are -1.67, 0, and 1.67, and the thresholds for the skewed data (i.e., unequal latent thresholds) are 0.4, 1.16, and 1.65. Thus, the continuous scores are manipulated to mimic responses on a rating scale while simultaneously modifying the distributional shape of the data.

IV. This data, which represents Group 1, is saved as a data file.

V. As in Step I, the population correlation matrix is created for *Group 2*.

VI. Step II is repeated wherein the specified sample size for Group 2 is 50 000 for data representing an equal sample size condition. Conversely, the specified sample size for data representing the unequal sample size conditions with ratios

[12] A grouping variable is created and saved in the data set.

of 1:9, 2:8, 3:7, and 4:6 are 90 000, 80 000, 70 000, and 60 000, respectively, for Group 2.

VII. Step III is repeated for Group 2 data.

VIII. The data, which represent Group 2, is saved as a data file.

IX. The data from Group 1 is appended to the data from Group 2 to create a population data set with a total of 100,000 simulees for the appropriate design cell.

Given the simulation design, there are four population data sets created for the *equal* sample size conditions (i.e., Part A): (i) 10-item symmetric, (ii) 10-item skewed, (iii) 20-item symmetric, and (iv) 20-item skewed. Likewise, there are 16 population data sets created for the *unequal* sample size conditions (Part B):

Ratio 1:9	Ratio 2:8	Ratio 3:7	Ratio 4:6
• 10-item symmetric	• 10-item symmetric	• 10-item symmetric	• 10-item symmetric
• 10-item skewed	• 10-item skewed	• 10-item skewed	• 10-item skewed
• 20-item symmetric	• 20-item symmetric	• 20-item symmetric	• 20-item symmetric
• 20-item skewed	• 20-item skewed	• 20-item skewed	• 20-item skewed

3. *Importing data into PRELIS.* The simulated categorical population data is imported into the PRELIS software program. Using PRELIS, the covariate (i.e., grouping variable) is declared continuous, and the ordinal variables are declared ordinal. The data are saved as a PRELIS data file. This is repeated for all the population data sets as described in the previous step.

4. *Create and save a data file containing all of the bootstrap data.*

I. *Choose sample sizes.*

i. For the first sub-study (Part A), which investigates the *equal* sample size condition, the sample size is set using the sampling fraction PRELIS syntax command.

ii. For Part B, the sub-study investigating the *unequal* sample size combinations, the sampling fraction syntax command is set to 1000 because the ratio of simulees in each group is already represented in the generated population data.

II. *Bootstrap the data.* In accordance with recommendations in SEM Monte Carlo studies (e.g., Skrondal, 2000), 1000 replications are conducted for each study condition to improve the precision of the results. The bootstrapping procedure in PRELIS is used to generate 1000 replications and this data is saved as a data (*.DAT) file. Although the term "bootstrap" is used, this is really a Monte Carlo method because samples are generated from an "assumed" true model in the population.

5. *Compute and save the polychoric and asymptotic covariance matrices.* For each of the 1000 random samples, the joint covariance matrix (CM) of the variables underlying the ordinal variables and the covariate (e.g., grouping variable) is estimated and saved in a stacked data file (CM=BOOT.CM)[13]. As stated by Jöreskog (2002), the covariance matrix is an "unconstrained covariance matrix just as a sample covariance matrix for continuous variables" (p. 16). In addition, the corresponding

[13] The Fixed variables (FI) command is used in PRELIS to denote the appropriate variable that represents the covariate (e.g., variable 11).

asymptotic covariance matrix (AC) is computed and saved for each of the random samples in a stacked data file (AC=BOOT.ACC).

6. *No-DIF MIMIC model.* Using LISREL 8.54 (Jöreskog & Sörbom, 2003b), the *no-DIF* MIMIC model with Jöreskog's (2002) ML or WLS estimation method (using the asymptotic covariance matrix of the joint unconditional covariance matrix) is modeled for each cell in the design, for all 1000 replications. Using the saved matrices from the previous step, LISREL is used to run the MIMIC method as modeled with *no group to item path* (i.e., no direct effects estimate). The t-values and goodness-of-fit values are saved as text files.
7. *DIF MIMIC model.* Using the saved matrices from Step 5, LISREL is used to run the DIF MIMIC method as modeled *with the group to item path* (as illustrated previously in Figure 2.1). This path represents a systematic difference in item responses, controlling for the latent variable – hence it is the path that models DIF. The t-values and goodness-of-fit index values are saved as text files.
8. *Assessing Model Fit.* In accordance with McDonald and Ho's (2002) recommendations as well as their choice of goodness of fit indices in their summary table for a path model fit, the Chi-squared value and root mean square of approximation (RMSEA; Steiger & Lind, 1980) are used to assess model fit. Moreover, the χ^2 value is presented in accordance with conventional practice in which applied researchers often assess fit using this test statistic. In terms of fit, a nonsignificant Chi-squared value (χ^2) is an indication that the model "fits" the data. That is, a nonsignificant χ^2 indicates "there is no significant discrepancy between the covariance matrix implied by the model and the population covariance matrix"

(Kelloway, 1998, p.25). In terms of model fit using the RMSEA, Steiger (1990) suggests that a RMSEA value below 0.10 indicates a good fit to the data, and values below 0.05 indicate a very good fit to the data. Using the RMSEA as a fit index in MIMIC models is also in line with findings by Yu (2002).

9. *Evaluating the DIF MIMIC model for Type I error rates.* From the saved LISREL output in Steps 7 and 8, the goodness-of-fit indices and *t*-values are used to evaluate the DIF MIMIC model. For each combination of conditions described in the study design, the Type I error rates for the DIF MIMIC model are analyzed using the mean rejection rates of the models. Type I error is defined as the proportion of times that a null-DIF item was falsely rejected at the 0.05 level. In other words, the empirical Type I error rates are computed as the number of rejections divided by 1000 replications.

   Based on Bradley's (1978)[14] liberal criteria, an empirical Type I error rate exceeding 7.5% (i.e., > 0.075 level of significance) will be considered to be inflated. Bradley's liberal criterion for robustness of validity requires Type I error values of p to lie between 0.025 and 0.075. Note that both the *t*-test and the Chi-squared tests are investigated. The Chi-square test is a more general (i.e., omnibus) test that can be used to test several items at a time, whereas the *t*-test (t-value) is a one-degree of freedom test and can therefore only test one item at a time. In this case, however, because there is a large number of degrees of freedom the t-statistic "operates as a z-

[14] Bradley's (1978) moderate criterion for robustness of validity requires Type I error values of p to lie between 0.04 and 0.06; whereas his fairly stringent criterion requires values of p to lie between 0.045 and 0.055.

statistic in testing that the estimate is statistically different from zero" (Byrne, 1998, p. 104).

CHAPTER IV: RESULTS

This chapter discusses the results of the DIF MIMIC simulation study.

Psychometric properties of the population data

Before sampling from the population data files it is important to verify that the simulated data has the desired psychometric properties.

Unidimensionality

Given that the data for the simulation is mimicking real response data (i.e., the factor loadings from the real data were used) using the 10 and 20 item CES-D scales, a confirmatory factor analysis (CFA) was computed to confirm the data fit a unidimensional model. It was assumed that if the real response data fit the model well, then the simulated data would fit the same model because the simulated data were derived directly from the real data. Thus, a confirmatory factor analysis, computed using LISREL 8.54 (Jöreskog & Sörbom, 2003b) was used to support the hypotheses that the CESD-20 and CESD-10 were unidimensional (i.e., general depression latent variable underlying responses to the items). Both scales were hypothesized to be unidimensional because the majority of studies have used a single, summated score from the scales to measure depressive symptomology. Furthermore, a single factor model was implied by the common use of the cut-off criterion score of 16 on the summated score from the CESD-20 to indicate "case" depression (Radloff, 1977). Moreover, the majority of studies have consistently shown moderate to high correlations between item pairs, which suggest, "a single underlying theoretical variable may be responsible for these correlations" (Sheehan, Fifield, Reisine & Tennen, 1995, p. 509).

Given the item rating format of the CES-D (i.e., a four-point ordered scale), a polychoric correlation matrix was used so that the correct standard errors were produced (Zumbo, Gelin & Hubley, 2002). Moreover, given that the model was based on categorical data, the estimation of parameters was determined using the weighted least squares (WLS) estimation procedure with the asymptotic covariance matrix (Byrne, 1998)[15]. Accordingly, the input data were in the form of two matrices – the polychoric correlation matrix (PM) and the asymptotic covariance matrix (AC). Using a WLS estimation method in LISREL required a two step process in which both the PM and AC matrices were initially computed and saved using PRELIS, and then the CFA analyses were conducted. The goodness-of-fit statistics suggest that both the 10 ($\chi^2(35) = 110.82$, RMSEA = .06) and 20 ($\chi^2(170) = 442.47$, RMSEA = .052) item one-factor models have a reasonable fit to the data.

Reliability of the population data

As mentioned in the methods section, different population data sets were created for the equal and unequal sample size conditions. Four population data sets[16] were created for the *equal* sample size conditions (i.e., Part A). For each of these population data files, the reliability, as computed using alpha, was as follows: the 10-item symmetric data $\alpha = .86$, the 10-item skewed data $\alpha = .85$, the 20-item symmetric data $\alpha = .92$, and the 20-item skewed data $\alpha = .92$. As expected, the longer scales (i.e., the 20-item scales) had better reliabilities.

---

[15] "The generally weighted least squares method is asymptotically distribution free, yielding more accurate estimates of standard errors and model fit than the maximum likelihood techniques" (Wong, 2000, p. 73).

[16] Two levels of the number of items in the scale by two levels of item distributions.

Likewise, there were 16 population data sets[17] created for the *unequal* sample size conditions (Part B). It was necessary to create 16 population data sets for the unequal sample size conditions because the ratio of group sizes needed to be incorporated in the population data structure (see Appendix C). Again, the shorter scales (i.e., 10-item scales) had lower reliabilities than the longer 20-item scales. The four 10-item symmetric and skewed population data files (each file had a different sample size ratio) had reliabilities equal to .86 and .85, respectively. Similarly, the 20-item symmetric and skewed population data files had reliabilities equal to .92.

Results of the Monte Carlo study

For each of the 1000 replications, the model fit and test statistics (t and χ^2) were recorded. As previously mentioned, the asymptotic covariance matrix[18] is used for the WLS and ML estimation. More specifically, the computation of WLS takes the *inverse* of the asymptotic covariance matrix. If this matrix is "not positive definite" there is no inverse matrix and thus the computation either fails entirely or gives results that are statistically incorrect. This problem is identified by (1) a warning message in the LISREL software output file and (2) an examination of the results where incorrect statistical values are revealed (e.g., negative chi-square values are incorrect because squared values, by definition, must be positive). In examining the results below, it should be noted that there are a few simulation cells in which the first run of the simulation resulted in all of the replications being non-computable (i.e., the results are not interpretable because they are statistically incorrect). For these cases the simulation was re-run; however, the results were the same – the solution was

[17] Two levels of the number of items in the scale by two levels of item distributions by four ratios of group sizes.
[18] More technically, it is the asymptotic covariance matrix of the estimated coefficients.

not valid. The solution was not valid because the matrix was not positive definite and therefore the inverse of the asymptotic covariance matrix could not be computed which is needed in order to implement the WLS method for covariance and correlation structures (for a discussion on not positive definite matrices see Wothke, 1993). The computation of ML, on the other hand, does not require the inverse of this matrix. To get ML estimates, you maximize the likelihood of the parameters given the data; thus, it does not involve a direct inversion of the asymptotic covariance matrix. Hence, the results using ML, as shown below, were computable.

There are a number of reasons why the asymptotic covariance matrix is "not positive definite." One possible reason could be due to sampling variation. When sample size is small, a sample covariance or correlation matrix may be not positive definite due to mere sampling fluctuation (see Anderson & Gerbing, 1984). A second reason could be due to poor parameter values at the start of the iteration process (Byrne, 1998). For example, if the start value is a positive number but the true estimated value is negative, the solution may be unable to continue iterations or may not converge. Thus, it is a problem when there is a wide discrepancy between the start values and the true estimates. Another explanation "is that the model is empirically underidentified in the sense that the information matrix is nearly singular (i.e., it is close to be nonpositive definite)" (Byrne, 1998, p. 68). Given the problem of a not positive definite matrix, one limitation with this DIF MIMIC approach is that errors are inevitable. One should therefore be cautious and always check that the matrix being analyzed is correct. With this in mind, the following results for the equal sample size condition (Part A) and the unequal sample size condition (Part B) are presented below.

Part A: Equal sample size condition

Model fit

The mean model fit values over the 1000 replications for the DIF MIMIC models are presented in Tables 4.1 and 4.2. The first table shows fit statistics for the DIF MIMIC model conducted with Jöreskog's (2002) ML estimation method. These results suggest that the overall model for each cell of the 10- and 20-item scales fits at least adequately. For the 10 and 20-item scales, the RMSEA values are all less than .10 suggesting the data have a good fit to the model.

Table 4.2 displays the mean fit statistics for the DIF MIMIC model conducted with Jöreskog's WLS estimation method. For the 10-item skewed scale data with a sample size combination of 500:500 the fit values were not computed because the asymptotic covariance matrix was not positive definite. Likewise, the 20-item symmetrical and skewed 200:200 scale data with WLS estimation did not produce any valid data because of the not positive definite matrix. A further discussion of this problem is located at the end of the results section of this dissertation. For the cells that had valid data, the RMSEA values were reasonable (i.e., less than .10). Given that the models fit adequately, the DIF MIMIC model is consistent with our use.

Table 4.1. Mean fit indices for the DIF MIMIC model using *ML* estimation and *equal* sample size combinations for the **10- and 20-item** scales.

		10-item scale		20-item scale	
	Sample size	**RMSEA**	χ^2 (*df*=43)	**RMSEA**	χ^2 (*df*=188)
Symmetric distribution	200:200	.07	98.98	.08	519.45
	500:500	.05	92.13	.05	445.07
	1000:1000	.03	90.33	.03	427.24
Skewed distribution	200:200	.07	121.96	.08	642.72
	500:500	.04	115.96	.05	555.14
	1000:1000	.03	113.39	.03	528.86

Table 4.2. Mean fit indices for the DIF MIMIC model using *WLS* estimation and *equal* sample size combinations for the **10-and 20-item** scales.

		10-item scale		20-item scale	
	Sample size	**RMSEA**	χ^2 (*df*=43)	**RMSEA**	χ^2 (*df*=188)
Symmetric distribution	200:200	.09	113.28	--	--
	500:500	.05	96.19	.07	694.77
	1000:1000	.03	92.35	.04	517.10
Skewed distribution	200:200	.09	111.27	--	--
	500:500	--	--	.07	607.47
	1000:1000	.03	90.26	.04	481.97

Type I error rates

The DIF MIMIC model was evaluated based on its ability to control Type I error rates under a variety of conditions. For the individual parameters, the chi-square values and t-values were examined. Thus, the chi-square value used for examining the Type I error rate is the difference in chi-squares between the MIMIC model with no group to the item path and the MIMIC model with the group to item path (e.g., λ_{12} in Figure 2.1). Using this chi-square value, the proportion of rejections was counted based on the chi-square *p*-value, with *p*-

values less than 0.05 leading to a decision not to reject the hypothesis. Likewise, the t-values were saved from the MIMIC model with the group to item path. The t-value in LISREL represents the parameter estimate divided by its standard error. Based on a level of .05, the t-statistic had to be greater than the absolute value of 1.96 to be rejected (Byrne, 1998). Accordingly, the proportion of rejections was counted, which represent the Type I error rates. Both the chi-square and t-value rejection rates (i.e., Type I error rates) for the *equal* sample size conditions are shown in Tables 4.3 through 4.6.

For the symmetrically distributed 10-item data using ML estimation (see Table 4.3), the Type I error rate was inflated (7.7% - 10.3%) for all four sample size conditions. Likewise, for the skewed 10-item data using ML estimation (see Table 4.3), the Type I error rate was also inflated (12.5% to 14.8%) for all sample size conditions. Table 4.4 shows that the empirical Type I error rates for the symmetrically distributed 20-item data using ML estimation were also inflated (10.8% - 14.7%) for all four sample size conditions. As shown in the same table, the Type I error rates for the skewed 20-item data using ML estimation were even more inflated than the symmetrically distributed data and ranged from 11.6% to 16.3% for all sample size conditions.

Table 4.3. Empirical Type I error rates of the DIF MIMIC model using *ML* estimation method across distributional condition, and *equal* sample size combinations for the **10-item** scale.

	Decision based on		**Sample size combinations**		
			200:200	**500:500**	**1000:1000**
Symmetric distribution	chi-square	*Reject*	.103	.093	.077
		Valid reps	964	995	993
	t-value	*Reject*	.098	.090	.076
		Valid reps	964	995	993
Skewed distribution	chi-square	*Reject*	.126	.148	.125
		Valid reps	957	991	995
	t-value	*Reject*	.122	.145	.124
		Valid reps	957	991	995

'*Valid reps*' is shorthand for the number of valid replications.

Table 4.4. Empirical Type I error rates of the DIF MIMIC model using *ML* estimation method across distributional condition, and *equal* sample size combinations for the **20-item** scale.

	Decision based on		**Sample size combinations**		
			200:200	**500:500**	**1000:1000**
Symmetric distribution	chi-square	*Reject*	.118	.108	.147
		Valid reps	626	508	470
	t-value	*Reject*	.112	.108	.147
		Valid reps	626	508	470
Skewed distribution	chi-square	*Reject*	.162	.163	.116
		Valid reps	660	575	481
	t-value	*Reject*	.153	.160	.116
		Valid reps	660	575	481

In terms of the 10-item scale with WLS estimation (see Table 4.5), the symmetrically distributed data showed inflated Type I error rates ranging from 9.9% to 23.5%. Likewise, the skewed data was also inflated (14.7% to 28.3%). It should also be noted that there were

no valid cells for the 10-item scale with skewed data for the 500:500 sample size combination because the matrix was not positive definite.

The 20-item scale using WLS estimation (see Table 4.6) showed even higher Type I error rates ranging from 24.9% - 46.7%. As one can also see, there were no valid chi-square or t-values for the 200:200 sample sizes combinations due to the problem of a non-positive definite matrix.

Table 4.5. Empirical Type I error rates of the DIF MIMIC model using *WLS* estimation method across distributional condition, and *equal* sample size combinations for the **10-item** scale.

	Decision based on		**Sample size combinations**		
			200:200	**500:500**	**1000:1000**
Symmetric distribution	chi-square	*Reject*	.235	.131	.099
		Valid reps	948	996	997
	t-value	*Reject*	.225	.131	.100
		Valid reps	891	959	982
Skewed distribution	chi-square	*Reject*	.283	Not	.147
		Valid reps	972	computable	991
	t-value	*Reject*	.275	Not	.149
		Valid reps	912	computable	957

Table 4.6. Empirical Type I error rates of the DIF MIMIC model using *WLS* estimation method across distributional condition, and *equal* sample size combinations for the **20-item** scale.

	Decision based on		**Sample size combinations**		
			200:200	**500:500**	**1000:1000**
Symmetric distribution	chi-square	*Reject* *Valid reps*	Not computable	.341 988	.249 993
	t-value	*Reject* *Valid reps*	Not computable	.350 956	.251 964
Skewed distribution	chi-square	*Reject* *Valid reps*	Not computable	.467 959	.305 957
	t-value	*Reject* *Valid reps*	Not computable	.463 924	306 930

Part B: Unequal sample size condition

Model fit

The mean model fit values for the DIF MIMIC models are shown in Tables 4.7 and 4.8. The first table shows the fit statistics for the DIF MIMIC model conducted with Jöreskog's ML estimation method. These results suggest that the overall model for each cell of the 10- and 20-item scales fit adequately. For both the scale lengths, the RMSEA values are all <.05 suggesting the data fit the model very well. In addition, the RMSEA fit statistic for the DIF MIMIC models conducted with the WLS estimation also suggest that the data fit the model adequately.

Table 4.7. Mean fit indices for the DIF MIMIC model using *ML* estimation and *unequal* sample size combinations for the **10 and 20 item** scales.

	Sample size	10-item scale RMSEA	10-item scale χ^2 (*df*=43)	20-item scale RMSEA	20-item scale χ^2 (*df*=188)
Symmetric distribution	1:9	.03	90.20	.03	425.98
	2:8	.03	90.06	.03	427.21
	3:7	.03	90.06	.03	426.65
	4:6	.03	89.87	.03	430.49
Skewed distribution	1:9	.03	113.47	.03	531.78
	2:8	.03	112.39	.03	527.44
	3:7	.03	114.31	.03	528.40
	4:6	.03	113.72	.03	528.93

Table 4.8. Mean fit indices for the DIF MIMIC model using *WLS* estimation and *unequal* sample size combinations for the **10 and 20 item** scales.

	Sample size	10-item scale RMSEA	10-item scale χ^2 (*df*=43)	20-item scale RMSEA	20-item scale χ^2 (*df*=188)
Symmetric distribution	1:9	.03	94.07	.04	528.27
	2:8	.03	93.65	.04	531.69
	3:7	.03	93.79	.04	532.83
	4:6	.03	93.24	.04	535.24
Skewed distribution	1:9	.03	92.77	.04	491.34
	2:8	.03	93.06	.04	494.32
	3:7	.03	94.51	.04	499.09
	4:6	.03	93.72	.04	497.22

Type I error rates

As in Part A, the chi-square values and t-values were examined and used to evaluate the Type I error rates of the DIF MIMIC model under a variety of conditions. Both the chi-

square and t-value rejection rates (i.e., Type I error rates) for the *unequal* sample size conditions are shown in Tables 4.9 through 4.12.

For the symmetrically distributed 10-item data using ML estimation (see Table 4.9), the Type I error rate was inflated (9% - 11.7%) for all four sample size conditions. Likewise, the skewed 10-item data using ML estimation also showed inflated Type I error rates (13.4% to 14.6%) for all sample size conditions.

For the symmetrically distributed 20-item data using ML estimation (see Table 4.10), the Type I error rate was also moderately inflated (9.8% - 12.7%) for all four sample size conditions. The Type I error rate for the skewed 20-item data using ML estimation was even more inflated than the symmetrically distributed data and ranged from 11.3% to 16.3% for all sample size conditions.

Table 4.9. Empirical Type I error rates of the DIF MIMIC model using *ML* estimation method across distributional condition, and *unequal* sample size combinations for the **10-item** scale.

	Decision based on		**Sample size ratios**			
			1:9	**2:8**	**3:7**	**4:6**
Symmetric distribution	chi-square	*Reject*	.097	.116	.090	.103
		Valid reps	982	988	996	996
	t-value	*Reject*	.096	.117	.092	.099
		Valid reps	982	988	996	996
Skewed distribution	chi-square	*Reject*	.136	.134	.134	.143
		Valid reps	974	983	991	994
	t-value	*Reject*	.146	.143	.137	.144
		Valid reps	974	983	991	994

Table 4.10. Empirical Type I error rates of the DIF MIMIC model using *ML* estimation method across distributional condition, and *unequal* sample size combinations for the **20-item** scale.

	Decision based on		**Sample size ratios**			
			1:9	**2:8**	**3:7**	**4:6**
Symmetric distribution	chi-square	*Reject*	.123	.105	.098	.124
		Valid reps	528	513	479	467
	t-value	*Reject*	.127	.105	.100	.126
		Valid reps	528	513	479	467
Skewed distribution	chi-square	*Reject*	.159	.113	.143	.163
		Valid reps	536	503	490	491
	t-value	*Reject*	.157	.117	.143	.159
		Valid reps	536	503	490	491

In terms of the 10-item scale with WLS estimation (see Table 4.11), the symmetrically distributed data showed inflated Type I error rates ranging from 11.4% to 13%. Likewise, the skewed data was also inflated (13.8% to 17.8%). The 20-item scale using WLS estimation (see Table 4.12) showed even higher Type I error rates for both the symmetrically distributed data (18.8% to 23.4%) and the skewed data (22.4% to 32.3%).

Table 4.11. Empirical Type I error rates of the DIF MIMIC model using *WLS* estimation method across distributional condition, and *unequal* sample size combinations for the **10-item** scale.

	Decision based on		**Sample size ratios**			
			1:9	**2:8**	**3:7**	**4:6**
Symmetric distribution	chi-square	*Reject*	.115	.126	.114	.125
		Valid reps	979	994	999	995
	t-value	*Reject*	.122	.130	.115	.126
		Valid reps	979	994	999	995
Skewed distribution	chi-square	*Reject*	.138	.162	.171	.178
		Valid reps	982	995	996	998
	t-value	*Reject*	.146	.167	.174	.177
		Valid reps	982	995	996	998

Table 4.12. Empirical Type I error rates of the DIF MIMIC model using *WLS* estimation method across distributional condition, and *unequal* sample size combinations for the **20-item** scale.

	Decision based on		**Sample size ratios**			
			1:9	**2:8**	**3:7**	**4:6**
Symmetric distribution	chi-square	*Reject* *Valid reps*	.188 903	.211 966	.207 998	.232 999
	t-value	*Reject* *Valid reps*	.195 903	.215 966	.209 998	.234 999
Skewed Distribution	chi-square	*Reject* *Valid reps*	.224 991	.259 999	.279 999	.320 982
	t-value	*Reject* *Valid reps*	.226 991	.261 999	.283 999	.323 982

Of course, given that the Type I error rates are inflated, this precludes us from investigating the statistical power of this DIF detection method. In addition, although both the chi-square and the t-values were inflated, the t-value was more inflated in some conditions described above.

CHAPTER V: SUMMARY AND DISCUSSION

Review of the research questions and novel contribution

Given that short scales are typically found in the educational and psychological disciplines and the MIMIC method is the most appropriate method for investigating DIF in short scales, the primary purpose of this dissertation was to investigate the Type I error rates for this DIF method. This investigation was conducted using Jöreskog's (2002) ML and WLS estimation methods. As mentioned in the introduction of this dissertation, no previous study had examined the Type I error rates for the DIF MIMIC method let alone its implementation in Jöreskog's LISREL software. Accordingly, the primary focus of this dissertation was to examine the Type I error rate of the proposed MIMIC approach under a variety of study conditions including seven sample size combinations, two item response distributions, two scale lengths, and two estimation methods.

Investigation of the Type I error rates in relation to the seven design conditions was needed for a better understanding of how the DIF MIMIC approach might perform in applied testing conditions. In doing so, this study adds to the measurement literature about the performance of this method with short scales. It should be kept in mind that, given the widespread use of Jöreskog's LISREL software, his methods of estimating for the DIF MIMIC model have the potential for wide impact.

Summary of simulation results

The results of this study found that the DIF MIMIC model has inflated Type I error rates with both the 10- and 20-item scales with ML and WLS estimation methods under all study design conditions. The Type I error rates were more inflated for the skewed data than

the symmetric data and the Type I error rates were more inflated for WLS compared to ML estimation. The results also illustrated that a limitation of the DIF MIMIC method with Jöreskog's WLS estimation method is that it produced not positive definite asymptotic covariance matrices. As discussed in the results section, the matter of a not positive definite matrix is problematic for WLS estimation (as opposed to ML) because the inverse of the asymptotic covariance matrix is needed in order to implement the method for covariance and correlation structure.

Given that there has been no empirical investigation of Jöreskog's MIMIC method for dealing with ordinal and continuous variables (i.e., joint covariance matrix), a suggestion as to what may be problematic with the MIMIC method was explored by investigating related research. Based on a review of the literature regarding the use of the Pearson and polychoric correlation matrices, it is possible that the Type I error rates were inflated because the MIMIC model with Jöreskog's (2002) estimation methods may produce incorrect standard errors for the asymptotic covariance matrix. This hypothesis is supported by several simulation studies by Olsson (1979), Babakus, Ferguson and Jöreskog (1987), Bollen and Barb (1981), Rigdon and Ferguson (1991) and others who have shown that estimates of the standard errors are incorrect under a variety of conditions when Pearson and polychoric correlations are calculated from discrete variables. For example, Olsson (1979) examined ordinal data with a Pearson Product-moment correlation and ML estimation and found that it could lead to incorrect standard errors of estimates. Likewise, Babakus, Ferguson and Jöreskog (1987) found inflated estimates of the standard errors using ML estimation when the data were ordinal (5-point rating scale) with both Pearson and polychoric correlations in LISREL. Furthermore, Rigdon and Ferguson (1991) found that WLS and ML fitting

functions with the polychoric correlation coefficient and ordinal data (5 response categories) produced biased standard errors. Of course, the current use of the DIF MIMIC method with Jöreskog's estimation procedure does not use a Pearson or polychoric matrix, but rather a joint covariance matrix of continuous and ordinal variables. Therefore, this standard error explanation is speculative but worthy of future consideration. In other words, the findings described above are only suggestive of why the DIF MIMIC approach has inflated Type I errors and hence may be a good starting point for future research. In the same light, biased parameter estimates should be investigated.

Based on the results from the current study, this dissertation contributes to the measurement literature by cautioning researchers against the use of the DIF MIMIC method with Jöreskog's estimation methods in LISREL. Without the results from this dissertation, it is likely that researchers would use this approach without realizing it has inflated Type I errors (at least for sample sizes less than 1000). Accordingly, given that this simulation study was motivated by practical contexts wherein the data were reflective of real test data and the design conditions were chosen based on practical contexts, this author recommends avoiding the MIMIC approach with Jöreskog's ML and WLS estimation procedures for investigating DIF. Moreover, for studies that have used the DIF MIMIC method with Jöreskog's ML and WLS estimation, it is likely that too many DIF items were flagged as functioning differently between groups because of the inflated Type I error rate of this method. Thus, for these studies, it is difficult to determine which items are truly functioning differently from those items that are falsely flagged as functioning differently. Hence, one should be cautious when interpreting results from these studies.

What should researchers use for DIF detection with short scales?

While Millsap and Meredith (Meredith, 1993; Meredith & Millsap 1992; Millsap & Meredith, 1992) are correct that a latent matching variable is in line with the formal definition of DIF and is therefore most appropriate for DIF methods compared to observed score matching variables, no latent variable approach for short scales currently exists that has Type I error rates near the nominal range of .05. Thus, until a viable latent variable approach is available, I would recommend using an unconstrained cumulative logits approach to Ordinal Logistic Regression (UCLOLR; French & Miller, 1996; Zumbo, 1999) or the generalized Mantel-Haenszel (GMH; Mantel & Haenszel, 1959) for Likert response format items. This recommendation is based on the results of a recent simulation study by Kristjansson, Aylesworth, Boss, McDowell and Zumbo (in press) who found that these observed score methods controlled Type I error well and had statistically high power for detecting uniform DIF (power for UCLOLR = .99; power for GMH = .96) and non-uniform DIF (UCLOLR & GMH had nearly perfect power at 1.0) with short scales and polytomous item formats. Although the UCLOLR and GMH are observed score methods and hence are, at best, approximations to the latent variable approaches, the simulation results in Kristjansson et al. show that the statistical power is substantial and thus these methods are able to detect DIF items.

Limitations and future directions

One limitation of the present simulation study is that it only investigated scales with a 4-point item response format. It is unknown what the Type I error rates would be if the DIF MIMIC approach was used with variables of different scale formats (e.g., binary). In this author's opinion, it is likely that two, three and five point response formats would result in

similar inflated Type I errors for the DIF MIMIC model. However, it is unknown what would happen in the case with greater than five scale points because adding scale points eventually becomes more similar to continuous item data. Although this is a limitation in the present study, the investigated 4-point scale format is commonly used in the social sciences and, given that the DIF MIMIC approach has inflated Type I error rates for this scale format, there is little reason to explore this further as, to be useful, the DIF MIMIC method should be able to operate properly on commonly encountered data formats.

One must also assess the validity of the current DIF MIMIC approach because variations in the implementation of DIF statistical methods, such as choice of model, computer algorithm and DIF statistical indices, affect the reliability and validity of DIF statistics. Moreover, given that this was a Monte Carlo study and not all possible models were studied, caution should be used in generalizing the results and conclusions of the DIF MIMIC approach with Jöreskog's estimation methods beyond the models and conditions investigated. For example, the results cannot be generalized to other DIF MIMIC approaches that use different estimation methods (e.g., UWLS). Thus, further research is necessary to determine the operating characteristics (e.g., Type I error rates and power) of the DIF MIMIC model with other estimation methods. For example, a simulation study is needed that examines the operating characteristics of Muthén's estimation method, which is different than Jöreskog's method (see Muthén, 1989; Muthén, Kao & Burstein, 1991). If Type I error rates are found to be equal to or less than the nominal range (i.e., 5%) using other estimation procedures, the power of the MIMIC approach must also be investigated. In doing so, the power must be tested under a variety of conditions, such as varying the number of DIF items in a scale (e.g., systematic evaluation of one, two, three, and more DIF items in a scale).

In regards to this author's prediction that the inflated Type I errors of the DIF MIMIC model investigated in this dissertation may be partly due to incorrect standard errors, future research needs to investigate this conjecture. If the standard errors are indeed incorrect, is there an adjustment that researchers can use to correct the standard errors? It should be noted that a simple application of a known correction to the standard errors is not appropriate because Jöreskog's estimation method uses a joint covariance matrix which is a more complex matrix than the Pearson and Polychoric matrices for which a correction has already been developed. Lastly, more work on an effect size measure must be undertaken; as in other statistical procedures, DIF statistics should be accompanied by an effect size measure (Kirk, 1996; Zumbo, 1999).

Validity and implications of Type I error rates for DIF

The main focus of this dissertation has been on investigating the Type I error rates of the DIF MIMIC model. Given that the results of this dissertation study found an inflated Type I error rate for the DIF MIMIC approach, a discussion of the implications and consequences of this type of error, and particularly an inflated Type I error rate, is necessary. Accordingly, this discussion focuses on a broader perspective in terms of Type I error rates and their importance in identifying differentially functioning test or scale items. In addition, given that statistical DIF indices do not provide researchers or test developers with an explanation for what to do with an item, and that "there has been little progress in identifying the causes or substantive themes that characterize items exhibiting DIF" (AERA, APA & NCME, 1999, p. 78), decisions as to what to do with items exhibiting DIF will also be addressed.

It is well known that the results of all DIF analyses must be interpreted within the context of the intended purpose of the test or measure (Camilli & Shepard, 1994; Clauser & Mazor, 1998; Zumbo, 1999) and the purpose of a test is shaped by policy. For example, it was policy that was used to instigate mental testing for the United States army. These tests were deemed necessary in order to help determine job placements or discharge of people in World War I. Moreover, it is often policy that defines the groups (e.g., visible minorities) to be investigated. Thus, when an item is identified as functioning differentially among groups of test takers, it is really a matter of policy as to what should be done with the item(s) (Clauser & Mazor, 1998). There are three standard practices for dealing with items that are identified as displaying DIF: (1) review the item and potentially remove, retain or revise the item, (2) remove the item, or (3) retain the item. One key difference among these practices is "whether identified items are considered biased until proven valid or valid until proven biased" (Clauser & Mazor, p. 40). To motivate this discussion, let us consider an item that is identified as displaying DIF. The questions are, 'how does one go about deciding what to do with that item?' and 'what are the implications if the DIF method has a Type I error rate that is inflated?'

Before discussing the implications of these practices, a brief comment on the consequences of a low Type I error rate for DIF methods is needed. When thinking about statistical power, one must keep in mind the Type I error rate of the method. Power is defined as the probability of correctly rejecting a false null hypothesis. If the Type I error rate is below the nominal value (i.e., one is less likely to find a difference that is not there) and there is adequate power, there is no real concern. However, if the Type I error rate is below the nominal value and the power is low, we would miss identifying DIF items (i.e., undetected

DIF). Having said this, we really need to investigate how low the power can be before one would start missing items that truly function differentially. In view of this, the following discussion will focus only on the decisions and implications of DIF results when Type I error rates are inflated.

Review the item.

Reviewing an item that is identified as DIF is a common practice in large testing organizations (e.g., Educational Testing Service [ETS], 2003) and a common suggestion from researchers who find DIF items. For example, the majority of authors from the DIF MIMIC studies reviewed in this dissertation (see Appendix A) suggest that item(s) displaying DIF need further review by subject content experts. In general, this approach treats items that are flagged as DIF as neither valid nor biased until a proper review of the items is conducted. Accordingly, this stance is considered conservative because false conclusions of bias are protected against by requiring further evaluation of the meaning of DIF results before items are eliminated (Camilli & Shepard, 1994). During this evaluation one must keep in mind an essential part of validation: the consequences of test decisions (Messick, 1989; Zumbo, 1999). Regardless of whether a test is used for decision-making (e.g., entrance admission, job qualification) or for research, one needs to explore the consequential basis of validity. Messick explicitly states that validity requires the consequences of test use – both intended and unintended effects – to be assessed. With this in mind, it is in the hands of content experts to clearly understand *why* the item is endorsed differently for different groups. In doing so, their decisions must be made with the *purpose* of the test or measure in mind.

If the item is tapping a factor other than the one of interest, *item bias* may be present. According to the AERA, APA, and NCME (1999) *Standards for educational and*

psychological testing, sources of item bias include, but are not limited to (1) inappropriate selection of test content (e.g., offensive or emotionally disturbing material, or language that has different meanings in one subgroup than another), (2) lack of clear test instructions, (3) items eliciting various responses other than those intended (e.g., response acquiescence), (4) response formats, and (5) different constructs being measured by a test. An example of item bias arising from test response formats could be if the item relies on some capability (e.g., English language proficiency) that is irrelevant to the purpose of the measurement but causes impediments for some respondents. For example, an item measuring mathematics problem-solving that makes inappropriately heavy demands on verbal ability (e.g., if an obscure or difficult word is used as part of the problem-solving text) would probably be judged as biased against respondents whose first language is other than that of the item. In terms of the CES-D data that was used as an example to motivate this dissertation problem, item 17 "crying" was shown to exhibit DIF. This item would be judged as biased if the item was tapping a factor (e.g., socialization of stigma against crying for males) other than the one of interest (i.e., depression). One reason why an item may be judged as biased when a test is found to measure different constructs across subgroups is because "different components come into play from one subgroup to another. Alternatively, an irrelevant component may have a more significant effect on the performance of examinees in one subgroup than in another" (AERA, APA & NCME, 1999, p. 81).

On the other hand, if an item is tapping a relevant characteristic of the test and the item is flagged as displaying DIF, *item impact* may be present (i.e., the groups truly differ on the underlying factor being measured). This may occur, for example, in a science test wherein males are more likely than females to correctly answer questions based on a reading

passage related to technical science material. This would be item impact if "such passages are a legitimate part of the reading materials that students are expected to encounter in high school and college and it is reasonable to expect that they will need to be able to comprehend such reading material in college" (Linn, 1993, p.353). In terms of the CES-D "crying" item, item impact would be present if the differential reporting of crying was judged to be a relevant characteristic of the symptomology of depression for which the CES-D is purported to measure. In order to decide whether an item shows bias or impact requires follow-up investigations into whether the source of the differential item functioning is relevant or irrelevant to the construct being measured by the test.

After reviewing an item flagged as displaying DIF and reaching a logical and plausible explanation for the nature of the DIF, considerable judgment is required to determine whether the contributing factor affecting the item functioning is relevant to the test purpose. Based on the judgments and decisions from subject-matter and testing experts, the reviewed item could be removed from the test or measure, retained, or revised (e.g., modify wording).

Remove reviewed item.

If *item bias* is present, the most common practice is to remove the item from the test or measure. As Camilli and Shepard (1994) state, "only if the source of differential difficulty is irrelevant should DIF items be ruled as biased and eliminated from a test" (p. 145). As previously mentioned, an item is identified as biased if it is found to measure construct-irrelevant factors (i.e., factors that measure knowledge, skills, abilities or behaviors that are extraneous to the purpose of the test). The Educational Testing Service (2003) is one

example of a company who removes items that are judged to be unfair, and they define unfair test items as items that tap construct-irrelevant factors.

There are two main situations in which construct-irrelevance threatens the validity of a test. One situation is if the DIF item measures an extraneous skill that makes the item more difficult for some individuals or groups. An example of this would be a verbal analogy item that requires knowledge of sporting terms wherein such knowledge is irrelevant to the construct "verbal reasoning" measured by the test. A second situation is if extraneous clues in the DIF item (e.g., wording, item format) help some individuals or groups respond correctly or endorse items in ways that are irrelevant to the construct being assessed. For example, it has been found that math addition problems presented in a vertical format as opposed to a horizontal format is more difficult for young students to correctly answer (cited in Camilli & Shepard, 1994).

Retain reviewed item.

On the other hand, if *item impact* is evident, this author recommends keeping the item. This recommendation is supported by the Graduate Record Examinations Board (1998) who purports that an item is included "*only* if the [subject-matter and testing] experts agree that the question is substantively correct, correctly written, and important to the measurement purpose of the test" (p. 5-6). The guidelines from ETS (2003) also support this recommendation in that they state that "test items that cause group differences because of valid factors are fair" (p. 4) and "it is fair to measure valid knowledge, even if the knowledge is not equally distributed across groups" (p. 4).

Revise reviewed item.

In general, reviewed DIF items are either removed from or retained in a test. If it is decided that an item needs revision, this is generally because the item content or language has become outdated and, therefore, may consequently affect the interpretations made from the test score (i.e., validity).

Given the above discussion, a DIF method with an inflated Type I error rate can have further implications if one decides to review DIF items. An inflated Type I error rate will result in a large number of items flagged as exhibiting DIF, most of which are false occurrences of DIF. This will result in a review committee examining *many* potentially problematic items, which could have two possible consequences. First, reviewing a large number of items could be an expensive and time consuming process. Second, reviewing numerous items could impact experts' opinions and views as to what to do with the items. For example, reviewers may feel over-burdened by the copious items they need to assess. In order to reduce their workload and manage their time more efficiently, they might resort to practices like creating quick rules to keep good items or remove poor items. As a result, they are likely to skip basic item reviewing procedures and are thereby more likely to make errors of judgment. For example, truly biased items may remain in the test because of the lack of or inadequate follow-up studies.

Whether or not a DIF procedure has an inflated Type I error rate, the practice of reviewing DIF items in order to make informed decisions as to what to do with them is *not* universally accepted. The key arguments against this practice revolve around issues of validity. For example, Raju and Ellis (2002) suggest that although review committee experts may help identify reasons for DIF, these reasons may not always be plausible or valid explanations for DIF. Moreover, there is little evidence to support the validity of such

judgments (Clauser & Mazor, 1998). While some people argue that future studies investigating review committees' post hoc explanations for DIF are needed to help support the validity of experts' judgments, others take a firmer stance. For example, David Thissen, a prominent researcher and professor in the theory of educational and psychological testing at the University of North Carolina at Chapel Hill, would argue that current techniques devised to help explain and draw accurate conclusions about the causes of DIF are too imprecise to make a credible case for or against an item (cited from Camilli & Shepard, 1994). Such critics are in favour of the more liberal practice of removing items flagged as displaying DIF without review.

Remove the item.

The second strategy used to deal with items that are flagged as displaying DIF is to remove them until they can be shown to measure the intended test purpose and judged to be fair. That is, the items are considered biased until proven valid. This is a more liberal approach to eliminating items flagged as DIF in comparison to the more conservative approach described above (i.e., reviewing items). Followers of this approach believe that test items flagged by DIF statistics should be immediately eliminated from the test or measure and these items can be only re-introduced into the test if follow-up evidence that proves they are valid test items is presented.

Supporters of this approach generally establish a set of rules that define the exact type of DIF items to be removed until additional evidence can be collected to support the validity of the item(s). For example, ETS established a set of guidelines that identify certain content categories of questions that are sometimes problematic in terms of DIF. Any skills test

questions in these categories, regardless of their performance with respect to DIF, are removed from the test (Graduate Record Examinations Board, 1998).

However, critics of this liberal approach would argue that because DIF is a necessary but not sufficient condition for item bias, removing an item without review will likely lead to the deletion of valid items. It has often been reported that test developers "are often confronted by DIF results that they cannot understand; and no amount of deliberation seems to help explain why some perfectly reasonable items have large DIF values" (Angoff, 1993, p. 19). Therefore, considering an item as biased may be too premature. Moreover, one must consider the impact on the content validity of the test or measure when eliminating an item. Removing an item may result in narrowing the breadth of the construct or domain, resulting in construct under-representation. As Messick (1989) states, construct under-representation occurs when "the test is too narrow and fails to include important dimensions or facets of the construct" (p. 34). The problem of construct under-representation will be even greater if the DIF method has an inflated Type I error rate. If this occurs, many items will be flagged as displaying DIF resulting in far too many items being removed from the test or measure. Not only may this practice of removing items homogenize a test with the potential consequence of construct-underrepresentation, but it does so at the expense of validity.

Removing items in lieu of evidence does not ensure validity. Such a liberal approach leads one to presume that statistical DIF indices tell researchers and test developers that there is something real about the item that makes it function differentially for different groups of test takers. As stated in the introduction to this discussion section, this is simply not true. DIF detection methods only identify items that are not measuring the same dimension(s) as the remaining non-DIF items. In other words, DIF indices signal differential multi-

dimensionality in a test or measure. Certainly then, multi-dimensionality does not specify bias or impact. Moreover, DIF indices can be unreliable. Unreliability in DIF indices occur because they are derived from a variety of DIF approaches, computer algorithms, estimation methods, and critical values. Thus, some results may be caused by statistical artifacts. Camilli and Shepard (1994) give a personal account wherein they have seen "innumerable occurrences of DIF disappear in a replication study or even when slightly different variations of a statistical technique were implemented" (p. 150). Thus, they firmly believe that "the process of item screening should not be based on statistical criteria alone" (p. 151). Accordingly, and in line with this author's belief, the appropriate use of DIF requires accurate statistical methods as well as procedures that incorporate item reviews followed by judgments of trained subject-matter specialists.

Retain the item.

An alternative strategy to removing items is to leave the item in the test. In this case, items flagged as displaying DIF are retained in the test until they are judged to be biased and unfairly related to group membership or flagged in some other way. In this case, the item is left in the test regardless of whether the item measures impact or bias. Thus, items are considered valid until proven biased. A potential benefit of this practice is that it may boost group differences in the test. That is, items flagged as DIF can be thought of as having higher discrimination between high and low performing groups compared to other test items. A second advantage of this approach is that it will *not* result in construct under-representation. Although items flagged as displaying DIF suggest that members of one group do more poorly than a comparison group, proponents of this approach argue that such items are a legitimate part of the content domain.

Critics of this practice believe it opens the door for test developers to simply conclude that they do not understand the reasons for the DIF rather than conscientiously re-examining the item(s) in rigorous follow-up studies. In other words, it allows test developers an excuse to merely fail to find confirmatory evidence of bias and hence the items remain in the test. Moreover, test developers may too easily dismiss the statistical DIF index value as a Type I error. That is, they will claim the item is falsely flagged as displaying DIF. This latter excuse, although perhaps accurate, will result in items remaining in the test without any further follow-up or review. In this author's opinion, such a practice may decrease the validity of test score interpretation.

If the procedure used to analyze DIF has an inflated Type I error rate, many items will be flagged as displaying DIF, all of which will be kept in the test until there is substantial evidence against leaving them in the test. Given that the majority of these items will be falsely flagged as displaying DIF because of the inflated Type I error rate, leaving them in the test is good. However, there are some items that are flagged as displaying DIF that truly function differently; and it is unknown which items these are. That is, DIF procedures do not distinguish items that are flagged because of the Type I error rate compared to items that truly show DIF. Leaving the items in the test that truly function differently is problematic because it is unknown as to where the problem lies. All that is known is that certain items are problematic, and what is unknown is where the source of differences is on particular items. The item could be biased or demonstrate impact. Unfortunately there is little literature discussing the implications of this practice.

Conclusions

In terms of the broader view of validity and the implications of Type I error rates, all three standard practices (i.e., reviewing, removing, or keeping the item) have their consequences. One common theme among all three practices is that more work is needed that investigates reasons for DIF among various content areas. DIF is not a replacement for item reviews. That is, DIF indices by themselves cannot provide information about issues of test bias and fair test use. Rather, DIF statistics in combination with ongoing evaluation and judgment about the meaning of DIF in light of the intended test purpose are required to be able to address the larger social issues of test bias and fairness. Thus, the inferences that may be drawn from items flagged as displaying DIF are really a question of validity, and it is in this sense that impact, bias, and fairness can be defined – all of which have to do with inferences, evaluations, judgments, and ethical considerations as they affect the intended purpose of the test. In terms of better understanding DIF results, it has been suggested that using additional matching variables may help explain otherwise unexplainable results (e.g., Angoff, 1993). For example, contextual variables (e.g., income level, parental education level) may be added in hopes of explaining more of the item variance. Thus, additional matching variables would hopefully reduce the variance of DIF values previously found, assuming, of course, that the added matching variables have the same meaning for the two groups (e.g., education or other social status variables must have the same meanings for each group). However, as Angoff (1993) reminds us, any decision to revise, remove, or retain items flagged as DIF depends on the nature of the additional variables (e.g., contextual or background variables).

Currently, the majority of published DIF studies either investigates DIF in popular scales or examines the statistical properties of DIF techniques. Unfortunately, the majority of studies that investigate DIF in scales do not present potential explanations of DIF nor do they address the larger social issue of fairness. Likewise, studies exploring the statistical properties of DIF methodologies need to address the issue of Type I and Type II errors in conjunction with the power of the DIF procedure. Thus, this author implores that future DIF research address more substantive questions about the interpretation, implications, and validity of their findings. Moreover, although this dissertation found discouraging results in terms of the investigated DIF MIMIC model with Jöreskog's estimation methods, new DIF methods for short scales are still needed. In fact, new DIF methods are needed that have the ability to handle more complex item responses and measurement structures (e.g., multidimensionality, hierarchical structure, and method effects) for short scales. It is, however, imperative that new methods be rigorously tested for their statistical properties such as Type I and Type II error rates, power, and robustness. More work on effect sizes is also required in order to ensure that the amount of DIF is meaningful. Currently, no effect size measure has been developed for SEM DIF methods and thus it is hoped that future studies will develop and incorporate an effect size measure for SEM DIF approaches.

References

Note: An asterisks (*) in front of a reference indicates the reference is cited in Appendix A.

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

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*, 155-73.

Andresen, E.M., Malmgren, J.A., Carter, W.B., & Patrick, D.L. (1994). Screening for depression in well older adults: Evaluation of a short form of the CES-D. *American Journal of Preventative Medicine, 10*, 77-84.

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Erlbaum.

Babakus, E., Ferguson, C.E., & Jöreskog, K. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research, 24*, 222-228.

Babor, T. F., De la Fuente, J. R., Saunders, J., & Grant, M. (1992). *Programme on substance abuse: AUDIT--The Alcohol Use Disorders Test: Guidelines for use in primary health care* (an update of WHO Document No. WHO-MNH-DAT-89.4 under the same title). Switzerland: World Health Organization.

Bagby, R. M., Parker, J. D. A., & Taylor, G. J. (1994). The Twenty-Item Toronto Alexithymia Scale-I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research, 38*, 23-32.

Bagby, R. M., Taylor, G. J., & Parker, J. D. A. (1994). The Twenty-Item Toronto Aexithymia Scale-II. Convergent, discriminant, and concurrent validity. *Journal of Psychosomatic Research, 38*, 33-40.

Beck, A.T., Weissman, A., Lester, D., & Trexler, L. (1974). The measurement of pessimism: The Hopelessness Scale. *Journal of Consulting and Clinical Psychology, 42*, 861-865.

* Bedford, A., & Deary, I.J. (1997). The personal disturbance scale (DSSI / sAD): Development, use, and structure. *Personality and Individual Differences, 22,* 493-510.

Binet, A., & Simon, T. (1973). *The development of intelligence in children.* New York: Arno. (Original work published 1916).

Bollen, K.A., & Barb, K.H. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review, 46*, 232-239.

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

* Brandt, J., Spencer, M., & Folstein, M. (1988). The telephone interview for cognitive status. *Neuropsychiatry & Behavioral Neurology, 1*, 111-117.

Burisch, M. (1997). Test length and validity revisted. *European Journal of Personality, 11,* 303-315.

Butcher, J.N., Dahlstrom, W.G., Graham, J.R., Tellegen, A., & Kraemmer, B. (1989). *Manual for the restandardized Minnesota Multiphasic Personality Inventory: MMPI-2*. Minneapolis, MN: University of Minnesota Press.

Byrne, B.M. (1998). *Structural equation modelling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale. Hillsdale, NJ: Erlbaum.

Camilli, G., & Shepard, L.A. (1994). *Measurement methods for the social sciences series: Methods for identifying biased test items* (Vol. 4). Thousand Oaks, CA: Sage.

Christensen, H., Jorm, A.F., Mackinnon, A.J., Korten, A.E., Jacomb, P.A., Henderson, A.S., & Rodgers, B. (1999). Age differences in depression and anxiety symptoms: A structural equation modelling analysis of data from a general population sample. *Psychological Medicine, 29*, 325-339.

Clark, V.A., Aneshensel, C.S., Frerichs, R.R., & Morgan, T.M. (1981). Analysis of effects of sex and age in response to items on the CES-D scale. *Psychiatry Research, 5*, 171-181.

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

Cole, N.S. (1993). History and development of DIF. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 25-29). Hillsdale, NJ: Erlbaum.

Corulla, W.J. (1990). A revised version of the psychoticism scale for children. *Personality and Individual Differences, 11*, 65–76.

* Crosswhite, F.J., Dossey, J.A., Swafford, J.O., McKnight, C.C., & Cooney, R.J. (1985). *Second international mathematics study summary report for the United States.* Champaign, IL: Stipes.

Curran, P.J., Bollen, K.A., Paxton, P., Kirby, J., & Chen, F. (2002). The noncentral chi-square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation. *Multivariate Behavioral Research, 37,* 1-36.

De Champlain, A., & Gessaroli, M.E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. *Applied Measurement in Education, 11,* 231-253.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9,* 327-346.

Educational Testing Service. (2003). *Educational Testing Service fairness and review guidelines.* Princeton, NJ: Educational Testing Service.

Eells, K., David, A., Havighurst, R.J., Herrick, V.E., & Tyler, R.W. (1951). *Intelligence and cultural differences.* Chicago: University of Chicago Press.

Fidalog, A.M., Mellenbergh, G.J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online, 5*(3), 1-11. Retrieved October 25, 2004, from http://www.mpr-online.de

Fleishman, J.A. (2004). *Using MIMIC models to assess the influence of differential item functioning.* Retrieved August 22, 2004, from http://outcomes.cancer.gov/conference/irt/fleishman.pdf.

Fleishman, J.A., Spector, W.D., & Altman, B.M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journal of Gerontology: Social Sciences, 57B*, S275-284.

* Folstein, M.G., Folstein, S.E., & McHugh, P.R. (1975). "Mini-Mental State:" A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*, 189-198.

Fossum, T.A. (2002). Keeping it short and simple: A simple rating scale for personality. *Dissertation Abstracts International*, DAI-B 63/05, 2581. (UMI No. 3051649)

Francis, L.J. (1996). The development of an abbreviated form of the Revised Junior Eysenck Personality Questionnaire (JEPQR-A) among 13-15 year olds. *Personality and Individual Differences*, *21*, 835-844.

French, A.W., & Miller, T.R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*, 315-332.

Gallo, J.J., Anthony, J.C., & Muthén, B.O. (1994). Age differences in the symptoms of depression: A latent trait analysis. *Journal of Gerontology, 49,* 251-264.

Gallo, J.J., Rabins, P.V., & Anthony, J.C. (1999). Sadness in older persons: 13-year follow-up of a community sample in Baltimore, Maryland. *Psychological Medicine, 29,* 341-350.

Gelin, M. N., & Zumbo, B.D. (2003). DIF results may change depending on how an item is scored: An illustration with the Center for Epidemiological Studies Depression (CES-D) Scale. *Educational and Psychological Measurement*, *63,* 63-72.

Gillis, J.S. (1980). *Child Anxiety Scale*. Champaign, IL: Institute for Personality and Ability Testing.

Goldberg, D.P., & Williams, P. (1988). *A user's guide to the General Health Questionnaire*. Windsor, England: NFER–Nelson.

* Goldberg, D., Bridges, K., Duncan-Hones, P., & Grayson, D. (1988). Detecting anxiety and depression in general medical settings. *British Medical Journal, 297,* 879-899.

Golding, J.M. (1988). Gender differences in depressive symptoms. *Psychology of Women Quarterly*, *12*, 61-74.

Gould, S.J. (1996). *The mismeasure of man: Revised and expanded.* New York: W.W. Norton and Company.

Graduate Records Exam Board. (1998). *Sex, race, ethnicity, and performance on the GRE General Test 1998-99.* Princeton, NJ: Educational Testing Service.

Grayson, D.A., Mackinnon, A., Jorm, A.F., Creasey, H., & Broe, G.A. (2000). Item bias in the Center for Epidemiological Studies Depression Scale: Effects of physical disorders and disability in an elderly community sample. *Journal of Gerontology: Psychological Sciences, 55B*, P273-P282.

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

Hertzog, C., Van Alstine, J., Usala, P.D., Hultsch, D.F., & Dixon, R. (1990). Measurement properties of the Center for Epidemiological Studies Depression scale (CES-D) in older populations. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2*, 64-72.

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

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

Jones, R.N. (2003). Racial bias in the assessment of cognitive functioning of older adults. *Aging and Mental Health, 7,* 83-102.

Jones, R.N., & Gallo, J.J. (2002). Education and sex differences in the Mini-Mental State Examination: Effects of differential item functioning. *Journal of Gerontology: Psychological Sciences, 57B*, P548-P558.

Jöreskog, K.G. (2002, June). *Analysis of ordinal variables 5: Covariates*. Retrieved January 6, 2004 from http://www.ssicentral.com/lisrel/column11.htm.

Jöreskog, K.G., & Goldberger, A.S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70,* 631-639.

Jöreskog, K., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software International.

Jöreskog, K.G., & Sörbom, D. (2003a). *PRELIS* (Version 2.51) [Computer software]. Chicago, IL: Scientific Software International.

Jöreskog, K.G., & Sörbom, D. (2003b). *LISREL* (Version 8.54) [Computer software]. Chicago, IL: Scientific Software International.

Kelloway, E.K. (1998). *Using LISREL for structural equation modeling: A researcher's guide*. Thousand Oaks, CA: Sage Publications.

Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25,* 146-162.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746-759.

Koh, H.K. (2003). *Type I error rates for multi-group confirmatory maximum likelihood factor analysis with ordinal and mixed item format data: A methodology for construct comparability.* Unpublished doctoral dissertation, The University of British Columbia, British Columbia, Canada.

Kristjansson, E., Aylesworth, R., Boss, M., McDowell, I., & Zumbo, B.D. (in press). A comparison of four methods for detecting DIF in ordered response items. *Educational and Psychological Measurement.*

Li, Y., Cohen, A.S., & Ibarra, R.A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing, 4,* 115-136.

* Linacre, J.M., Heinemann, A.W., Wright, B.D., Granger, C.V., & Hamilton, B.B. (1991). *The Functional Independence Measure as a measure of disability* (Research Rep. No. 91-01). Chicago, IL: Rehabilitation Services Evaluation Unit, Rehabilitation Institute of Chicago.

Linn, R.L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349-364). Hillside, NJ: Lawrence Erlbaum Publishers.

Lu, I. R. R., Thomas, D. R., & Zumbo, B. D. (in press). Embedding IRT in structural equation models: A comparison with regression based on IRT scores. *Structural Equation Modeling.*

McDonald, R.P., & Ho, M.R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*, 64-82.

Mackintosh, N.J. (1998). *IQ and human intelligence*. New York: Oxford University Press.

Mantel, N., & Haenszel, W.M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

March, J. (1997). *Multidimensional Anxiety Scale for Children*. New York: North Tonawanda Multi-Health Systems Inc.

Mast, B.T., & Lichtenburg, P.A. (2000). Assessment of functional abilities among geriatric patients: A MIMIC model of the Functional Independence Measure. *Rehabilitation Psychology, 45*, 94-64.

Meredith, W. (1993). Measurement invariance, factor invariance and factorial invariance. *Psychometrika, 58*, 525-543.

Meredith, W., & Millsap, R. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57,* 289–311.

Messick, S.A. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S.A. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*, 35-44.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

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

Millsap, R., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement, 16,* 389–402.

Minton, H.L. (1998). *Commentary on: "New methods for the diagnosis of the intellectual level of subnormals" Alfred Binet & Theodore Simon (1905), "The uses of intelligence tests" Lewis M. Terman.* Retrieved September 30, 2004, from http://psychclassics.yorku.ca/Binet/commentary.htm

Minton, H.L. (1984). The Iowa child welfare research stations and the 1940 debate on intelligence: Carrying on the legacy of a concerned mother. *Journal of the History of the Behavioral Sciences, 20*, 160-176.

Muñiz, J., Hambleon, R.K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*, 115-135.

Muthén, B.O. (1989). Using item-specific instructional information in achievement modelling. *Psychometrika, 54*, 385-396.

Muthén, B.O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*, 1-22.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45,* 19-30.

Muthén, B.O., Tam, W.Y., Muthén, L.K., Stolzenberg, R.M., & Hollis, M. (1993). Latent variable modeling in the LISCOMP Framework: Measurement of attitudes toward career choice. In D. Krebs & P. Schmidt (Eds.), *New directions in attitude measurement, Festschrift for Karl Schuessler* (pp. 277-290). Berlin, Germany: Walter de Gruyter.

O'Brien, R.M. (1979). The use of Pearson's *r* with ordinal data. *American Sociological Review, 44*, 851-857.

Olsson, U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research, 14*, 485-500.

Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 3*, 385-401.

Raju, N.S., & Ellis, B.B. (2002). Differential item and test functioning. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 156-188). San Francisco, CA: Jossey-Bass.

Ramsay, J.O. (2001). *TestGraf: A Program for the Graphical Analysis of Multiple-Choice Test and Questionnaire Data* [software and manual]. McGill University, Montreal, Canada.

Rigdon, E.E., & Ferguson, C.E. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research, XXVIII,* 491-497.

* Robins, L.N., Helzer, J.E., Croughan, J., & Ratcliff, K.S. (1981). National Institute of Mental Health Diagnostic Interview Schedule: Its history, characteristics, and validity. *Archives of General Psychiatry, 38*, 381-389.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Scholte, R.H.J., & De Bruyn, E.E.J. (2001). The Revised Junior Eysenck Questionnaire (JEPQ-R): Dutch replications of the full-length, short, and abbreviated forms. *Personality and Individual Differences, 31*, 615–625.

Seong, T-J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*, 299-311.

Sheehan, T.J., Fifield, J., Reisine, S., & Tennen, H. (1995). The measurement structure of the Center for Epidemiological Studies Depression scale. *Journal of Personality Assessment, 64*, 507-521.

Shevlin, M., Miles, J.N.V., & Bunting, B.P. (1997). Summated rating scales: A Monte Carlo investigation of the effects of reliability and collinearity in regression models. *Personality and Individual Differences, 23*, 665-676.

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research, 35,* 137-167.

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika, 66*, 563-576.

Stanton, J.M., Sinar, E.F., Balzer, W.K., & Smith, P.C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology, 55,* 167-194.

Steiger, J.H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*, 173-180.

Steiger, J.H., & Lind, J.C. (1980). *Statistically-based tests for the number of common factors.* Paper presented at the annual Spring Meeting of the Psychometric Society in Iowa City.

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

Swaminathan, H., & Gifford, J.A. (1983). Estimation of parameters in the three-parameter latent trait model. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 31-49). New York: Academic Press.

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

Terman, L.M. (1916). *The measurement of intelligence*. Boston, MA: Houghton Miffin.

The University of British Columbia (2004). *Master and Doctoral thesis submission: Manuscript-based format.* Retrieved August 20, 2004, from The University of British Columbia, Faculty of Graduate Studies Web site: http://www.grad.ubc.ca/students/thesis/index.asp?menu=001,002,000,000.

Uttaro, T., & Millsap, R.E. (1994). Factors influencing the Mantel-Haenzel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18,* 15-25.

* Ware, J.E., Kosinski, M., & Keller, S.D. (1996). A 12-item short form health survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care, 34,* 220-233.

Witarsa, P.M. (2003). *Nonparametric Item Response Modeling for Identifying Differential Item Functioning in the Moderate-to-Small-Scale Testing Context*. Unpublished doctoral dissertation, The University of British Columbia, British Columbia, Canada.

* Wittchen, H.U. (1994). Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): A critical review. *Journal of Psychiatric Research, 28,* 57-84.

Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 256-293). Newbury Park, CA: Sage.

Yu, C.Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Doctoral dissertation, University of California, Los Angeles.

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

Zumbo, B.D. (2003, November). *Fundamentals of item response theory: With applications in social and health research.* Data analysis and statistics seminar, organized by CRISP, UNB. Presented in Vancouver, British Columbia, Canada.

Zumbo, B. D., & Gelin, M.N. (in press). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological / community moderated test and item bias. *Educational Research and Policy Studies.*

Zumbo, B.D., Gelin, M.N., & Hubley, A.M. (2002). The construction and use of psychological tests and measures. *Encyclopedia of Life Support Systems*. France: United Nations Educational, Scientific and Cultural Organization Publishing (UNESCO-EOLSS Publishing).

Zumbo, B.D., & Hubley, A.M. (2003). Item bias. In Rocío Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 505-509). Thousand Oaks, CA: Sage Press.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: important advances in reliability and validity theory. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 73-92). Thousand Oaks, CA: Sage Press.

Zumbo, B. D., & Witarsa, P.M. (2004). *Nonparametric IRT methodology for detecting DIF in moderate-to-small scale measurement: Operating characteristics and a comparison with the Mantel Haenszel.* Paper presented at the annual meeting of the American Educational Research Association (AERA), San Diego, CA.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185–197.

Appendix A

Literature review of articles using the DIF MIMIC approach

Source	Construct	Measure / Instrument	Covariates[19] (direct effects)	Sample	Software	Estimation method
Fleishman, J.A. (2004, August 11). *Using MIMIC models to assess the influence of differential item functioning*. Retrieved from http://outcomes.cancer.gov/conference/irt/fleishman.pdf	General health status	SF-12 (Ware, Kosinski, & Keller, 1996) • 12 items total • ordinal responses (2-6 categories) • 2 latent variables: physical health (5 items), and mental health (4 items). 3 items load on both variables	Age (18-39; 40-59; 60-69;70+) Gender Education (< high school; high school; some college) Race/Ethnicity (White; black; Hispanic; other)	N=11,682 Adults >17 years old 55% female	MPlus	WLS
Jones, R.N. (2003). Racial bias in the assessment of cognitive functioning of older adults. *Aging & Mental Health, 7*, 83-102.	Cognitive functioning	Telephone interview of modified version of the Telephone Interview for Cognitive Status (TICS; Brandt, Spencer, & Folstein, 1988) • 7 item parcels • dichotomous and polytomous scored items	Age (50-54; 55-59; 60-64; 65-69; 75-79; 80-84; 85-90; 90+) Gender Education (0; 1-7; 8; 9-11; 13+) Income (<5k; 5-10k; 10-<20k; 40k+)	N=15,257 Adults ≥50 years old n=13,167 white, of 60% female n=2,090 black of which 64% female	MPlus	WLS

[19] Unless otherwise stated, covariates are dichotomous

Source	Construct	Measure / Instrument	Covariates[19] (direct effects)	Sample	Software	Estimation method
			Occupation Health conditions Health behaviors			
Fleishman, J.A., Spector, W.D., Altman, B.M. (2002). Impact of differential item functioning on age and gender in functional disability. *Journal of Gerontology: Social Sciences, 57B*, S275-S284.	Functional disability	Activities of daily living (ADL) • 6 dichotomous items Instrumental activities of daily living (IADL) • 5 dichotomous items The ADL and IADL comprise a undimensional scale (11 items) from the 1994 National Health Interview Survey Disability Supplement	Age (18-39; 40-69; 70+) Gender	N=5,750 adults ≥18 years old	MPlus	WLS
Jones, R.N. & Gallo, J.J. (2002). Education and sex differences in the Mini-Mental state examination: Effects of differential item functioning. *Journal of Gerontology: Psychological Sciences, 57B,* P548-P558.	Cognitive functioning (mental status)	Mini-mental state examination (MMSE; Folstein, Folstein, & McHugh, 1975) • 31 dichotomous items o 10 orientation items o 7 action & memory items o 6 concentration	Age (50-64 yrs; 65-74 yrs; 75+ yrs) Gender Education Ethnicity	N=8,556 adults ≥50 years old 61.6% female	MPlus	WLS

Source	Construct	Measure / Instrument	Covariates[19] (direct effects)	Sample	Software	Estimation method
		items ○ 8 language & praxis items • undimensional scale				
Mast, B.T. & Lichtenberg, P.A. (2000). Assessment of functional abilities among geriatric patients: A MIMIC model of the Functional Independence Measure. *Rehabilitation Psychology, 45,* 49-64.	General motor and cognitive functioning	Functional Independence Measure (FIM; Linacre et al., 1991) • 18 polytomous items (7-point rating scale): 13 ability to perform tasks, 5 measure aspects of social and cog. functioning • 2 composite scales: motor functioning, cognitive functioning	Age Gender Depression	N=718 geriatric inpatients 60-103 years old 68% female 62% African Americans	Amos 3.6	ML
Grayson, D.A., Mackinnon, A., Jorm, A.F., Creasey, H., & Broe, G.A. (2000). Item bias in the Center for Epidemiologic Studies Depression Scale: Effects of physical disorders and disability in an elderly community sample. *Journal of Gerontology: Psychological Sciences,*	Depression	CESD-20 (Radloff, 1977) • 20 polytomous items (4-point rating scale) • undimensional scale	Age Gender Marital status Disability (4 dichotomous variables) Physical disorders (10 dichotomous variables)	N=506 Adults 75+ years old (mean age 80.86, SD=4.17) 48% female	Amos 3.6	ML

Source	Construct	Measure / Instrument	Covariates[19] (direct effects)	Sample	Software	Estimation method
55B, P273-P282.						
Christensen, H., Jorm, A.F., Mackinnon, A.J., Korten, A.E., Jacomb, P.A., Henderson, A.S., & Rodgers, B. (1999). Age differences in depression and anxiety symptoms: A structural equation modeling analysis of data from a general population sample. *Psychological Medicine, 29,* 325-339.	Depression & anxiety	1. Anxiety (9 items) and depression (9 items) scales of Goldberg, Bridges, Duncan-Hones, & Grayson, (1988) – binary items 2. Personal disturbance scale (sAD) of the Delusions-Symptoms-States Inventory (DSSI; Bedord & Deary, 1997). 14 items (4-point scale): 7 anxiety items, 7 depression items 2 latent variables: depression and anxiety	Age (18-34; 35-49; 50-64; ≥65) Gender Education level Marital status Financial status	N=2,622 Australian 18-79 years old 52% female	Amos 3.6.1	ML
Gallo, J.J., Rabins, P.V. & Anthony, J.C. (1999). Sadness in older persons: 13-year follow-up of a community sample in Baltimore, Maryland. *Psychological Medicine, 29,* 341-350.	Depression	Composite diagnostic interview (CIDI; Wittchen, 1994) • 9 item parcels	Age (<65 *vs.* ≥65) Gender Education level Marital status Employment status Cognitive impairment (MMSE) Minority Status (white *vs.* other)	N=1,548 n=1,248 <65 years old; n=300 ≥65 years old > 60% female	LISCOMP	Limited-information GLS[20] estimator for dichotomous response

Source	Construct	Measure / Instrument	Covariates[19] (direct effects)	Sample	Software	Estimation method
Gallo, J.J., Anthony, J.C., & Muthén, B.O. (1994). Age differences in the symptoms of depression: A latent variable trait analysis. *Journal of Gerontology: Psychological Sciences, 49,* P251-P264.	Depression	Diagnostic interview schedule (DIS; Robins, Helzer, Croughan, & Ratcliff, 1981)	Age Gender Marital status Employment status Minority status MMSE (continuous 0-30)	N=6,541	LISCOMP	Limited-information GLS estimator for dichotomous response
Muthén, B.O. (1989). Using item-specific instructional information in achievement modeling. *Psychometrika, 54,* 385-396.	Math achieve-ment	Second International Mathematics Study (SIMS: Crosswhite et al., 1985). • 8 dichotomously scored items	Instructional coverage (i.e., opportunity to learn)	N=4,129 8th grade students	LISCOMP	Limited-information GLS
Muthén, B.O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28,* 1-22.	Math achieve-ment	Second International Mathematics Study (SIMS: Crosswhite et al., 1985). • 40 dichotomously scored items	Gender Ethnicity Instructional coverage Class type (remedial; enriched; algebra) Family background (4 categories) Fathers occupation Educational	N=3,724 8th grade students	LISCOMP	Limited-information GLS

Source	Construct	Measure / Instrument	Covariates[19] (direct effects)	Sample	Software	Estimation method
			aspirations (5 categories) Attitudes toward math (4 categories)			
Muthén, B.O., Tam, T.W.Y., Muthén, L.K., Stolzenberg, R.M., & Hollis, M. (1993). Latent variable modeling in the LISCOMP Framework: Measurement of attitudes toward career choice. In D. Krebs & P. Schmidt (Eds.), *New directions in attitude measurement, Festschrift for Karl Schuessler* (pp. 277-290). Berlin: Walter de Gruyter.	Career choice preferences	The National Longitudinal Study (NLS) • 10 polytomous items (3-point rating scale)	Gender, race (white vs. black), father's education and undergraduate major. 3 continuous variables: SES; SAT quantitative and verbal scores.	N=2645 students with a 4-year college degree	LISCOMP	WLS

Appendix B

Center for Epidemiologic Studies Depression Scales: CESD-10 and CESD-20

INSTRUCTIONS: Using the scale below, please circle the number for each statement that best describes how often you felt or behaved this way during the past week.

0 = Rarely or none of the time (less than 1 day)
1 = Some or a little of the time (1-2 days)
2 = Occasionally or a moderate amount of time (3-4 days)
3 = Most or all of the time (5-7 days)

DURING THE PAST WEEK:	Less than 1 day	1-2 days	3-4 days	5-7 days	**Factor Loadings** 10 item scale	20 item scale
1. **I was bothered by things that usually don't bother me.**	0	1	2	3	.669	.698
2. I did not feel like eating; my appetite was poor.	0	1	2	3	--	.533
3. I felt that I could not shake off the blues even with help from my family or friends.	0	1	2	3	--	.918
4. I felt that I was just as good as other people.	0	1	2	3	--	.462
5. **I had trouble keeping my mind on what I was doing.**	0	1	2	3	.744	.692
6. **I felt depressed.**	0	1	2	3	.857	.856
7. **I felt that everything I did was an effort.**	0	1	2	3	.743	.697
8. **I felt hopeful about the future.**	0	1	2	3	.532	.554
9. I thought my life had been a failure.	0	1	2	3	--	.751
10. **I felt fearful.**	0	1	2	3	.653	.658
11. **My sleep was restless.**	0	1	2	3	.597	.584
12. **I was happy.**	0	1	2	3	.680	.708
13. I talked less than usual.	0	1	2	3	--	.671
14. **I felt lonely.**	0	1	2	3	.658	.713
15. People were unfriendly.	0	1	2	3	--	.505
16. I enjoyed life.	0	1	2	3	--	.749
17. I had crying spells.	0	1	2	3	--	.729
18. I felt sad.	0	1	2	3	--	.853
19. I felt that people dislike me.	0	1	2	3	--	.605
20. **I could not get "going".**	0	1	2	3	.775	.734

- All 20 items are part of the CESD-20, whereas only the **bold** formatted items are part of the CESD-10.
- For the CESD-20 the items are summed after reverse scoring of items 4, 8, 12, and 16. Total CESD-20 scores range from 0-60, with higher scores indicating higher levels of general depression. For the CESD-10 the items are summed after reverse scoring items 8 and 12.
- The factor loadings were computed from the example data reported in Chapter II of this dissertation

Appendix C

Simulation study design

The first sub-study (Part A) investigates the Type I error rates in which two groups have *equal* sample sizes. Part A has a 2 x 2 x 2 x 3 factorial design: two scale lengths, by two item response distributions, by three sample size combinations, by two estimation methods. The second sub-study (Part B) investigates the Type I error rates in which two groups have *unequal* sample sizes. Similarly, Part B has a 2 x 2 x 2 x 4 factorial design, of which the variables are the same as in Part A except there are four sample size combinations instead of three.

	Item scale points	Items in scale	Item Distribution	Sample size combination	Sample size per group (equal)	Sample size per group (unequal)	Estimation Method**
		Incorporated into population file (SPSS)			Modified in LISREL code		
Part A (**equal** *n*)	4	10, 20	**Symmetric**	N/A	200 500 1000	N/A	ML WLS
	4	10, 20	**Skewed**	N/A	200 500 1000	N/A	ML WLS
Part B (**unequal** *n*)	4	10, 20	**Symmetric**	100:900 200:800 300:700 400:600	N/A	1000	ML WLS
	4	10, 20	**Skewed**	100:900 200:800 300:700 400:600	N/A	1000	ML WLS
	1 level	*2 levels*	*2 levels*	*4 levels*	*3 levels*	*1 level*	*2 levels*

** The ML and WLS estimation methods are based on Jöreskog's (2002) estimation methods.

Part A: 2 x 2 x 2 x 3 = 24 cell design

Part B: 2 x 2 x 2 x 4 = 32 cell design