Journal of Modern Applied Statistical Methods

Volume 6 | Issue 2 Article 22

11-1-2007

Operating Characteristics Of The DIF MIMIC Approach Using Jöreskog's Covariance Matrix With ML And WLS Estimation For Short Scales

Michaela N. Gelin
University of British Columbia

Bruno D. Zumbo
University of British Columbia, bruno.zumbo@ubc.ca

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 573-588

1538 – 9472/07/$95.00

Operating Characteristics Of The DIF MIMIC Approach Using Jöreskog's Covariance Matrix With ML And WLS Estimation For Short Scales

Michaela N. Gelin Bruno D. Zumbo
University of British Columbia

Type I error rate of a structural equation modeling (SEM) approach for investigating differential item functioning (DIF) in short scales was studied. Muthén's SEM model for DIF was examined using a covariance matrix (Jöreskog, 2002). It is conditioned on the latent variable, while testing the effect of the grouping variable over-and-above the underlying latent variable. Thus, it is a multiple-indicators, multiple-causes (MIMIC) DIF model. Type I error rates were determined using data reflective of short scales with ordinal item response formats typically found in the social and behavioral sciences. Results indicate Type I error rates for the DIF MIMIC model, as implemented in LISREL, are inflated for both estimation methods for the design conditions examined.

Key words: Type I error, multiple-causes model for DIF, Monte Carlo simulation.

Introduction

A variety of statistical methods have been developed over the years to aid the researcher in identifying DIF items for the purposes of (a) fairness and equity in testing, (b) evidence during litigation, (c) investigating whether item properties are changing over time, (d) dealing with a possible "threat to internal validity," and (e) 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 (Shimizu & Zumbo, 2005; Zumbo & Gelin, 2005; Zumbo & Hubley, 2003; Zumbo, 2007).

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 test s that contain many items (e.g., greater than 30) for the results to be reliable (e.g., Fidalgo, Mellenbergh, & Muñiz, 2000). Measures used in educational, psychological, and more broadly social and health science research (e.g., Rosenberg's Self-Esteem Scale, RSE; Rosenberg, 1965; Center for Epidemiologic Studies Depression Scale, CESD; Radloff, 1977) tend to have relatively fewer items, typically ranging from 3 to 30 items.

Reliability decreases with shorter scales and hence measurement error increases. Observed score DIF methods, such as logistic regression (LogR) or Mantel-Haesnzel (MH) that match on the observed score (e.g., total score or corrected total score often called the rest score), which has measurement error, are of particular concern in short scales because of the lower reliability and error of measurement. 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.

A latent variable approach is in line with the formal definition of DIF in which the underlying variable is the conditioning variable. In addition, a latent variable approach is recommended by Zwick (1990), Meredith (1993), Meredith and Millsap (1992), and Millsap and Meredith (1992), who argued that

Michaela N. Gelin is a Research Scientist at CTB/McGraw-Hill in Monterey, California. Bruno D. Zumbo is Professor of Measurement, Evaluation, and Research Methodology, and Department of Statistics. Email him at bruno.zumbo@ubc.ca

observed variable matching DIF methods such as the MH and LogR are not generally diagnostic of item bias. These observed score matching variable DIF methods use the manifest matching variable as a proxy for the latent matching variable and will only be appropriate when the two (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. Under the Rasch model, the observed total score is a sufficient statistic for the latent variable score – assuring the correspondence between the observed and latent matching variables.

Another situation where the observed and latent matching variables correspond is with long scales (a measure or scale with more than 30 items being combined into the composite score) in which all of the items are strong indicators (high factor loadings) of one underlying latent variable, assuming a one-dimensional scale. Shorter scales, containing up to 30 items, do not share this property even though they also may display uni-dimensionality. This rests partly on the notion that in item response theory modeling it is necessary to estimate the latent distribution, and that requires long scales for unbiased estimation and precision. The latent variable approach for investigating DIF in short scales rests on the structural equation modeling (SEM) multiple indicators multiple causes (MIMIC) method (Muthén, 1989).

In this study, Muthén's MIMIC DIF method was implemented using a relatively new covariance matrix available in LISREL for factor models for ordinal variables with covariate effects on the manifest and latent variables (Jöreskog, 2002; Moustaki, Jöreskog, & Mavridis, 2004).

Given that (a) short scales are typically found in the educational and psychological disciplines, (b) the SEM 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 study is to investigate the statistical properties of a relatively new covariance matrix for the SEM DIF MIMIC method. The proposed MIMIC methodology uses Muthén's (1989) SEM model computed via Jöreskog's covariance matrix. The Type I error rate of this DIF approach have not been investigated. The primary focus of this study is to examine the Type I error rate of the proposed DIF MIMIC approach by means of a simulation study under a variety of study conditions designed to reflect real responses to short scales with ordinal item formats typically found in the social and behavioral sciences.

A statistical test that maintains its Type I error rate is a valid test of the hypothesis. Type I error rates are often referred to as operating characteristics of a test. A Type I error rate, the 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 the conclusion is reached 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, for example, making a Type I error may be of great concern because of the matter of test fairness. The Type I error rate 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).

DIF MIMIC model

Although technical descriptions of Muthén's approach can be found, the description below is intended to be less technical with a broader audience of researchers who may be interested in SEM but less familiar with the psychometrics of DIF. The DIF MIMIC model was first proposed by Muthén in 1989. In general, this 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. This is a multiple-indicators, multiple-causes

(MIMIC) model which is akin to a latent variable ANCOVA. As Zumbo and Hubley (2003) noted, DIF methods are akin to ANCOVA or attribute-by-treatment interaction (ATI) methodologies.

The MIMIC model was introduced by Jöreskog and Goldberger (1975). It contains one or more latent variables that are simultaneously identified by both multiple endogenous item indicators, which comprise the scale under consideration, and by multiple exogenous causal variables such as background variables of gender or ethnicity. The MIMIC model allows the regression of latent variables on the background variables. Several uses of the MIMIC approach were 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 that allow one to study the relative importance of the predictors. Including multiple exogenous variables provides extra information about the measurement, which is particularly useful in detecting population heterogeneity (see Mast & Lichtenberg, 2000) and provides information to help validate scales, permitting the testing of the factor structure of a measure (Zumbo, 2005). The MIMIC approach allows for the detection of item-level measurement non-invariance (i.e., DIF).

Muthén's (1989) modeling approach, the MIMIC model, can be thought of in the context of an example using a 10-item scale, in this case of depression. The MIMIC model consists of three components: (1) a measurement model, (2) a regression model, and (3) a direct effects estimate. Figure 1 is a conceptual, or path, diagram to assist in the description of each of the components of the MIMIC DIF model.

The measurement component refers to the hypothesized relationship between a latent variable and its indicators. The measurement model relates the observed indicators (items) to the continuous latent variable, representing 'depression'. The latent variable is defined for this analysis by the 10 items that form the 10-item scale measuring depression. The relationship between the latent variable and its indicators or factor loadings, which are associated with the endogenous measurement model, 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 or residuals are set free in this model. Similarly, the measurement errors for the endogenous latent factors are set free.

The regression model relates the latent variable depression to the covariate sex or gender. The effect of the grouping variable, assumed to influence the latent factor, on the underlying latent construct is represented by an arrow from the latent grouping variable, the covariate, to the latent variable depression. This single directional relationship is set free in this model. This is analogous to regression of a continuous outcome variable onto one or more covariates such as gender, marital status, and education level.

The interpretation of the regression coefficient for the grouping variable will depend, of course, 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, depression, on gender would indicate that females have lower underlying depression than males. The third component is a direct effect estimate that detects measurement invariance in an item response associated with group membership. In other words, adding direct effects from the covariate(s) to the observed indicators, unmediated by the latent factor, incorporates DIF.

It is possible to have a directional arrow pointing from the grouping variable to the individual item being analyzed. This analysis is repeated for each individual item on the scale that one wishes to investigate DIF. More than one item could be tested at a time by specifying more than one direction arrow at a time. This path, or paths for more than one item at a time, represents a systematic difference in responses, controlling for the latent variable.

Having described the DIF MIMIC method there are numerous advantages for using this method: (1) follows the formal definition of DIF, (2) allows for multiple conditioning variables, (3) the combination of covariates (e.g., demographics, attitudes) indirectly represent group membership and hence group

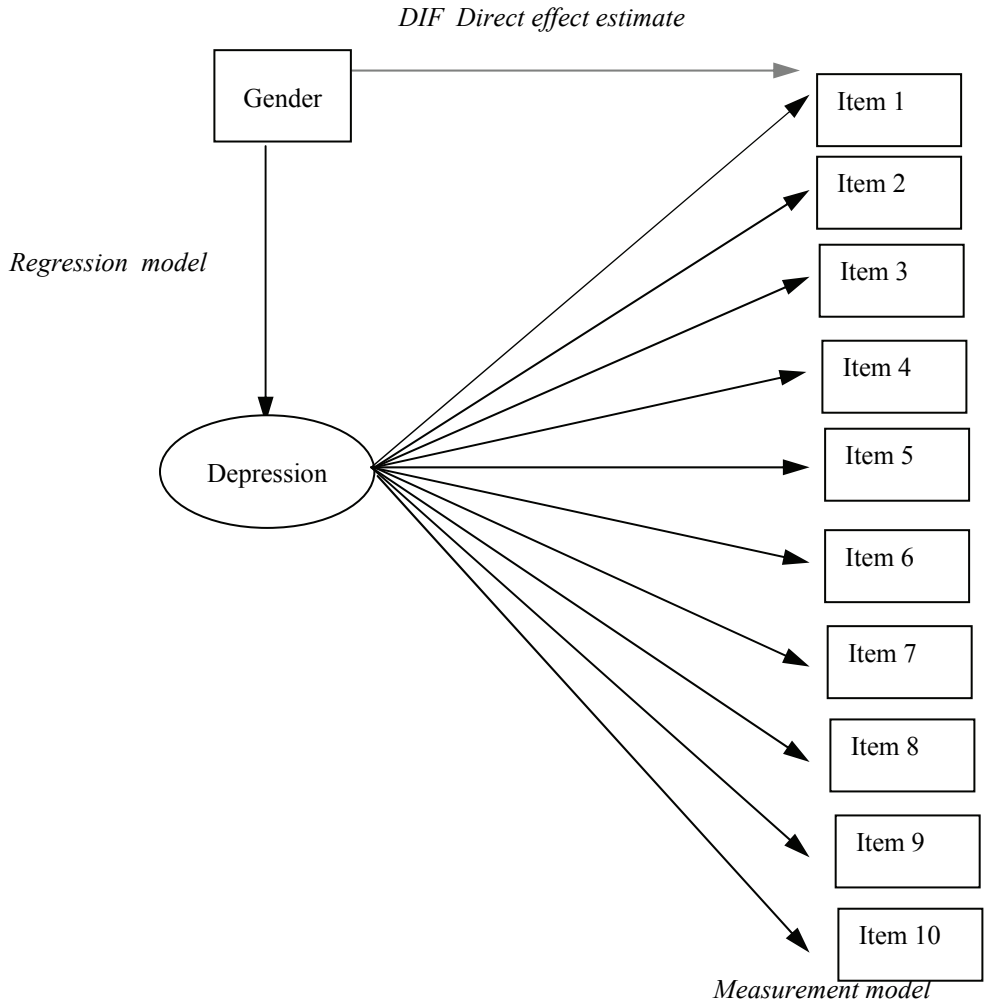

Figure 1. Conceptual (Path) diagram for the DIF MIMIC model for a 10-item scale.

membership does not have to be assigned a priori, (4) can be used with binary, ordinal, and mixed item formats, (5) can be used with multidimensional scales, (6) can model complex data structures involving complex item and test formats (testlets, item bundles, correlated errors), and (7) can be used with short scales. One limitation of this method is that it does not test for interactions (non-uniform DIF); it only investigates uniform DIF. The DIF MIMIC method only examines DIF that is attributable to differences in item difficulty (differences in thresholds). This method assumes the measurement model is the same in both groups (an implicit assumption in GLIM models such as LogR or MH, as well as conditional and unconditional DIF methods, see Zumbo & Hubley, 2003).

A Covariance Matrix for SEM DIF

Recently, Jöreskog (2002) and Moustaki, Jöreskog, and Mavridis (2004) described a new covariance matrix that takes

into consideration that one or more ordinal variables are observed jointly with a covariate(s) (possible explanatory variables). This covariance matrix makes it possible to implement Muthén's MIMIC DIF modeling approach in LISREL. 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. For technical details see Jöreskog (2002) and Moustaki, Jöreskog, and Mavridis (2004). The description below is intended to be less technical with a broader audience of researchers in mind.

In order to understand the advantage of Jöreskog's (2002) covariance matrix, a psychometric problem with be clarified. 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. It is also known from regression and correlation theory that for truly binary variables (e.g., grouping variables representing a contrast in a design matrix) the Pearson correlation can be used, and this models a difference in means for the continuous dependent or response variables in the model. The construction of a proper covariance matrix becomes a problem when there is a mix of ordinal and continuous data. Figure 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 in the specification of a design matrix representing group differences). The correct correlation between the test items in Figure 2, such as 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 (item1) and a continuous variable (age) is problematic because of their different variable formats.

If the data contain mixed variable formats, as is the case shown in Figure 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. This type of measurement error caused by using Pearson's correlation with ordinal data, such as Likert-type response formats, has long been debated in the literature (O'Brien, 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. 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.

Jöreskog's (2002) new method correctly treats the variables according to their variable type (see Figure 2). The ordinal item responses (items 1 through 3 in Figure 2) are correctly treated as ordinal variables, and the age and height variables are correctly treated as continuous covariates. This method allows computing the joint covariance matrix of the predictor and the variables underlying each of the ordinal variables (this is done simultaneously). Given that one or more ordinal item response variables are jointly observed with one or more manifest (observed) variables, such as 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. The joint covariance matrix may be computed for 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.

The statistical test of DIF is examined via (a) the t-statistic of the DIF direct effects coefficient, or (b) a Chi-squared difference test of two models, one with and a second without the DIF direct effects, wherein the nominal alpha of .05 is used in the test for DIF.

	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					

Figure 2. Example of a correlation matrix with mixed variable formats.

Methodology

Monte Carlo methods were used to examine the Type I error rates of Muthén's (1989) DIF MIMIC methodology computed via Jöreskog's (2002) covariance matrix with ML and WLS estimation methods. To provide a realistic set of values within the various study design variables described below in the simulation study, real item response data using the 10 and 20 item versions of the Center for Epidemiologic Studies Depression scale (CESD; Radloff, 1977) was used. The CESD is a widely used 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 items. The real response data came from 600 community-dwelling adults living in northern British Columbia (290 females; 310 males) who completed the 20-item CESD scale. 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. 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). This same item response data was also used to represent the short 10-item CESD scale. See Figure 3 for the specific items that make-up the 20- and 10-item versions.

Data from the CESD scale was chosen because it is a commonly used measure and hence is reflective of measures used in the social and behavioral sciences. Moreover, the scale and item characteristics (unidimensionality, scale length and item format) were representative of data typically found in psychological measures. Specifically, the 10 item short form (CESD-10: Andresen, Malmgren, Carter, Patrick, 1994) and the original 20 item (CESD-20: Radloff, 1977) CESD scales are essentially unidimensional (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.

The variables in this simulation study are seven sample size combinations (three equal and four unequal group combinations), two item response distributions (normal/symmetric and positively skewed), two scale lengths (10 and 20 items per scale), and two estimation methods (ML and WLS).

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 (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.

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. As found in the CESD 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. 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 behavioral sciences, the ordinal variables used in this study are conceptualized as observed ordered-categorical variables, y, wherein the underlying variable, y*, is completely unobserved (latent) and continuous. As the normally distributed latent variable increases beyond 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 one does 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 (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 approximately 0; Kurtosis approximately -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 approx. 1.6, kurtosis approx. 1.8).

Sample size combinations

Building on simulation designs in the literature (De Champlain & Gessaroli, 1998; Curran, Bollen, Paxton, Kirby, & Chen, 2002; Muñiz, Hambleton & Xing, 2001; Muthén & Kaplan, 1992), as well as from published

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	Factor Loadings 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

Note: 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.

Figure 3. Center for Epidemiologic Studies Depression Scales: CESD-10 and CESD-20.

literature using the CESD between 2000 and 2004 (PsycINFO search), seven combinations of equal and unequal sample sizes are considered.

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 (moderate-to-small-scale testing).

Estimation methods

Given that (i) the primary focus of this article 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) covariance matrix with ML (which involves the asymptotic covariance matrix and WLS estimation methods will be used. As previously described, Jöreskog's method was 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

First, 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. Based on the real data described above, the factor loadings for simulating the 10 and 20 item scales are listed in Figure 3. Using the population correlation matrix among the variables, continuous item response data, y*, of a specified population size, with normally distributed but independent (i.e., uncorrelated) continuous scores are generated and saved for each of two groups. A grouping variable is created and saved in the data set. For Group 1 in the equal sample size condition, the specified sample size is 50 000. However, for 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. For Group 2, the specified sample size 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 of 1:9, 2:8, 3:7, and 4:6 are 90 000, 80 000, 70 000, and 60 000, respectively, for Group 2. These normally distributed scores represent the (typically unobserved) latent scores from which ordered responses are generated.

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. The continuous scores are manipulated to mimic responses on a rating scale while simultaneously modifying the distributional shape of the data. Lastly, 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.

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 the number of replications. Based on Bradley's (1978) 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-statistic in testing that the estimate is statistically different from zero" (Byrne, 1998, p. 104).

Results

Psychometric properties of the data

Before sampling from the population data files it is important to verify that the simulated data has the desired psychometric properties. A confirmatory factor analysis (CFA) with the polychoric correlation matrix and weighted least squares (WLS) estimation procedure with the asymptotic covariance matrix (Byrne, 1998) was computed using LISREL 8.54 (Jöreskog & Sörbom, 2003b). 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 data

Different population data sets were created for the equal and unequal sample size conditions. Four population data sets (two levels of the number of items in the scale by two levels of item distributions were created for the equal sample size conditions (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 (the 20-item scales) had better reliabilities.

Monte Carlo

For each of the 1000 replications, the model fit and test statistics (t and χ^2) were recorded. The asymptotic covariance matrix of the estimated coefficients 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).

There are a few simulation cells in which the first run of the simulation resulted in all of the replications being non-computable, as 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 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 (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 really 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 overall model fit values over the 1000 replications for the DIF MIMIC models with ML estimation method 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.

The mean fit statistics for the DIF MIMIC model conducted with WLS estimation method showed that 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. Similarly, 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 article. 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.

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 were examined since there is only one path (direct effects estimate) one is able to also test if the path is equal to zero via the *t* statistic. As expected, the *t* statistic is also inflated and follows the same patterns as the chi-square statistic reported in the results tables.

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 (λ_{12} in Figure 1). Using this chi-square value, the proportion of rejections was counted, which represent the Type I error rates, based on the chi-square p-value, with p-values less than 0.05 leading to a decision not to reject the hypothesis. The chi-square rejection rates (Type I error rates) across estimation method, scale length, distributional condition, for the equal sample size combinations are shown in Table 1.

For the symmetrically distributed 10-item data using ML estimation, the Type I error rate was inflated (7.7% - 10.3%) for all three sample size conditions. Similarly, for the skewed 10-item data using ML estimation, the Type I error rate was also inflated (12.5% to 14.8%) for all sample size conditions. Table 1 also 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 three 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.

In terms of the 10-item scale with WLS estimation (Table 1), 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 1) showed even higher Type I error rates ranging from 24.9% - 46.7%. As one can also see, there were no valid chi-square for the 200:200 sample sizes combinations due to the problem of a non-positive definite matrix.

Table 1. Empirical Type I error rates of the Chi-squared Test of the DIF MIMIC model across estimation method, scale length, distributional condition, for the *equal* sample size combination.

Estimation method	Scale length	Distribution		Sample size combinations		
				200:200	500:500	1000:1000
ML	10-item	Symmetric	*Reject*	.103	.093	.077
			Valid reps	964	995	993
		Skewed	*Reject*	.126	.148	.125
			Valid reps	957	991	995
	20-item	Symmetric	*Reject*	.118	.108	.147
			Valid reps	626	508	470
		Skewed	*Reject*	.162	.163	.116
			Valid reps	660	575	481
WLS	10-item	Symmetric	*Reject*	.235	.131	.099
			Valid reps	948	996	997
		Skewed	*Reject*	.283	Not	.147
			Valid reps	972	computable	991
	20-item	Symmetric	*Reject*	Not	.341	.249
			Valid reps	computable	988	993
		Skewed	*Reject*	Not	.467	.305
			Valid reps	computable	959	957

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

Part B: Unequal sample size condition

Model fit

The fit statistics for the DIF MIMIC model conducted with ML estimation 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 <.5 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.

Type I error rates

As in Part A, the chi-square values were examined and used to evaluate the Type I error rates of the DIF MIMIC model under a variety of conditions. The chi-square rejection rates (Type I error rates) for the unequal sample size conditions are shown in Table 2.

For the symmetrically distributed 10-item data using ML estimation, the Type I error rate was inflated (9% - 11.6%) 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 14.3%) for all sample size conditions.

the MIMIC method is the most appropriate

Table 2. Empirical Type I error rates of the Chi-squared test of the DIF MIMIC model across estimation method, scale length, distributional condition, for the *unequal* sample size combinations.

Estimation method	Scale length	Distribution		Sample size combinations 1:9	2:8	3:7	4:6
ML	10-item	Symmetric	*Reject*	.097	.116	.090	.103
			Valid reps	982	988	996	996
		Skewed	*Reject*	.136	.134	.134	.143
			Valid reps	974	983	991	994
	20-item	Symmetric	*Reject*	.123	.105	.098	.124
			Valid reps	528	513	479	467
		Skewed	*Reject*	.159	.113	.143	.163
			Valid reps	536	503	490	491
WLS	10-item	Symmetric	*Reject*	.115	.126	.114	.125
			Valid reps	979	994	999	995
		Skewed	*Reject*	.138	.162	.171	.178
			Valid reps	982	995	996	998
	20-item	Symmetric	*Reject*	.188	.211	.207	.232
			Valid reps	903	966	998	999
		Skewed	*Reject*	.224	.259	.279	.320
			Valid reps	991	999	999	982

For the symmetrically distributed 20-item data using ML estimation, the Type I error rate was also moderately inflated (9.8% - 12.4%) 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.

In terms of the 10-item scale with WLS estimation (see Table 2), the symmetrically distributed data showed inflated Type I error rates ranging from 11.4% to 12.5%. Likewise, the skewed data was also inflated (13.8% to 17.8%). The 20-item scale using WLS estimation (see Table 2) showed even higher Type I error rates for both the symmetrically distributed data (18.8% to 23.2%) and the skewed data (22.4% to 32%).

Discussion

Given that short scales are typically found in the educational and psychological disciplines and method for investigating DIF in short scales, the primary purpose of this article was to investigate the Type I error rates for this DIF method as implemented using Jöreskog's (2002) covariance matrix with ML and WLS estimation methods. As mentioned in the introduction of this article, no previous study had examined the Type I error rates for the DIF MIMIC method let alone its implementation in the LISREL software. Accordingly, the primary focus of this article 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.

The results of this study clearly show 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 WLS estimation 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.

Based on the results from the current study we caution researchers against the use of the DIF MIMIC method with Jöreskog's methods in LISREL. 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 DIF MIMIC approach currently available in LISREL. Moreover, for studies that have used this DIF MIMIC method (with the new covariance matrix described above), 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.

References

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.

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.

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.

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

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.

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.

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

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.

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

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.

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.

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.

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.

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

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

Moustaki, I., Jöreskog, K. G., & Mavridis, D. (2004). Factor models for ordinal variables with covariate effects on the manifest and latent variables: A comparison of LISREL and IRT approaches. *Structural Equation Modeling, 11,* 487-513.

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., & 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.

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, 28,* 491-497.

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

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.

Shimizu, Y., & Zumbo, B. D. (2005). A Logistic Regression for Differential Item Functioning Primer. *Japan Language Testing Association Journal, 7,* 110-124.

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.

Zumbo, B. D. (2005). Structural Equation Modeling and Test Validation. In Brian Everitt and David C. Howell, *Encyclopedia of Behavioral Statistics*, (pp. 1951-1958). Chichester, UK: John Wiley & Sons Ltd.

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. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233.

Zumbo, B. D., & Gelin, M.N. (2005). A Matter of Test Bias in Educational Policy Research: Bringing the Context into Picture by Investigating Sociological / Community Moderated (or Mediated) Test and Item Bias. *Educational Research and Policy Studies, 4, 223-233.*

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.

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