SECCIÓN METODOLÓGICA

Psicológica (2009), 30, 343-370.

Impact of Differential Item Functioning on Subsequent Statistical Conclusions Based on Observed Test Score Data

Zhen Li & Bruno D. Zumbo[1]

University of British Columbia, Canada

This simulation study investigated the impact of differential item functioning (DIF) on the Type I error rate and effect size of the independent samples *t*-test on the observed total test scores. Five studies were conducted: studies one to three investigated the impact of unidirectional DIF (i.e., DIF amplification) on the Type I error rate and effect size of the independent sample t-test, studies four and five investigated the DIF cancellation effects on the Type I error rate, and effect size of the independent sample t-test. The Type I error rate and effect size were defined in terms of latent population means rather than observed sample means. The results showed that the amplification and cancellation effects among uniform DIF items did transfer to test level. These findings highlight the importance of screening DIF before conducting any further statistical analysis.

Differential item functioning (DIF) has been widely studied in educational and psychological measurement. For recent reviews please see Camilli (2006) and Zumbo (2007). Previous research has primarily focused on the definitions of and the methods for detecting DIF. It is well accepted that the presence of DIF might degrade the validity of a test. There is relatively little known, however, about the impact of DIF on later statistical decisions when one uses the observed test scores in data analyses and corresponding statistical hypothesis tests. For example, let us imagine that a researcher is investigating whether there are gender differences on a language proficiency test. What is the impact of gender-based differential item functioning on the eventual statistical decision of whether the group means (male versus female) of the observed scores on the language

[1] Send correspondence to: Professor Bruno D. Zumbo. Department of ECPS, 2125 Main Mall. University of British Columbia. Vancouver, B.C., Canada., V6T 1Z4. Email: bruno.zumbo@ubc.ca

proficiency test are equal? There is remarkably little research to help one directly answer this question.

DIF may be present in a test because either (a) DIF analyses have not been used as part of the item analyses, (b) it is there unbeknownst to the researcher, as an artifact of DIF detection being a statistical decision method, and hence true DIF items may be missed, or (c) as a result of the practice of leaving items flagged as DIF in a test. Irrespective of how the DIF items got there, it is still unknown how such DIF items affect the subsequent statistical results and conclusions, particularly, the Type I error rate and effect size of hypothesis tests from observed score test data.

In order to directly answer this research question of the effect of DIF items on the eventual statistical conclusions from the test total scores, we conducted five interrelated simulation studies wherein we simulated population test data using item response theory (IRT) with varying degrees of DIF -- i.e., number of items exhibiting DIF and the magnitude of DIF. In order to answer the hypothetical researcher's research question, the observed (number correct) scores were then subjected to a t-test to test for the equality of sample means. Throughout this research we focus on the (Type I) error rates and effect sizes of the t-test under the null hypothesis of equal means. We did not investigate the statistical power (i.e., the results under the case when the population means are not equal) due to space limitations and due to the fact that Type I error rates need to be established before one can interpret the results of statistical power. The statistical power study is forthcoming.

It is important to note that the Type I error rate herein, in essence, was the probability of rejecting the null hypothesis when the latent means (rather than the observed test score means) were equal across groups. That is, using IRT one notes that an item response is a function of item parameters and examinee ability. By definition, when DIF items were retained in a test, these DIF items might result in differences in item responses of different group of examinees of comparable abilities. Accordingly, our research question more formally can be stated as: What is the probability of rejecting the null hypothesis of equal observed test score means when the latent means are equal but DIF is present in the test? Likewise the effect size reflects those settings in which the latent variable population means are also equal.

Based on tangentially related research that investigates the impact of DIF on person parameter estimates (i.e., the latent variable score) from IRT, scale scores, and predictive validity (e.g., Drasgow, 1987; Maller, 2001; Roznowski, 1987; Roznowski & Reith, 1999; Rupp & Zumbo, 2003, 2006;

Shealy & Stout, 1991, 1993; Wells, Subkoviak, & Serlin, 2002) we predict that the Type I error rate and effect sizes will be inflated, however, the extent and under what conditions it will be inflated are unknown. To answer the question of how much DIF effects the eventual statistical conclusions we are interested in two testing situations: (a) several DIF items consistently favor one group, and hence, of course are against the other one; (b) some of the DIF items favor one group and some favor the other. The first situation represents what we refer to as DIF amplification effects (which were the focuses of studies one, two, and three) whereas the second situation as DIF cancellation effects (which were the focuses of studies four and five). Of course, other test data situations may arise but given that this is the first study of its kind we wanted to address two fairly straightforward, but of course plausible, testing situations.

Five inter-related computer simulation studies were conducted. The first study focused on the amplification effects of DIF on the Type I error rate of the hypothesis test of equality of means of the observed test scores. The second simulation study focused on the amplification effects of DIF on the effect size. The third simulation study investigated the impact of varying the test item parameter values on the Type I error rate of the subsequent t-test of the observed score means. Note that in studies one and two the items used to generate DIF were sampled in a fixed manner. Influences of the different values of the item parameters on the Type I error rate were not considered. Therefore, study three was added to confirm the generalizability of the results of this study. Study four focused on the impact of DIF cancellation effects on Type I error rate, and finally study five focused on the impact of DIF cancellation effects on the effect size. In order to organize the findings and convey them in a clear manner, we organized the five simulation studies into two sections: section one being the amplification effects and section two the cancellation effects. Each section will have a brief discussion section and then a general discussion will be reserved for the end.

We focused our research on the widely used two independent sample case of testing the equality of observed score group means; that is, the conventional (pooled variances) independent samples t-test. This scenario reflected the all too widely observed case wherein researchers investigate mean differences on their test data (a) without having first studied whether DIF exists, or (b) when one conducts DIF analyses but decides to retain the items even though DIF is found. It is important to note that the DIF was aligned with the hypothesis test of mean differences itself (i.e., there were potential *gender* DIF items when one was investigating *gender* differences on the observed test scores). Without loss of generality to other assessment

and measurement approaches (such as psychological or health measurement), we will use educational achievement testing and gender DIF as our example to contextualize our research. Of course, the DIF could be due to test translation or adaptation, or any other situation that results in a lack of measurement invariance (Zumbo, 2007).

SECTION ONE

Impact of Differential Item Functioning (DIF) on Statistical Conclusions, I: Amplification Effect

Study One: Type I Error Rates

The purpose of this first simulation study was to document the effect of DIF items on the eventual Type I error rates of the t-test on the observed (number correct) total test score data. In this study we focused on the amplification effect of DIF item. That is the situation where DIF items favor a group consistently.

METHODS

Simulation factors.

The simulation factors manipulated in this study were magnitude of DIF, number of DIF items, and the sample size. There are three levels of magnitude of DIF -- small, moderate, and large as defined by Raju's area statistic of .4, .6, and .8 (Raju, 1988), four levels of number of DIF items (1, 4, 8, and 16 items out of 38 items in the test), and four levels of sample size (25, 50, 125, and 250 examinees per group). In addition, for comparison purposes, we investigated the no DIF condition as a baseline for the four sample sizes – it is expected, of course, that in this condition the observed Type I error rate would be at the nominal level. Therefore, for the Type I error rate simulation, the resultant simulation experiment was a 3x4x4 completely crossed factorial design; and in addition 4 no-DIF conditions (for the four sample sizes) resulting in a total of 52 cells in our simulation design.

We focused our investigation on binary items. The data were generated using item response theory (IRT). In order to examine the amplification effect of DIF items, we focused on unidirectional uniform DIF. Unidirectional DIF (Shealy & Stout, 1991, 1993) occurs when the DIF items are against the same group for all levels of ability (θ). Thus in this study DIF items were simulated consistently favoring the reference group. In addition, we adopted Zumbo's (2003) simulation design and

therefore we did not vary the test length, and we used real item parameters based on the TOEFL test to generate our item data for a 38 item test.

The first factor of interest in this study was the magnitude of DIF. Theoretically, we expected larger magnitude of DIF would enlarge the differences in item responses between groups and hence the combined DIF effect across items might result in greater Type I error rate. Previous studies (French & Maller, 2007; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993) also simulated DIF in this manner. Following these earlier studies, the uniform DIF items were simulated by shifting the b-parameter in the focal group to manipulate the area between two item response functions. In the situations wherein there was more than one DIF item, all the items in that situation had the same magnitude of DIF. That is, for the ease of interpretation, we did not investigate the effect of having a mixed magnitude of DIF – e.g., for the situation in which more than one item had a DIF, all of the items had, for instance, a .40 DIF effect.

Similarly, we expected that the Type I error rate might be affected by the proportion of DIF items in the test. In the unidirectional uniform DIF case, the hypothesis was that the more DIF items were retained in the test the larger differences would be resulted in observed response data across groups, then the more likely that the Type I error rate would be affected. Note that following Zumbo (2003) we did not varying the total of number of items in the test – we investigated 1, 4, 8, and 16 DIF items, out of a total 38 items.

Sample size was another factor that might affect the Type I error rate in terms of latent means as the larger the sample size the more likely one is to reject the null hypothesis. Sample size was set equal in both comparison groups.

Simulation procedures.

Following Zumbo (2003), the 38 item parameter estimates for the Structure and Written Expression section of Test of English as a Foreign Language (TOEFL) were used to simulate the data in the various conditions. The means and standard deviations of item parameters in the reference and focal groups were presented in Table 1 and Table 2.

Table 1

Means (M) and Standard Deviations (SD) of Item Parameters in Reference Group

Item Parameter	M	SD	Minimum	Maximum
a-parameter	0.986	0.304	0.535	1.890
b-parameter	-0.133	0.861	-2.494	1.537
c-parameter	0.231	0.106	0.029	0.448

Table 2

Means (M) and Standard Deviations (SD) of b-parameters under Different Simulation Conditions in Focal Group

Number of DIF items	Small DIF (Raju's area = 0.4)		Moderate DIF (Raju's area = 0.6)		Large DIF (Raju's area = 0.8)	
	M	SD	M	SD	M	SD
1	-0.121	0.83	-0.115	0.816	-0.109	0.804
4	-0.081	0.875	-0.054	0.892	-0.028	0.915
8	-0.028	0.886	0.025	0.916	0.078	0.957
16	0.083	0.884	0.19	0.923	0.298	0.978

Examinee response data was generated using a three-parameter unidimensional logistic item response model (Birnbaum, 1968) as shown in equation (1),

$$P_i(\theta) = c_i + \frac{(1-c_i)}{1+\exp[-1.7a_i(\theta-b_i)]}, \quad (1)$$

where a_i, b_i, and c_i are the item discrimination, difficulty, and guessing parameters of item i, respectively. The latent variable is denoted as θ, whereas $P_i(\theta)$ denotes the probability of answering item i correctly with

ability θ. Five thousand replications were conducted in each cell (i.e., study condition) of the simulation design; therefore, each Type I error rate value was based on 5000 replications of the simulation experiment.

Three steps were conducted to generate the item response data for Type I error rate studies.

Step #1: In the first step we generated the reference population data. In this step, the ability values, θ, for the reference group were generated from a standard normal distribution (M = 0, SD = 1). The probabilities, $P_i(\theta)$, were calculated using equation (1) and the values of a, b, c, and the generated θ. Then, uniformly distributed random numbers with interval [0, 1] were generated. To obtain the binary item response, the item response probabilities, $P_i(\theta)$, were converted to *1*s when the probability was larger than the corresponding random number, whereas the probabilities were converted to *0*s otherwise (Hambleton & Rovinelli, 1986). Next, the observed total test scores (number correct) were computed. And finally, samples with a particular sample size were randomly sampled from the reference population.

Step #2: In step two we generated the focal population data using exactly the same procedures except that some of the item parameter values were changed to reflect DIF on selected items and depending on the cell of the simulation design.

Step #3: In step three the generated two populations were merged into one file and the independent sample t-tests were conducted, and the Type I error rates were computed as the number of rejections of the null hypothesis out of the 5000 replications. Our nominal significance level was 0.05 throughout this study. Therefore, empirically, the Type I error is defined as the proportion of times that a true null-hypothesis was falsely rejected at the 0.05 level.

Analysis of the Type I error rate simulation results.

We used the Bradley (1978) approach to documenting the inflation in Type I error rate. Bradley defined three different levels of Type I error rate robustness which he terms as fairly stringent, moderate, and very liberal. Thus, for a Type I error rate of .05, the fairly stringent criterion for robustness requires the empirical Type I error rate lie between .045 and .055. The moderate criterion requires the empirical Type I error rate lies between .040 and .060. And the very liberal criterion requires the empirical Type I error rate lies between .025 and .075. Please recall from the definition above that these proportions of rejected t-tests were the Type I

error rates because the population means for the latent variable, θ, were equal.

In addition to Bradley's descriptive method we also used regression modeling to investigate the effect of the simulation factors, treating the design and analysis as a type of response surface modeling (Zumbo & Harwell, 1999). The dependent variable is a proportion (i.e., the empirical Type I error rate based on 5000 replications) therefore the logit transformation was applied (Cohen, Cohen, West, & Aiken, 2003, p. 240). The regression modeling was conducted in two steps. In the first step a model was fit with main effects, and then an assessment was made whether the interactions statistically contributed to the model. In the second step graphical methods were used to describe the main effects and/or interactions.

RESULTS AND CONCLUSIONS

Table 3 lists the simulation results and the description based on Bradley's criteria. The Type I error rate for different sample sizes were computed for the no DIF conditions to establish baselines for comparisons with the conditions wherein different DIF conditions were manipulated. Under the no DIF condition, as shown in second column of Table 3, the Type I error rates range, as expected, from 0.052 to 0.053 for sample size from 25 to 250 per group (the column labeled 'No DIF' in Table 3). This also serves as a check on our simulation methodology.

Table 3 also displays the results of the Type I error rates for the case of the DIF amplification effect for the (a) different magnitudes of DIF, (b) number of DIF items, and (c) sample size combinations. Please recall that DIF items in this study were all simulated favoring the reference group. The far left column of Table 3 lists the sample sizes per group and next to it is the baseline Type I error rates. The remaining nine columns were divided into three magnitudes of DIF (i.e., Raju's area of .40, .60, and .80). Within each magnitude of DIF, the four columns represent the cases wherein there are 1, 4, 8, and 16 DIF items. For example, focusing on Raju's area of .40, with one uniform DIF retained in the test (column 2), the Type I error rates were between 0.45 and 0.56 for the sample sizes of 25 to 250 per group. In this situation none of the Type I error rates were inflated for all studied sample sizes using Bradley's (1978) moderate criterion. As the sample size increased to 250, the Type I error rate inflated with large DIF compared against the moderate criterion. In terms of categorizing the resultant Type I error rates:

- When only one of the 38 items had DIF, the Type I error rate met the moderate criterion, except for Raju's area of .80 with 250 examinees per group wherein the Type I error rate only met the liberal degree of robustness.
- Irrespective of the magnitude of DIF and sample size, with 16 out of 38 of the items having DIF the Type I error rate was inflated. Likewise, for Raju's area of .60 and .80 with 8 out of 38 of the items having DIF the Type I error rate was inflated, and
- For 4 or 8 out of 38 items having DIF, the classification of the Type I error rates were dependent on the sample size and the magnitude of DIF – ranging from moderate to liberal inflated Type I error rates.

These classifications of DIF are informative in terms of deciding whether one should treat the Type I error rate as too large and hence invalidating the t-test of the hypothesis of equal population observed score means, but these classifications do not clearly provide a description of how the simulation factors, as main effects or interactions, effect the Type I error rate of the t-test. To address this latter question, we conducted the regression analysis of the simulation results treating each factor in the simulation (Number of DIF items, Magnitude of DIF, Sample size) and the interactions among them as explanatory variables in the multiple regression analysis (Zumbo & Harwell, 1999). In the first step of the modeling, the main effects were entered into the model with a resultant R-squared of 0.782 ($F(3, 44) = 52.53$, $p < .0001$), then the three two-way interactions were entered into the model for a resulting R-squared of 0.973 (R-squared change was statistically significant, $F(3, 41) = 96.24$, $p < .0001$), and finally the three-way interaction was entered into the model resulting in an eventual model R-squared of 0.985 (R-squared change was statistically significant, $F(1,40) = 33.67$, $p < .0001$). Please note that because of the use of interaction terms all of the explanatory variables were first centered before product terms were computed for the interactions. Clearly, the three-way interaction was statistically significant. Upon close inspection of Table 3, for each magnitude of DIF, it can be clearly seen that the inflation of the Type I error rate increases as the number of DIF items and the sample size increase.

Table 3

Type I Error Rates of t-test under Amplification Effect for Different Sample Sizes, Number of DIF, and Magnitude of DIF Combinations

	Magnitude of DIF												
	No DIF	Raju's area 0.4				Raju's area 0.6				Raju's area 0.8			
Number of DIF	0	1	4	8	16	1	4	8	16	1	4	8	16
N(per group)													
25	0.053	0.045	0.058	0.060↑	0.104↑↑	0.057	0.051	0.079↑↑	0.172↑↑	0.047	0.059	0.095↑↑	0.265↑↑
50	0.052	0.047	0.052	0.074↑	0.165↑↑	0.045	0.060↑	0.105↑↑	0.308↑↑	0.048	0.068↑	0.131↑↑	0.488↑↑
125	0.052	0.054	0.056	0.120↑↑	0.341↑↑	0.055	0.077↑↑	0.181↑↑	0.628↑↑	0.054	0.092↑↑	0.278↑↑	0.853↑↑
250	0.053	0.056	0.069↑	0.175↑↑	0.601↑↑	0.057	0.094↑↑	0.319↑↑	0.898↑↑	0.063↑	0.126↑↑	0.505↑↑	0.990↑↑

Note:

Type I error rate	Bradley (1978) criterion
$\alpha < 0.055$	Meet the stringent criterion
$0.055 \leq \alpha < 0.060$	Meet the moderate criterion
$0.060 \leq \alpha < 0.075$, ↑	Violates the moderate but meets the liberal criterion
$\alpha \geq 0.075$ ↑↑	Violates the liberal criterion, therefore inflated

Study Two: Effect Size

A second computer simulation study was designed to investigate the effect of DIF on the effect size of the independent sample t-test when DIF items were retained in the tests. Cohen's d is the appropriate effect size measure to use in the context of a t-test of independent means; d is defined as the difference between two means divided by the pooled standard deviation for those means. We compute d for both the observed total scores and latent variables; which allow us to index the impact of DIF on the effect size. For the observed score d, the means and standard deviations are computed from the observed total test scores, whereas for the latent variable d the mean and standard deviations are computed from the latent variable scores.

As in study one, the observed score effect size is computed when the latent means (rather than the observed group test score means) were equal across groups. Therefore, our research question can be stated as: What is the effect size for the observed test score means when the latent means are equal but DIF is present in the test?

METHODS

The simulation factors manipulated in this study, as well as the simulation methodology, were the same as those in study one except for one experimental factor, sample size. That is, we manipulated number of DIF items and the magnitude of DIF. As in study one, there were four levels of number of DIF items (1, 4, 8, and 16 items out of 38 items in the test), and three levels of magnitude of DIF -- small, moderate, and large as defined by Raju's area statistic of .4, .6, and .8 (Raju, 1988). In addition, we investigated the no DIF condition as a baseline for the four sample sizes for comparison purposes. This resulted in a 4x3 completely crossed factorial design and an additional no-DIF condition resulting in a total of 13 cells in our simulation design.

Note that like Zumbo (2003) we were not interested in the sample-to-sample variation in effect size estimates but instead focused on (the population analogue of) the bias in effect size. With this in mind we simulated 10,000 examinees in each cell of the simulation design for our pseudo-populations. For each cell we computed the effect size for the observed total test score mean difference and for the latent mean difference (and their corresponding standard deviations). Because both the observed score and latent variable effect size values are on the same metric (both being standardized) we were able to compute the difference between them as an index of how much the DIF biases the effect size.

RESULTS AND CONCLUSIONS

As was noted above, because the effect sizes are on the same metric (i.e., both are standardized), Table 4 lists differences between the effect sizes of the observed and the latent variable score for the three magnitudes of DIF and the four different number of DIF items (1, 4, 8, 16). One can see that when no DIF exists the effect sizes of the latent mean and observed mean are, as expected, equal (to the third decimal point and hence within sampling). Again, this serves as a check of the simulation methodology. However, when DIF (unidirectional uniform DIF) appeared in the test, the effect size differences increase. The more DIF items one has in their test and the larger the DIF, the greater the effect size differences with the observed mean differences being spuriously inflated by the presence of DIF.

Using the same analysis methodology used in study one, the simulation results were analyzed using regression analysis with effect size differences as the dependent variable and magnitude of DIF, number of DIF item, and their interaction as independent variables. The model is statistically significant ($F(3, 8) = 287.9$, $p < .0001$) with an R-squared of 0.991 and an adjusted R-squared of .987. All the predictors are statistically significant, including the interaction term. Upon close inspection of Table 4, it can be clearly seen that the effect size differences increase as the number of DIF items and the magnitude of DIF increase.

A research question naturally arises from our findings to this point. Given that in studies one and two we treated the item parameter values as fixed values we do not know the impact of varying item difficulty, discrimination and guessing on the Type I error rate.

Study Three: Impact of Item Parameter Values on Type I Error Rates

Studies one and two investigated the impact of DIF on the Type I error rate and effect size; however, the items used to generate DIF were sampled from 38 items in a fixed form of a test. Influences of the values of the item parameters were, therefore, not considered in either of the first two studies. Study three focuses on the impact of varying item parameter values on the Type I error rate. In essence, study three is an empirical check as to whether the findings in study one are generalizable to other item parameter values than just the ones under investigation therein – in essence, an investigation into the external validity of the findings in study one.

Table 4

Differences Between Latent and Observed Effect Size

Number of DIF Items	ES(Observed Score) – ES(Latent Variable Score)
0	-0.001
	Small DIF (Raju's area = 0.4)
1	0.003
4	0.029
8	0.088
16	0.196
	Moderate DIF (Raju's area = 0.6)
1	0.021
4	0.049
8	0.136
16	0.297
	Large DIF (Raju's area = 0.8)
1	0.033
4	0.063
8	0.167
16	0.396

In this study we investigated the impact of item properties (values of a-, b-, and c-parameters), magnitude of DIF (quantified by Raju's area) and Δb (*b*-parameter differences between groups) and sample size on Type I error rate. This study focused on the case of one DIF item (i.e., the case in which the Type I error rates are protected in study one). We did not investigate the case of more than one DIF item because in those cases the Type I error rate is already inflated and hence of little practical value to investigate how the item parameter values may further inflate the error rates.

This study is different in purpose and design than the typical computer simulation study in psychometric research. The typical psychometric simulation study, such as studies one or two, have, in experimental design terms, fix experimental factors. Therefore, as is well known in experimental design, generalizing beyond the values of the fixed factors is not recommended. If one wants to investigate the generalizability of findings from a fixed factor (computer simulation) experiment one needs to randomly sample the values of the levels of the manipulated factors; hence, in essence, creating a random factor. The present study does just that by sampling item parameter values and magnitudes of DIF to investigate whether the protected Type I error rate when one has only one item exhibiting DIF generalizes to other item parameter values than those used in study one.

METHODS

Therefore, different from study one in which item parameters for the item exhibiting DIF were real parameters from TOEFL, DIF item parameter values and the b-parameter differences between groups in this study were randomly generated from normal and uniform distributions.

Let us first describe the simulation design in broad strokes with the details to follow. One can think of the simulation running in three large steps.

Step 1: We followed study one and used the same number of items and item parameters (1 DIF item out of 38 total items) and sample sizes (25:25, 50:50, 125:125, and 250:250).

Step 2: For each sample size condition we generated 50 random item parameter values for the DIF item – recall that the other 37 items parameters were the same as those in study one. This resulted, in essence, in 50 runs for each sample size condition. For each of these runs, as in study one, 5000 replications were conducted using IRT to generate the data to computed the resultant Type I error rate for that run. Note that there are 50

runs for each sample size condition resulting in a total of 200 Type I error rates (one for each run) from the simulation.

Step 3: The resultant Type I error rates and their respective DIF item parameter values for the 200 runs (50 runs for each sample size combination) were then read into a statistical software package for statistical analysis.

Following similar approaches in the research literature (e.g., Hambleton & Rovinelli, 1986; Hambleton & Swaminathan, 1985; Zwick, Donoghue, & Grima, 1993), item parameter values were selected from probability distributions with specified means and variances – e.g., a-parameters were selected from a uniform distribution. For each run the DIF item a-parameter values were generated from a uniform distribution ($M = 0$, $SD = 2$), b-parameter values were generated from normal distribution ($M = 0$, $SD = 0.75$), and c-parameter values were generated from uniform distribution ($M = 0$, $SD = 0.50$). Note that as in study one the θ values were generated from a normal distribution ($M = 0$, $SD = 1$) for each group; hence, like study one, the resultant proportion of rejected t-tests was the empirical Type I error rate.

Again, as in study one, given that this study focuses on uniform DIF, another factor manipulated in this study is the difference in *b*-parameter values between the focal and reference groups. The difference in b-parameters, Δb, were generated from a normal distribution ($M = 0$, $SD = 0.50$). With *b, c* and Δb, Raju's areas were calculated using equation (2) to quantify the magnitude of the uniform DIF,

$$Area = (1-c)\left|b_2 - b_1\right|. \qquad (2)$$

As a descriptive summary of our simulation data, the generated values of the b-parameter ranged from -2.691 to 2.066 (M=-0.106, SD=0.863). Likewise, the a-parameter values ranged from 0.006 to 1.993 (M= 0.987, SD= 0.594); and c-parameters ranged from 0.002 to 0.300 (M=0.161, SD=0.085). Furthermore, Raju's area ranged from 0.004 to 1.156 (M=0.327, SD=0.254), and the difference in b-parameters as an index of DIF (the delta-b) ranged from -1.068 to 1.300 (M=-0.031, SD=0.494). Finally, the a-, b-, and c-parameter values were not statistically significantly correlated with each other; ranging from -0.048 to 0.036.

The impact of varying item parameters on Type I error rate was then analyzed by statistical modeling using the resultant data from above simulation. The dependent variable for these analyses is the Type I error rate whereas the explanatory variables are: a-parameter, b-parameter, c-

parameter, Δb, sample size, and the magnitude of DIF calculated by Raju's formula in equation (2).

RESULTS AND CONCLUSIONS

Table 5 listed the minimum and maximum values, means, standard deviations of the resultant Type I error rates for different sample sizes. The minimum and maximum values, mean, and standard deviation of the Type I error rate for sample size of 25 and 50 per group are almost same. As the sample size increases to 125 per group and above, the maximum values of Type I error rate tend to inflated beyond the Bradley's moderate and liberal criteria. The means and the standard deviations, however, are same as those of small samples.

Table 5
Type I Error Rates for Different Sample Sizes

n (per group)	Minimum	Maximum	M	SD
25	0.044	0.057	0.050	0.003
50	0.044	0.058	0.051	0.003
125	0.042	0.065	0.051	0.004
250	0.044	0.070	0.052	0.004

Table 6 provided the percentage of Type I error rates, out of the 50 runs in that cell, for each sample size that meet Bradley's (1978) various criteria for acceptable Type I error rates. As an example to guide the reader in how to interpret Table 6, for a sample size of 25 per group the Type I error rate met the moderate criterion with all Type I error rates less than .060, and 47 of 50 meet stringent criterion with values less than .055. In general, from Table 6 it is clear that with increasing sample size the number of the Type I error rate values that meet the moderate criterion decreased – this is also true of the stringent criterion.

To investigate the effect of sample size on the Type I error rate in this study, one-way ANOVA was conducted with sample size as the independent variable. The effect of sample size was not significant,

$F(3,196) = 2.075$, $p = 0.105$. This result indicates that the sample size effect was trivial in the study situation. To investigate the association among the dependent and explanatory variables in this study, we conducted correlation analyses. Table 7 provided the Pearson correlation and Spearman's *rho* between Type I error rate and *a, b, c*, Δb, and Raju' area. The results indicate that the Type I error rate is only statistically significantly correlates with Raju's area.

Table 6

For Different Sample Sizes, the Percentage of Type I Error Rates that Meet Bradley's (1978) Criterion

	Sample size (per group)			
Criterion	25	50	125	250
Stringent	94	88	88	88
Moderate	100	100	98	94
Liberal	100	100	100	100

It should be noted that the finding in this study, Study three, that sample size was not significant is not the same as the finding in Study one above. The difference is most likely due to the fact that in Study three we only investigated one DIF item, whereas in Study one we investigated from one to 16 DIF items. Furthermore, focusing only on the case of one DIF item in Study one, Table 3, findings similar to Study three can be seen – a trivial effect of sample size, and only in the case of large magnitude of DIF and number of DIF items.

The above descriptive information indicated that in general (a) the magnitude of DIF (Raju' area) is the only factor that significantly correlated with Type I error, and (b) a-, b-, c-parameters are not significantly related to Type I error rate. It should be noted that because the Type I error rate is a proportion we investigated whether using logit transformation would change the conclusions. The transformation did not change the conclusions, so the analysis was reported using untransformed data.

Table 7

Pearson Correlations and Spearman's rho Between Type I Error Rate and a-, b-, c-parameter, Δb, and Raju's Area

	Type I error rate	
	Pearson correlation	Spearman's rho
a	0.07	0.02
b	-0.01	-0.05
c	-0.09	-0.09
Δb	0.03	-0.04
Raju's area	0.35**	0.29**

**P < 0.01.

Study three was conducted to investigate the effects of varying item parameter (*a, b,* and *c*) values, Δb, sample size and magnitude of DIF (quantified by Raju's area) on the Type I error rate in the one DIF item situation. The results indicated that in this study the values of item parameters are not related to the inflation of Type I error rate. The only influential factor is the magnitude of DIF (Raju's area). This result confirms what we found in study one: that the magnitude of DIF is a significant explanatory variable for increases in subsequent Type I error rates for the t-test based on the observed total test score, and confirm the generalizability of results in study one: the Type I error may be protected with one small DIF item retained in the test.

SECTION ONE DISCUSSION

It was found, as predicted, that DIF did have an effect on the statistical conclusions; both the Type I error rate and effect size index of the observed score differences were inflated. The effect size results are informative for the Type I error findings because they, in essence, show that when one has DIF that the observed score effect sizes are non-zero (when they should be zero in the Type I error situation). That is, the observed score effect sizes are inflated by the DIF. This highlights our earlier statement that our Type I error rates (and effect sizes) reflect the probability

of rejecting the null hypothesis of equal observed test score means when the latent means are equal but DIF is present in the test. The Type I error rate (and zero effect size) are defined relative to the latent variable in the IRT model, not the observed variable; hence the inflation in Type I error rate. In short, DIF creates mean differences in the observed scores when none exist on the latent variable.

However, remarkably and not predicted from research, DIF also had little to no effect in some conditions. That is, if one has one DIF item out of 38 items the Type I error rate of their subsequent hypothesis test is not substantially inflated above the nominal level. Furthermore, the subsequent effect size from the mean comparison is only inflated less than 0.03 on a standardized metric. In fact, this little effect of DIF also held up when there were 4 (out of 38) DIF items with small magnitude of DIF. Study three shows that the conclusions are not restricted to the specific item parameter values for the DIF item.

The first section of this paper only addresses the matter of amplification, unidirectional DIF. The next section moves to the question of what happens to the Type I error of subsequent hypothesis tests when DIF items show cancellation patterns.

SECTION TWO

Impact of Differential Item Functioning (DIF) on Statistical Conclusions, II: Cancellation Effects

The results in Section one were based on an amplification view of DIF – i.e., all the DIF items were in the same direction. Section two will build on section one's findings and focus on investigating potential cancellation effects. A cancellation effect (also called a multi-directional DIF) occurs when some DIF items favor one group and other DIF items favor the other group and the overall DIF effect cancels each other out. Of course, one can have partial cancellation wherein the overall DIF effect does not cancel out entirely but rather to some degree. For example, if one imagines gender DIF then one would have cancellation if the items favoring boys cancelled out the items against boys (i.e., favoring girls). Of course, as noted above, this gender DIF cancellation effect can be to only some degree, and hence only partial cancellation.

Building on the previous three studies, two computer simulation studies are reported below. Study four reports on the Type I error rates and study five on the effect sizes of subsequent statistical tests of mean differences when some degree of cancellation effects are present among

non-unidirectional DIF items. Therefore, the general simulation methodology is the same as the one used in studies one and two, respectively, for studies four and five.

The research question addressed in this section was concerned with identifying under what conditions the Type I error rate of subsequent hypothesis tests of equality of observed score means was inflated above the nominal level (i.e., 0.05), and under what conditions the effect size is biased.

Study Four: Impact of Cancellation DIF on Type I error rates

METHOD

Give that this is the first study of its kind, and part of a larger series of studies, we limited our simulation to some idealized situations in which the magnitude of DIF is the same for each of the items. Future studies will be able to build on this idealized experimental work to more generalized situations. We chose to focus on the case of 16 DIF items (out of a total of 38) because (a) research reported above shows that this number of DIF items substantially inflated the Type I error rate, and (b) the 16 items allowed us to investigate a large number of degrees of partial cancellation -- as compared to, for example, 4 DIF items which would only allow us to investigate a quarter, a half, or three quarters of the items favoring one group and the remaining items favoring the other group. Therefore, for example, in our simulation design for a small magnitude of DIF (Raju's area of 0.40), we simulated the situation in which 8 items were favoring one group (e.g., boys) and 8 items were favoring a second group, (e.g., girls). We denoted this situation as 8:8; which is the balanced DIF situation in which there is complete cancellation. We expect in this situation that the DIF effects will be balanced out. Next we simulated the same situation expect for the DIF items being distributed as 7:9, 6:10, 5:11, 4:12, 3:13, 2:14, 1:15, and 0:16 DIF items per group. For each of these nine simulation conditions the same simulation procedures were conducted to generate the item response data for this study as in study one. Continuing with the same data analysis strategies, the descriptive information was presented based on Bradley's (1998) criterion followed by regression analysis.

RESULTS AND CONCLUSIONS

Tables 8 to 10 list the Type I error rates and the robustness information based on Bradley's criterion for small, moderate and large magnitude of DIF, respectively. One can see from these tables that when one has complete cancellation (i.e., 8:8) the Type I error rate is, as expected, not inflated. One can also see that, as in study one, as the sample size and magnitude of DIF increase so does the Type I error rate. However, depending on the magnitude of DIF, the Type I error rate can be protected for some partial cancellation conditions. For example, one can see from Table 8 (small DIF, Raju's area of 0.40) that for a sample size of 50 per group the subsequent t-test of equal means has a protected Type I error rate in partial cancellation of 5 DIF items favoring one group and 11 items favoring the other group (i.e., a six item difference in the number of DIF items). Furthermore, for Tables 9 and 10 (i.e., moderate and large DIF), for a sample size of 25 per group the t-test is protected for as much as 6 items favoring one group and 10 items favoring the other (i.e., a 4 item difference in the number of DIF items).

Table 8

Type I Error Rates of t-test under Cancellation Effect among Items with Small DIF

	Number of DIF items against reference and focal groups (reference/focal)								
N	8:8	7:9	6:10	5:11	4:12	3:13	2:14	1:15	0:16
25/25	0.049	0.052	0.053	0.057	0.074↑	0.065↑	0.076↑↑	0.097↑↑	0.104↑↑
50/50	0.048	0.054	0.057	0.062↑	0.077↑↑	0.093↑↑	0.108↑↑	0.133↑↑	0.165↑↑
125/125	0.055	0.060↑	0.069↑	0.082↑↑	0.107↑↑	0.168↑↑	0.215↑↑	0.285↑↑	0.341↑↑
250/250	0.049	0.052	0.093↑↑	0.127↑↑	0.184↑↑	0.288↑↑	0.393↑↑	0.497↑↑	0.601↑↑

Note:

Type I error rate	Bradley (1978) criterion
$\alpha < 0.055$	Meet the stringent criterion
$0.055 \leq \alpha < 0.060$	Meet the moderate criterion
$0.060 \leq \alpha < 0.075$, ↑	Violate the moderate criterion but meet the liberal
$\alpha \geq 0.075$ ↑↑	Violate the liberal criterion, Inflated

Table 9

Type I Error Rates of t-test under Cancellation Effect among Items with Moderate DIF

	Number of DIF items against reference and focal groups (reference/focal)								
N	8:8	7:9	6:10	5:11	4:12	3:13	2:14	1:15	0:16
25/25	0.046	0.049	0.055	0.066↑	0.084↑↑	0.102↑↑	0.121↑↑	0.148↑↑	0.172↑↑
50/50	0.049	0.056	0.075↑↑	0.084↑↑	0.109↑↑	0.163↑↑	0.209↑↑	0.255↑↑	0.308↑↑
125/125	0.052	0.058	0.100↑↑	0.136↑↑	0.214↑↑	0.334↑↑	0.436↑↑	0.535↑↑	0.628↑↑
250/250	0.049	0.056	0.134↑↑	0.227↑↑	0.376↑↑	0.562↑↑	0.730↑	0.837↑↑	0.898↑↑

Note:

Type I error rate	Bradley (1978) criterion
$\alpha < 0.055$	Meet the stringent criterion
$0.055 \leq \alpha < 0.060$	Meet the moderate criterion
$0.060 \leq \alpha < 0.075$, ↑	Violate the moderate criterion but meet the liberal
$\alpha \geq 0.075$ ↑↑	Violate the liberal criterion, Inflated

Table 10

Type I Error Rates of t-test under Cancellation Effect among Items with Large DIF

	Number of DIF items against reference and focal groups (reference/focal)								
N	8:8	7:9	6:10	5:11	4:12	3:13	2:14	1:15	0:16
25/25	0.051	0.051	0.060↑	0.076↑↑	0.098↑↑	0.145↑↑	0.166↑↑	0.214↑↑	0.265↑↑
50/50	0.048	0.053	0.067↑	0.101↑↑	0.170↑↑	0.259↑↑	0.280↑↑	0.381↑↑	0.488↑↑
125/125	0.048	0.062↑	0.098↑↑	0.173↑↑	0.356↑↑	0.531↑↑	0.588↑↑	0.747↑↑	0.853↑↑
250/250	0.055	0.073↑	0.135↑↑	0.318↑↑	0.595↑↑	0.819↑↑	0.866↑↑	0.963↑↑	0.990↑↑

Note:

Type I error rate	Bradley (1978) criterion
$\alpha < 0.055$	Meet the stringent criterion
$0.055 \leq \alpha < 0.060$	Meet the moderate criterion
$0.060 \leq \alpha < 0.075$, ↑	Violate the moderate criterion but meet the liberal
$\alpha \geq 0.075$ ↑↑	Violate the liberal criterion, Inflated

To investigate which experimental factors influence the Type I error rate a regression analysis was conducted with magnitude of DIF, sample size, and difference in the number of DIF items between the two groups as independent variables. The resultant model with two-way and three-way interactions was statistical significant, $F(7, 100) = 274.1$, $p < 0.0001$, R-squared = 0.950. All the main effects and the three-way interaction were statistically significant. One can see from a careful review of the tables that the relationship between the difference in the number of DIF items (i.e., a proxy for the degree of partial cancellation) is more pronounced (i.e., a higher correlation) for larger magnitudes of DIF, and this relationship increases with increasing sample size.

Clearly then the Type I error rate depends not only on the degree of partial cancellation but also on magnitude of DIF and sample size, and that in some cases the Type I error rate is protected even when one has partial cancellation.

Study Five: Impact of Cancellation DIF on Effect Size

Study five was designed to investigate the cancellation effect of DIF on the population effect size of the independent sample t-test. As in study two, Cohen's d was used as the measure of effect size and, the effect size difference, ΔES, was computed as the difference between the observed mean effect size and latent mean effect size so that a positive difference means that the effect size was larger for the observed scores. Our research question in this study is: What is the effect size for the observed test score means when the latent means are equal but DIF cancellation effect is present in the test?

METHODS

The simulation factors manipulated in this study, as well as the simulation methodology, were the same as those in study two except for the experimental factor, number of DIF items. That is, within each magnitude (small, moderate and large) of DIF, we manipulated number of DIF items (out 16 DIF items) against focal group and reference groups. As in study four, the simulated number (out 16 DIF items) of DIF items against reference and focal groups were as follows: 8:8, 7:9, 6:10, 5:11, 4:12, 3:13, 2:14, 1:15, and 0:16. It should be noted that the 0:16 condition does, of course, not reflect cancellation but was included for comparison purposes. Three levels of magnitude of DIF -- small, moderate, and large as defined by Raju's area statistic of .4, .6, and .8 (Raju, 1988) were investigated. This resulted in a 9x3 completely crossed factorial design resulting in a total of 27 cells in our simulation design (including completely balanced, 8:8, and

completely unbalance cases, 0:16). As in study two, we simulated 10,000 examinees in each cell of the simulation design for our pseudo-populations. For each cell we computed the effect size for the observed total test score mean difference and for the latent mean difference (and their corresponding standard deviations). Because both the observed score and latent variable effect size values are on the same metric (both being standardized) we computed the difference between them (ΔES) as an index of how much the DIF biases the effect size.

RESULTS AND CONCLUSIONS

Table 11 lists differences between the effect sizes of the observed and the latent variable score for the three magnitudes of DIF and the 9 number of DIF situations. One can see that when the number of DIF items present in each group are totally balanced (8:8), the effect sizes of the latent mean and observed mean are, as expected, almost equal – i.e., -0.008. However, as the number of DIF items against each group is not balanced, the ΔES increase; the more unbalanced, and the larger the magnitude of DIF, the greater the ΔES – i.e., the observed mean differences being spuriously inflated by the presence of DIF.

The simulation results were analyzed using regression analysis with ΔES as the dependent variable and magnitude of DIF, number of DIF item difference between groups, and their two-way interactions as independent variables. The model is statistically significant ($F(3, 23) = 1621.08$, $p < .0001$) with an R-squared of 0.995. All the predictors are statistically significant, including the interaction term. The interactions among independent variables can be seen upon careful review of Table 11. Clearly, ΔES increases as the imbalance in DIF and magnitude of DIF increase.

SECTION TWO DISCUSSION

The results confirm the hypothesis that when there is a balanced number of DIF items between groups and when magnitude of DIF is close to zero, the Type I error rate is protected and ΔES was not biased no matter how large the magnitude of DIF and number of DIF items present. On the other hand, as the number of DIF items become more unbalanced between groups both the Type I error rate and the ΔES were inflated. Furthermore, the effect of imbalance was even more inflated by magnitude of DIF.

Table 11

Differences Between Observed and Latent Mean Effect Sizes for Varying Number of DIF Items in Reference and Focal Group for Different Magnitudes of DIF

Number of DIF items in each group	ES difference (ΔES)		
	Magnitude of DIF		
Reference vs. focal	Small	Moderate	Large
8 : 8	-0.008	-0.010	-0.009
7 : 9	0.023	0.032	0.035
6 : 10	0.052	0.077	0.082
5 : 11	0.083	0.109	0.147
4 : 12	0.100	0.158	0.198
3 : 13	0.124	0.193	0.266
2 : 14	0.149	0.225	0.315
1 : 15	0.174	0.270	0.354
0 : 16	0.196	0.297	0.396

GENERAL DISCUSSION

As we noted in the introduction, it is not uncommon for researchers to either not test for DIF before comparing groups, or if they test for DIF but they decide to leave DIF items in the test. Of course, DIF is a statistical characteristic of a sample so it is possible that DIF items are simply not detected during item analysis. In short, this leaves us with the question of the impact of DIF items on the eventual statistical tests conducted on the observed test (or scale) scores. To answer the above general questions we conducted five related simulation studies. To our knowledge, this is the first of a line of research that directly answers the often heard question: What is the impact of having DIF on the eventual statistical conclusions from my test scores? It offers advice to practicing researchers about when and how much the presence of DIF will effect their statistical conclusions based on the total observed test scores. Although, simulated in idealized

situations deliberately, the five related simulation studies provide researchers and practitioners with general guidelines.

In the case of Section I wherein the DIF items are all in one direction (e.g., the test favors girls consistently, amplification DIF), as expected, DIF results in inflation in Type I error rates of the eventual t-test. Likewise, of course, reflecting the inflation in Type I error rates, the observed score effect size is also inflated, sometimes substantially. The inflation of the effect size is important because it is now widely recommended that effect sizes be reported with statistical hypothesis test results. What was not expected, however, was that the Type I error rate and effect sizes were not biased by the presence of DIF when the number of DIF items is small (i.e., 1 DIF item out of 38 items, and even 4 DIF items out of a total of 38 items when the magnitude of DIF is small to moderate. This is important, and comforting, to researchers who do not typically screen for DIF or ones who do not remove DIF items from the test. However, what is not yet known is the impact of DIF, in these situations when the Type I error rate is protected, on the eventual statistical power. The issue of impact of DIF on statistical power will be investigated in forthcoming studies. Likewise, our studies should not be interpreted to suggest that one need not screen items for DIF. In fact, our conclusions are quite to the contrary because DIF analyses are needed because under many situations the Type I error rate and effect sizes are severely biased by the presence of DIF.

In Section II wherein one has an imbalance of DIF items, for example, some items favoring girls and others favoring boys, the effect of DIF depends on the degree of imbalance. As expected, when the DIF is balanced (e.g., 8 items favoring boys and 8 items favoring girls) the DIF effect cancels out and the Type I error rate and effect sizes are not biased by DIF. However the degree of imbalance and the magnitude of DIF interact to inflate both the Type I error rate and effect size. Again, the t-test was surprisingly robust in terms of Type I error rate and effect size with a small amount of imbalance (e.g., the t-test was not greatly effected when 6 items favored one group and 10 items the other).

Overall, these findings highlight why it is important to use DIF screening procedures before conducting group comparisons because one may find themselves in the situation wherein the Type I error rate of their hypothesis test, and the corresponding effect size reported, are highly inflated declaring group differences where none exist. Likewise, retaining DIF items in the test may also have significant effect on other psychometrical procedures, such as equating results when used in concert with DIF detection or more broadly in the use of linking and equating. That

is, several studies have investigated the effects of linking or equating methods on DIF detection (e.g., Candell & Drasgow, 1988; Cohen & Kim, 1992; Hidalgo-Montesinos & Lopez-Pina, 2002; Miller & Oshima, 1992); however, there is a need for more research on the effect of DIF on equating or linking (e.g., Chu & Kamata, 2005) in its more general use in large-scale testing much like we do for significance testing herein.

REFERENCES

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 34*, 144-152.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397 - 424). Reading, MA: Addison-Wesley.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221- 256). Westport, CT: Praeger.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12,* 253-260.

Chu, K.L., & Kamata, A. (2005). Test equating with the presence of DIF. *Journal of Applied Measurement, 6,* 342-354.

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

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19-29.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67*, 373-393.

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

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. Applied Psychological Measurement, 10, 287-302.

Hidalgo-Montesinos, M.D., & Lopez-Pina, J.A. (2002). Two-stage equating in differential item functioning detection under the graded response model with the Raju area measures and the Lord statistic. *Educational and Psychological Measurement, 62*, 32-44.

Kim, S.H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*, 51-66.

Maller, S. J. (2001). Differential item functioning in the WISC-III: Item parameters for boys and girls in the national standard sample. *Educational and Psychological Measurement, 61*, 793-817.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement, 16,* 381-388.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*, 257-274.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116.

Roznowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: Implications for measurement bias. *Journal of Applied Psychology, 72,* 480-483.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59*, 248-269.

Rupp, A. A., & Zumbo, B. D. (2003). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *Alberta Journal of Educational Research, 49,* 264-276.

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

Shealy, R., & Stout, W. (1991). *An item response theory model for test bias* (Office of Naval Research Tech. Rep. No. 4421-548). Champaign-Urbana: University of Illinois, Department of Statistics.

Shealy, R., & Stout, W. (1993). An item response theory model for test and differential item functioning. In H. Wainer & P. Holland (Eds.), *Differential item functioning* (pp. 197-240). Hillsdale, NJ: Lawrence Erlbaum.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*, 77-87.

Zumbo, B. D., & Harwell, M. R. (1999). *The methodology of methodological research: Analyzing the results of simulation experiments.* (Paper No. ESQBS-99-2). Prince George, B. C.: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? *Implications for translating language tests. Language Testing, 20,* 136-147.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao and S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 45-79). Elsevier Science B.V.: The Netherlands.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233 – 251.

(Manuscript received: 26 August 2008; accepted: 22 October 2008)