Psychological Test and Assessment Modeling, Volume 56, 2014 (1), 25-44

The impact of group pseudo-guessing parameter differences on the detection of uniform and nonuniform DIF

W. Holmes Finch[1] *& Brian F. French*

Abstract

Differential item functioning (DIF) is an important aspect of item development and validity assessment. Traditionally DIF is divided into two broad types, focusing on conditional group differences of the item difficulty (uniform DIF) and discrimination (nonuniform DIF) parameters. Relatively little attention has been given to group differences on the probability of answering an item correctly due to chance. The goal of this study was to investigate the influence of such group differences on the detection of uniform and nonuniform DIF, and on the accuracy of the estimation of item difficulty and discrimination parameters. Results demonstrate that when groups differed on the pseudo-guessing parameter in a 3 parameter item response theory model, Type I error rates for both uniform and nonuniform DIF were elevated, and that these differences appear to be due to parameter estimation bias for both item difficulty and discrimination. Implications of these results are discussed.

Key words: Differential Item Functioning, Guessing, Type I Error, Validity, Item Response Theory, 3 parameter logistic model

[1] *Correspondence concerning this article should be addressed to:* Holmes Finch, PhD, Department of Educational Psychology, TC 521 Ball State University, Muncie, IN 47306, USA; Brian F. French, PhD, Department of Educational Leadership and Counseling Psychology, Cleveland Hall, Washington State University, Pullman, Washington, 99164, USA; email: frenchb@wsu.edu

Differential item functioning (DIF) continues to receive attention in the measurement literature both in applied and methodological studies. The latter has begun to focus more on the accuracy of detection and explanation of why DIF is occurring, rather than the creation of new methods for detection. This focus is justified given the number of methods available. Moreover, the importance of locating differentially functioning items revolves around the issue of valid score inferences. The importance of this issue cannot be overstated given the dependence on test scores to drive decisions at all levels of the nation's educational system, including school accountability for student growth (Dunn & Allen, 2009). Thus, accurate detection and explanation of why DIF might be occurring is paramount so that items and then test scores can accurately reflect examinee ability.

Typically DIF is discussed in two broad classes: (a) uniform DIF where, conditioning on the measured trait, the probability of a correct item response, given dichotomous items, is uniformly higher for members of one group (e.g. reference) compared to members of another (e.g. focal), and (b) nonuniform DIF where the probability of a correct item response differs between groups across all levels of ability, but the magnitude of this difference is not consistent; i.e. there is an interaction between ability level and group membership (Camilli & Shepard, 1994). An alternative way to view DIF for dichotomous items, as presented by Lord (1980) can be through the 3-parameter logistic model (3-PL), which takes the form:

$$P\left(U_i = 1 \mid \theta\right) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}} \tag{1}$$

where

θ = examinee ability

c_i = pseudo-guessing parameter for item i

a_i = discrimination parameter for item i

b_i = difficulty parameter for item i

In this context, uniform DIF represents a difference in item difficulty (*b*) between the two groups, while nonuniform DIF is a group difference in item discrimination (*a*). Other researchers have defined DIF in terms of differences in the Item Response Functions (IRF's) (see Camilli & Shepard, 1994). Specifically, uniform DIF has been described as the case when group specific IRF's for an item differ across values of the latent trait, and this difference is consistent. In contrast, non-uniform DIF can be thought of as occurring when the IRF's for the groups differ from one another, but this difference is not consistent across all values of the latent trait. A special case of non-uniform DIF, commonly referred to as crossing DIF, occurs when the IRF's for the groups actually cross one another.

DIF in the pseudo-guessing parameter

A form of DIF that has not received much attention, and which is the focus of this study, is that which occurs when the probability of a correct item response due to chance differs

between the two groups. That is, there is a group difference on the pseudo-guessing or *c* parameter in the 3PL item response theory model seen in Equation 1. Consider, for example, a multiple choice mathematics item with 5 possible responses (4 distracters and the correct response). Next, consider that this item is worded so that members of the reference group (but not the focal group) could rule out one of the distracters due to information not based on the measured trait, mathematics. Members of the focal group who did not know the correct answer would therefore have a 20% chance of obtaining a correct answer through random guessing, whereas members of the reference group would have a 25% chance of a correct random guess because they were able to rule out one of the distracters. We do acknowledge that the probabilities may not be exact as some examinees may be able to rule out a different number of choices, hence pseudo-guessing.

Such a differential guessing situation might be seen as exemplary of the influence of a secondary nuisance dimension. In this case the secondary dimension would not have a direct impact on item difficulty or discrimination; e.g. the item would remain equally difficult in terms of mathematics for both groups. However, the nuisance dimension would impact the probability of a correct response due to guessing by providing examinees with clues regarding which distracter(s) could be ignored. This *c*-DIF then denotes group differences on the 3-PL *c* parameter. While here we refer to the difference in group *c* values as a unique type of DIF, it is also possible to view it as a variation of nonuniform DIF, in that the IRFs for the groups are different from one another, but this difference is not consistent across levels of the latent trait being measured. However, we will refer to *c*-DIF in this manuscript in order to differentiate between this condition, and other more traditional types of nonuniform DIF described in the literature.

Methods for detecting DIF

A variety of methods are available to test for DIF, including the Mantel-Haenszel (MH) statistic, logistic regression (LR), SIBTEST, crossing SIBTEST, and the Item Response Theory Likelihood Ratio (IRTLR) among others (Camilli & Shepard, 1994); Camilli, 2006; Osterlind & Everson, 2009; Penfield & Camilli, 2007). The literature is bountiful with studies applying these techniques for DIF detection as well as studies examining the accuracy of uniform and nonuniform DIF detection across methods and data conditions. The two methods used for DIF detection in this study, IRTLR and LR, were selected because they can assess an item for both uniform and non-uniform DIF. In addition, IRTLR can be used to test for the presence of *c*-DIF. Following is a brief description of each method.

Likelihood Ratio statistic

The IRTLR test statistic (Thissen, 2001; Thissen, Steinberg &, Wainer, 1988; Thissen, Steinberg &, Wainer, 1993) is calculated by fitting a series of nested models to the data using an appropriate form (e.g. 2PL, 3PL). The initial model constrains all item parame-

ters to be equal between the reference and focal groups, yielding a log-likelihood (LL_{equal}) statistic:

$$LL_{equal} = \sum_{G=1}^{2}\sum_{p=1}^{N} \ln\left[\sum_{1}^{q}\prod_{i=1}^{nitems}\left(T_{iG}\left(u_{ipG}\right)\varphi_G\left(\theta\right)d\theta\right)\right] \tag{2}$$

where

$T_{iG}\left(u_{ipG}\right)$ = Parameters for ICC for group G; constrained equal for both groups

$\varphi_G\left(\theta\right)$ = Distribution of the latent trait for group G.

The second IRT model fit to the data again constrains item parameters equal between the two groups, except for those associated with the target item for which DIF is being assessed. Again, a log-likelihood value ($LL_{unequal}$) is calculated for this second model. The difference between these two log-likelihoods then serves as the test statistic, $G^2 = -2\left(LL_{equal}\right) - \left(-2LL_{unequal}\right)$ for DIF. It is distributed as a chi-square with 3 degrees of freedom, and tests for overall DIF across all of the item parameters (e.g., *a, b* and *c* in the 3PL context). If this test is statistically significant, then individual tests are conducted for each item parameter by unconstraining individual parameters and comparing the resulting log-likelihood values in a manner analogous to that described above, in order to identify the nature of the DIF.

Logistic regression

LR takes the form of a model linking a categorical outcome with one or more predictor variables, which can be either continuous or categorical. As outlined by Swaminathan and Rogers (1990) the logistic regression model for DIF detection is:

$$p(u_i = 1 \mid \theta, g) = \frac{e^{\beta_0 + \beta_j\theta + \beta_j g + \beta_j(\theta g)}}{1 + e^{\beta_0 + \beta_j\theta + \beta_j g + \beta_j(\theta g)}} \tag{3}$$

where

p_i = Probability of correctly responding to item i

β_0 = Intercept

β_j = Slope for model term j

θ = Ability level, typically the total score

g = Group membership.

A significant interaction between the ability estimate and group membership would indicate the presence of non-uniform, whereas a significant group term would indicate uniform DIF. As noted above, one advantage of LR is that the method can assess both uni-

form and nonuniform DIF (Swaminathan & Rogers, 1990), and has been found reasonably effective in several previous studies, as will be discussed in detail below.

Prior research on c-DIF

Little work has examined the influence of *c*-DIF on 3-PL item parameter estimation or uniform and specifically non-uniform DIF detection. However, results that are available do show that the presence of *c*-DIF may cause problems for DIF detection methods and influence the probability of a correct response (Bao, Dayton &, Hendrickson, 2009; Walstand & Robson, 1997). Additionally, *c*-DIF also may be more difficult to detect compared to uniform and non-uniform DIF (Schumacker, 2005). That said, Thissen, Steinberg and Wainer (1988) demonstrated how IRTLR could be used successfully to test for items containing *c*-DIF. In particular, to test for *c*-DIF, two models are fit to the item response data. In the first model, the pseudo-guessing parameter for the target item is held equal between the groups, while in the second model the pseudo-guessing parameter values for the target item are allowed to vary by group. The test statistic for the null hypothesis of *c*-DIF is then calculated as the difference between the -2 log-likelihood values of the constrained and unconstrained models, as described above in general.

DeMars and Wise (2007) argue that differential rapid-guessing among groups on a timed test could lead to uniform DIF. That is, what might be characterized as uniform DIF could actually appear because groups have differential guessing rates. Examinees tend to exhibit this type of behavior in high-stakes testing environments (Schnipke & Scrams, 1997) when time to respond is ending and in low-stakes untimed testing environments (Wise & Kong, 2005). Such guessing behavior appears to be unrelated to ability levels (Wise & Kong). However, and possibly of serious consequence to the examinee, this type of differential guessing behavior, including rapid responding, can lead to lower scores for rapid responders as compared to examinees not exhibiting such behavior (Osborne & Blanchard, 2011).

More recently, Demars & Wise (2010), through a series of studies, demonstrated that there are situations where DIF may be detected due to guessing behavior and not due to content of the items. That is, DIF may manifest itself as a direct result of an examinee behavior (i.e., rapid response) and not an examinee characteristic (e.g, sex, race). These authors also demonstrated that the presence of groups displaying more rapid guessing and lower ability resulted in an overall lower proportion of items being flagged for DIF. However, the data conditions only focused on uniform DIF.

In addition to guessing behavior influencing DIF detection Type I error rates, Finch and French (2007) found that the presence of one type of DIF (e.g. uniform) elevated the probability of incorrectly identifying the other type of DIF (e.g. non-uniform). Given this result, one could assume that similar inaccuracy in uniform/non-uniform DIF detection may be occurring as a result of guessing and rapid response behavior in general, although very little direct research has been conducted in this area. The current study addresses this problem from a different angle compared to previous research by examining not only the Type I error rates for both uniform and nonuniform DIF in the presence of differen-

tial guessing rates between groups (*c*-DIF) but also investigating the bias in item parameter estimates in order to better understand the impact of group differences in guessing on model estimation.

Goals of the current study

Given the prior results discussed above, it is hypothesized that the presence of *c*-DIF will impact the probability of identifying either uniform or non-uniform DIF for an item, even when neither type is actually present in the population. In practice, such a result could lead to unnecessary item reviews and/or incorrect item revisions. Beyond the few studies reviewed above, there has been little published research systematically examining the influence of *c*-DIF on parameter estimates and DIF detection or even how accurate *c*-DIF can be detected. Therefore, the goals of this study were to examine the influence of *c*-DIF on the estimation of 3-PL item parameters, and on the accuracy of the detection of uniform and non-uniform DIF by the two common DIF methods which allow for simultaneous testing of uniform and non-uniform DIF, logistic regression (LR) and the Item Response Theory Likelihood ratio test (IRTLR). These two methods were selected for inclusion in the study because they have repeatedly been shown to be effective tools for uniform and non-uniform DIF detection (Finch & French, 2007; Thissen, Steinberg, &, Wainer, 1993; Rogers & Swaminathan, 1993). In addition, the performance of IRTLR in identifying *c*-DIF, the only major DIF detection method designed to test for group differences in *c*, was assessed, which has not been done extensively in prior work.

Method

A simulation study was conducted to investigate the influence of *c*-DIF on uniform and non-uniform DIF detection, and on the accuracy of the estimation of item parameter values. A total of 1000 replications were generated and analyzed for each combination of simulation conditions, which are described below. This number of replications has been shown to be more than sufficient to ensure accurate estimation of model parameters, as well as Type I error and power rates (e.g., Bandalos & Leite, 2013; Cohen, Kane, & Kim, 2001; Harwell, Stone, Hsu, & Kirisci, 1996). Dichotomous data (20 items) were simulated with IRT-LAB (Penfield, 2003) using the 3-PL with item parameter values (Table 1) drawn from a national licensing exam calibration sample. The first item was selected as the target for which *c*-DIF was simulated, as is described below. Outcome variables of interest in the simulation study included bias (calculated as the mean absolute deviation between the population and model estimated values) in the estimates of *a, b* and *c* for the target item for the reference and focal groups, as well as the mean standard errors for these estimates. Item parameter estimation and accompanying standard errors were obtained using BIOLOG-MG, v. 3 (Zimowski, Muraki, Mislevy &, Bock, 2003). Estimation was conducted separately for each group so that we could isolate the impact of various *c* parameter values on estimation of the other item parameters. This

Table 1:
Item Parameter Values For Generating Simulated Data

Item	Discrimination	Difficulty	Pseudo-guessing
1	0.8	-0.25	0.20
2	1.51	1.10	0.19
3	0.81	0.17	0.27
4	0.46	1.52	0.26
5	0.67	1.49	0.26
6	0.35	-0.04	0.38
7	0.52	-1.77	0.27
8	0.58	0.83	0.28
9	1.17	1.09	0.28
10	0.62	0.15	0.22
11	0.49	1.06	0.23
12	1.04	0.36	0.14
13	0.41	-0.34	0.12
14	0.76	0.13	0.14
15	0.99	-0.40	0.12
16	0.62	-1.52	0.32
17	0.39	-0.35	0.15
18	0.41	0.71	0.29
19	0.47	0.36	0.26
20	0.61	0.52	0.33

was particularly important in attempting to explain the cause of DIF results that were obtained. In addition, the rejection of the null hypothesis of no group difference for *a (*non-uniform DIF) and *b* (uniform DIF) detection using both LR and IRTLR and the IRTLR rejection rate for the *c* parameter of the target item also served as outcomes of interest. A finding of significant group differences for *a* and *b* represents a Type I error in this study, because these parameters were simulated to be equal between the groups.

The following factors were manipulated to evaluate their influence on DIF detection.

Sample size/sample size ratio

Focal and reference group sample sizes were simulated as 250/250, 250/500, 500/500, 500/1000, and 1000/1000. These were selected to study the impact of *c*-DIF with a relatively small (500) to a fairly large overall sample (2000). These values are also in keep-

ing with prior published research in both the applied and simulation literature (e.g., Naryanan & Swaminathan, 1994, 1996; Rogers & Swaminathan, 1993). The smallest sample size conditions (250/250 and 250/500) are very small for use with a 3PL model. These were included to provide information regarding the impact of *c*-DIF at a lower bound sample size value. However, in practice using this model with such small sample sizes is probably not appropriate with IRT methods. Although, this size is commonly used with other methods (e.g., LR, MH, SIBTEST).

Group ability differences (impact)

Prior research has demonstrated that group mean differences in the latent trait are associated with Type I error inflation for uniform DIF (Clauser et al., 1993; Cohen & Kim, 1993; Roussos & Stout, 1996). Thus, in order to ascertain the impact of such differences in the case of *c*-DIF, the mean abilities of the focal and reference groups were set either as 0/0 (equal) or -0.5/0 (unequal).

c-DIF

c-DIF was simulated at 4 levels for the target item. For the reference group, *c* was held constant at 0.2, while for the focal group the value of *c* was set to 0.2 (no *c*-DIF), 0.15, 0.1 or 0.05. The latter three conditions represent increasing levels of differences between the *c*- parameter between groups. While we recognize that many patterns in *c*-DIF are possible, those included in this study were selected in order to clearly demonstrate what impact, if any, group differences in the pseudo-guessing parameter have on the testing of uniform and non-uniform DIF. That is, the magnitude of the difference was the critical manipulation and not the direction. Clearly, future research may focus on other patterns of *c*-DIF.

DIF detection methods

Uniform and non-uniform DIF were tested using both LR and IRTLR as mentioned previously. These methods were selected because they have been shown to work well under many conditions for detecting both uniform and nonuniform DIF in one analysis (e.g., Finch & French, 2007). In addition, IRTLR is the only commonly used method available to test for *c*-DIF.

Results

Rejection rates

Table 2 contains Type I error rates for uniform and non-uniform DIF for LR and IRTLR by the level of *c*-DIF, sample size, and group latent variable means, along with rejection rates for IRTLR comparing *c* between groups. Perhaps the most obvious pattern in these results is the inflation of Type I error rates for uniform and non-uniform DIF for both LR and IRTLR as the presence of *c*-DIF increased. The error rates, for example, were close to the nominal level for LR when there was no *c*-DIF (0.2/0.2 condition). However, these error rates rose concomitantly with an increase in the magnitude of *c*-DIF. This error inflation for LR was mild when the groups' *c* parameters differed by 0.05 (0.2/0.15 condition), but became quite severe in the 0.2/0.1 and 0.2/0.05 cases. A similar pattern of higher Type I error rates for greater *c*-DIF was also present for IRTLR, although this effect was impacted to a greater degree by the combination of sample size and group mean differences on the latent trait. When there was no *c*-DIF, IRTLR demonstrated somewhat inflated uniform and non-uniform Type I error rates for the smallest sample size conditions, particularly when ability differences were present. As sample size increased, these Type I error rates declined. This pattern was also evident when *c*-DIF was present, although the Type I error rates were generally inflated even when no ability differences were present. As with LR, the error rates increased concomitantly with the degree of *c*-DIF, although they tended to be lower compared to those for LR under most conditions, except for the two smallest sample size conditions in the presence of ability differences. There were not large differences in error inflation for uniform and nonuniform DIF for either LR or IRTLR.

Table 2:
Uniform And Non-uniform Rejection Rates* For Logistic Regression And The IRTLR Test By Level Of c-DIF

c-DIF **Ref/Foc**	**n**	**Impact**	**LR-[1]**	**LR[2]**	**IRTLR[1]**	**IRTLR[2]**	**IRTLR *c***
0.2/0.2	250/250	0/0	0.07	0.05	0.10	0.10	0.03
		0/-.5	0.02	0.05	0.15	0.11	0.04
	250/500	0/0	0.05	0.05	0.10	0.09	0.03
		0/-.5	0.06	0.06	0.16	0.09	0.01
	500/500	0/0	0.06	0.06	0.03	0.06	0.02
		0/-.5	0.06	0.08	0.07	0.08	0.02
	500/1000	0/0	0.03	0.02	0.04	0.05	0.04
		0/-.5	0.04	0.07	0.08	0.08	0.04
	1000/1000	0/0	0.04	0.04	0.05	0.06	0.04
		0/-.5	0.06	0.09	0.07	0.07	0.05

⇒

c-DIF Ref/Foc	n	Impact	LR-[1]	LR[2]	IRTLR[1]	IRTLR[2]	IRTLR *c*
0.2/0.15	250/250	0/0	0.06	0.07	0.12	0.11	0.05
		0/-.5	0.06	0.07	0.18	0.12	0.05
	250/500	0/0	0.06	0.05	0.07	0.07	0.04
		0/-.5	0.06	0.05	0.17	0.12	0.05
	500/500	0/0	0.07	0.05	0.09	0.07	0.05
		0/-.5	0.15	0.11	0.11	0.08	0.05
	500/1000	0/0	0.09	0.06	0.09	0.06	0.06
		0/-.5	0.12	0.10	0.14	0.10	0.05
	1000/1000	0/0	0.08	0.07	0.08	0.09	0.07
		0/-.5	0.11	0.15	0.15	0.12	0.07
0.2/0.10	250/250	0/0	0.08	0.06	0.11	0.10	0.06
		0/-.5	0.10	0.07	0.17	0.15	0.07
	250/500	0/0	0.12	0.13	0.09	0.10	0.07
		0/-.5	0.13	0.13	0.19	0.20	0.07
	500/500	0/0	0.15	0.15	0.08	0.12	0.08
		0/-.5	0.23	0.22	0.21	0.20	0.08
	500/1000	0/0	0.20	0.21	0.13	0.13	0.10
		0/-.5	0.31	0.28	0.23	0.19	0.09
	1000/1000	0/0	0.25	0.22	0.20	0.18	0.11
		0/-.5	0.27	0.27	0.30	0.29	0.12
0.2/0.05	250/250	0/0	0.18	0.20	0.14	0.13	0.08
		0/-.5	0.30	0.31	0.18	0.19	0.09
	250/500	0/0	0.23	0.21	0.15	0.16	0.10
		0/-.5	0.30	0.29	0.18	0.21	0.10
	500/500	0/0	0.28	0.24	0.21	0.19	0.12
		0/-.5	0.33	0.36	0.31	0.33	0.13
	500/1000	0/0	0.41	0.38	0.32	0.34	0.13
		0/-.5	0.51	0.52	0.44	0.44	0.15
	1000/1000	0/0	0.58	0.59	0.43	0.37	0.15
		0/-.5	0.58	0.56	0.54	0.49	0.16

[1] = Uniform, [2] = Non-uniform

*For uniform DIF, Non-uniform DIF and Pseudo-guessing DIF in the 0.2/0.2 condition, a significant result represents a Type I error.

With regard to the IRTLR test for *c*-DIF, the Type I error rate was always at or below the nominal 0.05 level. Furthermore, this test had very low power for detecting *c*-DIF, even presenting the presence of the biggest difference (0.2/0.05). Indeed, the highest level of power was 0.16 for a total sample size of 2000 when the groups' means on the latent trait were different. Generally speaking, power for detecting the presence of *c*-DIF was lower than the Type I error rates for both uniform and nonuniform DIF, across the conditions simulated.

Parameter estimation bias

Table 3 contains the item parameter bias results by the level of *c*-DIF, sample size and ability differences, based on estimates obtained using Bilog-MG (Zimowski, Muraki, Mislevy, &, Bock, 1996). In order to obtain these values, the model for the two groups was estimated without constraints for the target item. In the population, the item discrimination value was simulated to be 0.8. The estimate of *a* for the reference group was consistently higher than this, by a magnitude of between 0.02 to 0.10, with somewhat

Table 3:
Parameter Estimation Bias (Mean Standard Error) By Level Of c-DIF, Sample Size And Impact

c-DIF **Ref/Foc**	**n**	**Impact**	**Ref *a***	**Focal *a***	**Ref *b***	**Focal *b***	**Ref *c***	**Focal *c***
0.2/0.2	250/250	0/0	0.06 (0.20)	0.03 (0.19)	0.10 (0.24)	0.10 (0.24)	0.01 (0.08)	0.01 (0.08)
		0/-.5	0.08 (0.24)	0.07 (0.20)	0.10 (0.23)	0.58 (0.23)	0.01 (0.07)	0.01 (0.08)
	250/500	0/0	0.07 (0.16)	0.05 (0.20)	0.08 (0.21)	0.09 (0.24)	0.01 (0.08)	0.01 (0.08)
		0/-.5	0.09 (0.16)	0.05 (0.21)	0.08 (0.21)	0.59 (0.24)	0.01 (0.08)	0.01 (0.08)
	500/500	0/0	0.04 (0.16)	0.03 (0.16)	0.06 (0.20)	0.08 (0.21)	0.0 (0.08)	0.01 (0.07)
		0/-.5	0.07 (0.19)	0.05 (0.16)	0.07 (0.19)	0.59 (0.21)	0.01 (0.07)	0.01 (0.08)
	500/1000	0/0	0.06 (0.13)	0.04 (0.17)	0.08 (0.18)	0.09 (0.20)	0.01 (0.08)	0.01 (0.07)
		0/-.5	0.07 (0.13)	0.04 (0.19)	0.08 (0.18)	0.59 (0.19)	0.01 (0.07)	0.01 (0.07)

⇒

c-DIF Ref/Foc	n	Impact	Ref *a*	Focal *a*	Ref *b*	Focal *b*	Ref *c*	Focal *c*
	1000/1000	0/0	0.02 (0.13)	0.04 (0.13)	0.10 (0.18)	0.07 (0.19)	0.01 (0.07)	0.01 (0.07)
		0/-.5	0.05 (0.14)	0.04 (0.16)	0.09 (0.18)	0.58 (0.18)	0.01 (0.06)	0.01
0.2/0.15	250/250	0/0	0.07	0.11	0.08	0.20	0.0	0.05
		0/-.5	0.05	0.12	0.06	0.67	0.01	0.05
	250/500	0/0	0.07	0.09	0.09	0.17	0.01	0.05
		0/-.5	0.07	0.10	0.09	0.65	0.01	0.04
	500/500	0/0	0.05	0.09	0.06	0.16	0.01	0.05
		0/-.5	0.07	0.09	0.06	0.59	0.01	0.05
	500/1000	0/0	0.05	0.07	0.08	0.15	0.01	0.03
		0/-.5	0.05	0.09	0.08	0.63	0.01	0.04
	1000/1000	0/0	0.02	0.08	0.07	0.17	0.01	0.03
		0/-.5	0.04	0.08	0.08	0.60	0.01	0.04
0.2/0.10	250/250	0/0	0.07	0.19	0.08	0.20	0.01	0.08
		0/-.5	0.07	0.11	0.06	0.67	0.01	0.09
	250/500	0/0	0.07	0.13	0.09	0.17	0.01	0.08
		0/-.5	0.07	0.16	0.09	0.65	0.01	0.07
	500/500	0/0	0.05	0.13	0.09	0.16	0.01	0.08
		0/-.5	0.07	0.13	0.09	0.66	0.01	0.08
	500/1000	0/0	0.04	0.11	0.08	0.15	0.01	0.07
		0/-.5	0.05	0.12	0.08	0.63	0.01	0.05
	1000/1000	0/0	0.03	0.10	0.07	0.17	0.01	0.07
		0/-.5	0.06	0.12	0.08	0.60	0.02	0.06
0.2/0.05	250/250	0/0	0.10	0.19	0.05	0.32	0.02	0.12
		0/-.5	0.07	0.21	0.08	0.77	0.01	0.10
	250/500	0/0	0.07	0.17	0.08	0.29	0.01	0.11
		0/-.5	0.07	0.20	0.08	0.75	0.01	0.09
	500/500	0/0	0.05	0.17	0.09	0.29	0.01	0.12
		0/-.5	0.10	0.16	0.09	0.61	0.02	0.10
	500/1000	0/0	0.05	0.14	0.08	0.25	0.01	0.10
		0/-.5	0.05	0.17	0.07	0.69	0.01	0.07
	1000/1000	0/0	0.04	0.16	0.08	0.26	0.02	0.10
		0/-.5	0.02	0.16	0.07	0.74	0.01	0.08

*Note that *c* for the reference group was always simulated to be 0.2, while for the focal group it was simulated to be 0.2, 0.15, 0.1 or 0.05, as indicated in the table.

less bias for larger samples. The presence of group mean differences in theta (i.e., ability) was associated with a small increase in bias for item discrimination parameter estimate for the reference group. On the other hand, the level of *c*-DIF present did not appear to lead to item discrimination parameter estimation bias. This latter result would be expected given that the *c* parameter was always set at 0.2 for this group. For the focal group, estimates of *a* displayed comparable levels of positive bias in the no *c*-DIF (0.2/0.2) condition. However, as the degree of *c*-DIF increased, the severity of this positive bias increased as well, with the largest values occurring in the 0.2/0.05 *c*-DIF condition. As with the reference group, the extent of *a* bias for the focal group was ameliorated to some extent by larger sample sizes, and was unrelated to the presence of impact. Nonetheless, in all *c*-DIF conditions, item discrimination bias was greater for the focal compared to the reference group.

The population value of item difficulty was simulated to be -0.25. As was true for the *a* parameter, the estimates of *b* were somewhat positively biased for the reference group. However, contrary to what was found for discrimination, sample size was not associated with changes in estimation bias for the reference group, nor was the presence of impact or *c*-DIF. Neither of these latter results is surprising, as the mean of the latent trait and the value of *c* remained unchanged for the reference group across these conditions. The bias results for the focal group were substantially different compared to those of the reference. First, the presence of ability differences resulted in positive bias in the estimation of *b* regardless of sample size and *c*-DIF condition. Indeed, when comparing results for the focal and reference groups, *b* for the former was consistently close to 0.5 larger than that of the latter, corresponding to the simulated difference in the group means on the latent trait. In other words, the -0.5 mean on θ for the focal group was expressed in the item being approximately 0.5 more difficult for this group. In addition, the presence of *c*-DIF also was associated with increased levels of estimation bias for *b*. In other words, when the population value of *c* declined for the focal group, there was an increase in the overestimation of item difficulty. As was the case for the reference group, sample size did not influence the degree of item difficulty overestimation.

With regard to the *c* parameter, estimates for the reference group were consistently very close to the population value of 0.2, regardless of the sample size, ability difference condition, or *c*-DIF. For the focal group, parameter estimates were close to the population value of 0.2 in the no *c*-DIF case, as well. However, in the presence of *c*-DIF, pseudo-guessing parameter values were consistently overestimated for the focal group. For example, in the 0.2/0.15 condition, where the population focal group *c* was 0.15, the estimates obtained from BILOG were positively biased by between 0.03 and 0.05, meaning that the actual estimates were near 0.20. A similar and more severe pattern of positive bias was evident for greater levels of *c*-DIF, indicating that even when the population *c* was very low (e.g. 0.05) the estimate for the focal group continued to be above 0.10 and often above 0.15. These results, taken together with those for item difficulty, suggest that in the presence of *c*-DIF, the probability of an examinee obtaining a correct answer to an item by chance is consistently overestimated, which is associated with a concomitant overestimation in the difficulty of the item as well. Thus, it appears that when the popu-

lation probability of a chance right answer to an item was less than 0.20, this value inflated item difficulty, and the estimated pseudo-guessing value was too high.

Item parameter standard errors

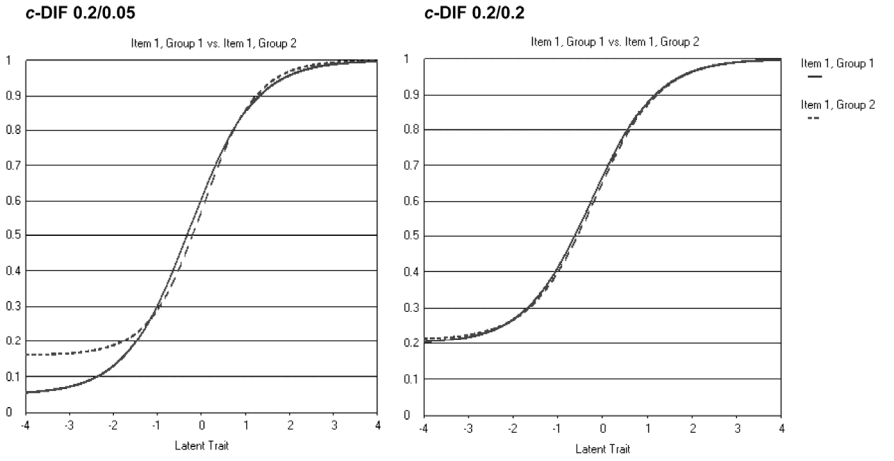
The mean of the item parameter estimate standard errors by the sample size and level of *c*-DIF appear in Table 3. Standard errors of all three item parameters for both groups were smaller for larger sample sizes. For the focal group, the mean standard error for each parameter estimate declined concomitantly with increases in the level of *c*-DIF, regardless of sample size. In contrast, the mean focal standard error for item discrimination increased with increasing *c*-DIF, particularly for the smaller sample size conditions. Finally, the focal group standard errors declined somewhat as the *c* parameter declined, particularly for the smaller sample sizes. The mean standard errors for the reference group item parameters were largely unaffected by the level of *c*-DIF.

Further investigation of parameter estimation bias

To gain a deeper understanding of the item parameter bias results, we can refer to item characteristic curves (ICC's) for specific items. These curves, which display in graphical format the item response functions (IRF's), may provide further insights into the impact of *c*-DIF on the estimation of other IRT model parameters. Figure 1 includes the ICC for the population parameters used to simulate the data and the mean sample estimates under the *c*-DIF=0.2/0.2 and *c*-DIF=0.2/0.05 conditions for the focal group. Group 1 refers to the population generating IRF, while Group 2 refers to the mean of the sample IRF's for the focal group. In general, the population and mean sample ICC's for the focal group for *c*-DIF=0.2/0.2 were extremely similar. On the other hand, when *c*-DIF=0.2/0.05 there was a marked difference between the focal group population and mean sample estimate ICC's, particularly at the lower end of the latent trait scale, with these two curves even crossing at two points. These ICC's highlight the distortion of the focal group IRF that is caused by the overestimation of *c* and *b* in particular, in the presence of *c*-DIF.

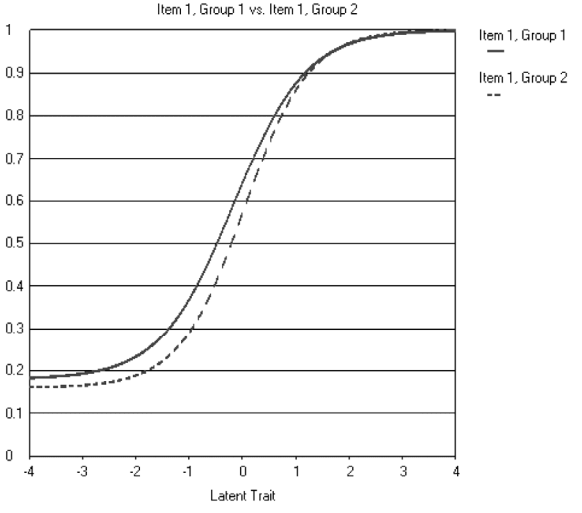
Figure 2 contains the sample ICC's for the reference and focal groups in the *c*-DIF=0.2/0.05 condition. When interpreting the two ICC's, it is important to keep in mind that the *a* and *b* for the two groups were simulated to be the same value, with only *c* simulated to be different. However, upon examination of these curves it is clear that there were substantial differences between the two groups, particularly for examinees with lower values of ability. Indeed, in this *c*-DIF condition, LR found uniform DIF to be present in 0.18 (18%) to 0.58 (58%) of simulated cases, and non-uniform DIF to be present at approximately the same rates, as presented in Table 2. IRTLR had uniform and non-uniform detection rates somewhat lower than those of LR in the 0.2/0.05 condition of *c*-DIF, though still elevated well above the nominal 0.05 level. These results were present despite the fact that neither uniform nor non-uniform DIF were simulated.

Figure 1:
Item Characteristic Curves for population and estimated models of target item for focal group at the c-DIF 0.2/0.05 and the c-DIF 0.2/0.2 conditions



(Group1=ICC based on population parameters; Group2=ICC based on sample estimates for focal group)

Figure 2:
Item Characteristic Curves for mean estimated IRF's of target item for Reference and Focal groups at the c-DIF 0.2/0.05 condition



(Group1=Reference group; Group2=Focal group)

Discussion

DIF analysis is a staple of the test development process, serving to ensure assessment fairness and validity. Given this importance, it is crucial that DIF detection be accurate and specific not only as to the presence of DIF, but also in regard to the type. Only in this way can measurement specialists and psychometricians begin to pinpoint changes to items and exams that might be needed to ensure fairness. The purpose of this study was to examine the impact of *c*-DIF on the estimation of 3-PL item parameters, and on the testing of uniform and non-uniform DIF for dichotomous items. The results demonstrate that when *c* differs between groups, the consequence is poor estimation of item difficulty and discrimination parameters and inflation of Type I error for uniform and non-uniform DIF detection. This pattern of parameter estimation bias and Type I error inflation held across sample sizes and differences (or not) between group means on the latent trait being measured. In addition, the severity of estimation bias increased concomitantly with increases in the level of *c*-DIF.

The results of this study present psychometricians and measurement practitioners with several implications for practice. Perhaps foremost among these is that the presence of *c*-DIF will have a direct impact on the estimation of item discrimination and particularly item difficulty parameters, which in turn may lead to incorrect findings of uniform and non-uniform DIF. In this study, despite the fact that item discrimination and difficulty parameters were simulated to be equal for the two groups, both LR and IRTLR had elevated rates of DIF detection for uniform and non-uniform DIF in the presence of *c*-DIF. Further investigation of this result revealed that bias in difficulty and discrimination for the focal group became more pronounced as the *c* parameter for the focal group became smaller or showed a larger difference between groups. In addition, the estimate of *c* was biased upward for the focal group when it was below 0.2 in the population. Therefore, it appears that inaccuracy in the estimation of lower *c* values led to an overestimation of other item parameters, particularly difficulty, for the focal group. In turn, this distortion of the IRF resulted in elevated findings of uniform and non-uniform DIF for both IRTLR and LR. Of additional interest in this regard was that IRTLR had difficulty correctly detecting the presence of *c*-DIF, which was likely caused by the upward bias in the focal group *c* parameter estimate.

These results point to the need for researchers to think carefully about both the estimation of item parameters and the interpretation of DIF results. As an example, when using the BILOG software package to estimate item parameters, it is possible for the researcher to set prior distributions for discrimination, difficulty and pseudo-guessing that may reflect the actual item parameters more accurately than the program default values. Indeed, for this study a very small number of example datasets were analyzed to assess the impact on the estimates of all three IRT parameters for the focal group of changing priors on *c* to more accurately reflect what was true in the population. We found that indeed, estimation bias in *a*, *b*, and *c* could be reduced through judicious selection of priors. However, in practice researchers may not know or even suspect that the value of *c* for some members of the sample might be radically different from that of others. In addition, changing these prior distributions in the process of parameter estimation will not address

the problem of inflated Type I error rates for testing uniform and non-uniform DIF using LR. On the other hand, if a researcher conducted the IRTLR tests for DIF manually using MULTILOG, rather than the IRTLR software, it would be possible to adjust the priors for *c*, if it were known that *c*-DIF were present.

Knowing if *c*-DIF is present will require developing data screening methods to detect if groups are exhibiting differently guessing behavior or a certain response bias related to rapid guessing. Future work focused on such screening measures will be important, as there is evidence that groups exhibiting such behavior can differ in their means on measured traits compared to groups not exhibiting such behavior (Osborne & Blanchard, 2011). These misleading differences may be a result of this differential guessing behavior leading to DIF and incorrect score comparisons across groups, which in turn has implications for assessing group differences in many environments from educational achievement tests to examination of the influence of intervention programs. We want to be sure we are comparing the best estimate of ability and not one clouded with many sources of error, which we may be able to control, such as differential guessing behavior.

Another implication of these results is with regard to the interpretation of significant uniform and non-uniform DIF results. It seems possible that in some cases, findings of DIF for *a* and *b* could actually be caused, at least in part, by the presence of *c*-DIF. Therefore, researchers should follow up such significant DIF results with a careful examination of the groups' item parameter estimates, including *c*. While such recommendations for carefully examining IRF's as standard follow up in DIF studies have been made previously (Finch & French, 2007), these typically revolve around difficulty and discrimination, and not the pseudo-guessing parameter. However, clearly when groups have different pseudo-guessing population values, all DIF testing may be called into question. Indeed, it is possible that in some cases where uniform or non-uniform DIF are found but the substantive cause of this DIF are unclear, *c*-DIF might be the culprit. The importance of this issue extends beyond the widely used 3-PL model to the constrained 3-PL model described by Kubinger and Draxler (2007). For this difficulty plus guessing (DGPL) model, item discrimination is constrained to be equal across items, but item difficulty and pseudo-guessing are each estimated. Given that one of the primary impacts of differential pseudo-guessing is on the distortion of the item difficulty parameter estimate, it would seem reasonable to assume that this distortion would also be present for the DGPL model, as it was for the 3-PL. However, future research should investigate this issue to ensure that it is indeed the case.

Test developers and analysts also may consider the possibility of *c*-DIF when writing and reviewing individual items. For example, attention may be needed to ensure that inadvertent clues are not embedded in incorrect options in multiple-choice tests that might allow members of some examinee groups but not others to eliminate these options from consideration and thus improve their likelihood of correctly guessing an item. Referring to item writing guidelines does assist this process and the number of guidelines to consider is numerous. Haladyna and Downing (1989a, 1989b), for instance suggest 43 item-writing guidelines. However, much like the lack of attention to guessing behavior, there is little, if any, empirical support for over half of these guidelines (Haladyna, 2004). Future work should focus on systematic empirical examinations of writing guidelines

that deal with distracters and guessing behavior in multi-choice items, for example, with the goal of reducing this probability of this behavior due to the item characteristics.

A final implication to be drawn from this study is that the detection of *c*-DIF using statistical tools may be problematic. The only one of the major DIF detection procedures that allows for such testing, IRTLR, had very low *c*-DIF detection rates across all of the conditions simulated in this study, despite the fact that in some cases the focal group's *c* value was one-fourth that of the focal group. Indeed, the highest rate of such detection was 0.16, which occurred for the largest sample size and greatest level of *c*-DIF. In addition, estimates of the focal group *c* using BILOG tended to be upwardly biased toward the population value of the Reference group (0.2), which likely led to the low detection rates for IRTLR, and which would make even a visual comparison of the groups' *c* values for DIF somewhat difficult. As discussed above, this problem could be ameliorated to some extent by the judicious selection of a prior distribution for *c*. Such a selection would require that the researcher be cognizant of the possibility that the *c* parameter for one or more groups in the sample may vary to some extent, and some notion as to what the appropriate prior distribution should be. While neither of these expectations is unreasonable, they also are not typically central to an IRT analysis, which tends to be focused more on item difficulty and discrimination. Future DIF work may need to focus on more accurate estimation of the *c*-parameter and methods to increase accurate detection of *c*-DIF.

In summary, measurement professionals must be sensitive to the possibility of *c*-DIF and its potential impact on both IRT parameter estimation and testing for both uniform and nonuniform DIF. Ignoring the potential for *c*-DIF could lead to poor estimation of item difficulty and discrimination parameters and false findings of uniform/non-uniform DIF. In turn, such results could lead to the needless changing of existing items and development of new items to replace those falsely identified as containing uniform or non-uniform DIF. In addition, the root cause of the problem, *c*-DIF, would be ignored and not corrected. More importantly, such differences in guessing behavior, as captured in *c*-DIF, may lead to uniform and non-uniform DIF items which could, in turn, accumulate to influence mean score levels for groups. This artificial change in scores calls into question the validity of score inferences and hence decisions we make about groups and individuals. The extent to which this can occur is critical to understand in future work.

References

Bandalos, D.L. & Leite, W. (2013). Use of Monte Carlo studies in Structural Equation Modeling research. In G.R. Hancock & R.O. Mueller (Eds.), *Structural Equation Modeling: A second course*. Charlotte, NC: Information Age Publishing.

Bao, H., Dayton, C.M., &, Hendrickson, A.B. (2009). Differential item functioning amplification and cancellation in a reading test. *Practical Assessment, Research & Evaluation, 14(19),* 1-27.

Cohen, A. S. & Kim, S.-H., (1993). A comparison of Lord's chi-square and Raju's area measures in detection of DIF. *Applied Psychological Measurement, 17,* 39-52.

Cohen, A.S., Kane, M.T., &, Kim, S-H. (2001). The precision of simulation study results. *Applied Psychological Measurement, 25(2),* 136-145.

Clauser, B., Mazor, K., &, Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6*, 269-279.

DeMars, C. E. & Wise, S. L. (2010). Can differential rapid-guessing behavior lead to differential item functioning? *International Journal of Testing, 10,* 207-229.

DeMars, C. E. & Wise, S. L. (2007, April). *Can differential rapid-guessing behavior lead to differential item functioning?* Paper presented at the American Educational Research Association Conference, Chicago.

Dunn, J. L. & Allen, J. (2009). Holding Schools Accountable for the Growth of Nonproficient Students: Coordinating Measurement and Accountability, *Educational Measurement: Issues and Practice,* 28*,* 27-41.

Finch, W.H. & French, B.F. (2007). Detection of Crossing Differential Item Functioning: A Comparison of Four Methods. *Educational and Psychological Measurement, 68,* 742-759.

Haladyna, T.M. (2004). *Developing and validating multiple-choice test items (3rd ed.)*. Mahwah, NJ: Lawrence Earlbaum Associates.

Haladyna, T.M. & Downing, S.M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1*, 37-50.

Haladyna, T.M. & Downing, S.M. (1989b). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1*, 51-78.

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

Kubinger, K.D. & Draxler, C. (2007). A comparison of the Rasch model and constrained item response theory models for pertinent psychological test data. In M. von Davier & C.H. Carstensen, Ed.s, *Multivariate and Mixture Distribution Rasch Models: Extensions and applications.* New York: Springer.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

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

Narayanan, P. & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and Simultaneous Item Bias procedures for detecting Differential Item Functioning. *Applied Psychological Measurement*, *20,* 257-274.

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

Osborne J. W. & Blanchard, M. R. (2011). Random responding from participants is a threat to the validity of social science research results. *Frontiers in Psychology, 21,* 1-7.

Penfield, R. D. (2003). IRT-Lab: Software for research and pedagogy in item response theory. *Applied Psychological Measurement*, *27*, 301–302.

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.

Roussos, L. A. & Stout, W. F. (1996). Simulation studies of the effects of small sample and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215-230.

Schumacker, R. E. (2005). *Test bias and differential item functioning.* Applied Measurement Associates.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-stage mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34,* 213-232.

Thissen, D., Steoinber, L., &, Wainer, H. (1988). Use of item response theory in the study of group difference in trace lines. In H. Wainer & H. Braun (Eds.), *Test Validity,* (pp. 147-170). Hillsdale, NJ: Lawrence Erlbaum.

Walstad, W.B. & Robson, D. (1997). Differential Item Functioning and Male-Female Differences on Multiple-Choice Tests in Economics. *The Journal of Economic Education, 28(2)*, 155-171.

Wise,. S. L. & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18,* 163-183.

Zimowski, M. F., Muraki, E., Mislevy, R. J., &, Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago: Scientific Software International.