Differential Item Functioning: Its Consequences

Yi-Hsuan Lee

Jinming Zhang

January 2010

ETS RR-10-01

Differential Item Functioning: Its Consequences

Yi-Hsuan Lee

Educational Testing Service, Princeton, New Jersey

Jinming Zhang

University of Illinois at Urbana-Champaign

February 2010

Abstract

This report examines the consequences of differential item functioning (DIF) using simulated data. Its impact on total score, item response theory (IRT) ability estimate, and test reliability was evaluated in various testing scenarios created by manipulating the following four factors: test length, percentage of DIF items per form, sample sizes of different examinee groups, and types of responses. The results indicate that the greatest score difference was observed between the examinee groups on forms with DIF items, and the magnitude was less than 2 points on the 0–60 total score scale and .15 on the IRT ability scale. The influence on reliability was rather limited.

Key words: Differential item functioning, total score, expected a posteriori, reliability

Acknowledgments

The authors thank Neil Dorans, Dan Eignor, Shelby Haberman, and Hyeon-Joo Oh for very helpful comments and Kim Fryer for her editorial assistance. Any opinions expressed in this report are those of the authors and not necessarily of ETS.

Introduction

Fairness concerns are often framed in terms of test bias favoring or disadvantaging groups of examinees by gender, native language, ethnicity, or socioeconomic status (Roever, 2005). The issue of item/test bias received extensive attention in admission testing and licensing exams through the Golden Rule Settlement in 1984. Nowadays, item bias is initially investigated through the detection of differential item functioning (DIF). As described by Dorans and Holland (1993), DIF refers to a difference in item performance between two comparable groups of examinees, usually named as *reference group* and *focal group*. In other words, DIF is an unexpected difference in performance among groups who are supposed to be comparable. The logical first step in detecting bias is to find items on which one group performs better than the other after matching on the construct being measured by the test. However, the presence of DIF is not equivalent to test/item bias. If, and only if, the presence of DIF can be attributed to unintended item content is the item said to be unfair (Penfield & Camilli, 2007). As Zieky (1993) pointed out,

> it is important to realize that DIF is not a synonym for bias.... The judgment of fairness is based on whether or not the difference in difficulty is believed to be related to the construct being measured.... The fairness of an item depends directly on the purpose for which a test is being used. For example, a science item that is differentially difficult for women may be judged to be fair in a test designed for certification of science teachers because the item measures a topic that every entry-level science teacher should know. However, that same item, with the same DIF value, may be judged to be unfair in a test of general knowledge designed for all entry-level teachers. (p. 340)

In other words, DIF itself does not indicate whether an item is fair or not; fairness depends on use. Discussions about DIF should be put in a fairness framework that includes the fair use of test scores (see, e.g., Dorans, 2004).

Statistical procedures are routinely employed for the identification of DIF. Examples are the Mantel-Haenszel method (Holland & Thayer, 1988), the standardization procedure

(Dorans & Kulick, 1986), the general item response theory (IRT) likelihood ratio approach (Thissen, Steinberg, & Wainer, 1988), and the simultaneous item bias test (Shealy & Stout, 1993a, 1993b). A summary of DIF methodologies can be found in Mapuranga, Dorans, and Middleton (2008).

The Mantel-Haenszel procedure is typically used in DIF analyses at ETS. A corresponding categorization scheme (Zieky, 1993) classifies items into three categories : A (items exhibiting negligible levels of DIF), B (items exhibiting moderate levels of DIF), and C (items exhibiting large levels of DIF). Statistical measures of DIF are used as an empirical check on the fairness of items (ETS, 2009). If DIF data are available, tests are assembled following rules that keep DIF low. If data are unavailable at assembly, DIF is calculated after test administration. Items with C DIF are reviewed for fairness by panels of people who have no vested interest in the test, and those items are left in or dropped on the basis of judgment about the construct relevance of the items. In practice, only a few items with C DIF are typically removed before scoring and reporting. Items displaying B DIF may remain in the test form without challenge. On the other hand, items with true DIF may not be flagged (significant) because of type II error of a statistical procedure. Responses to those items will be included in scoring and reporting. Thus there is cause for concern as to the impact of DIF on measurement consequences in different testing scenarios. In this study, *measurement consequences* refer to the reliability of a test form and ability estimation of examinees; the latter involves the total number of right scores and IRT ability estimates. With the use of simulated data, it is possible to compare measurement consequences not only across examinee groups but also across different test forms that diverge only in the presence of DIF items. Results of this study can provide bounds on the likely effects of DIF.

It is worth noting that all introduced DIF items were generated to favor the reference group uniformly across the ability continuum. If items in one test form display uniform DIF but favor different groups, there will likely be variations between examinee groups in item performance but not in test performance. Items exhibiting crossing DIF would yield weak or ignorable score differences across examinee groups even at the item level.

The remainder of this section introduces notation. Let θ denote the IRT ability parameter, which is assumed to be unidimensional, and let $\hat{\theta}$ be its estimate. Parameters a, b, and c stand for the item discrimination, difficulty, and guessing parameters, respectively. A test form with M items is administered to N_g examinees in group g; $g = R$ for the reference group and $g = F$ for the focal group. A DIF-free item refers to an item that is free of DIF. A DIF-free form is a test form in which all items are DIF-free, whereas a DIF form is one that contains some DIF items. Let K denote the number of DIF items in the M-item test form, $0 \leq K \leq M$. The Y_{nm} is the dichotomous response of person n in group g to item m, $1 \leq n \leq N_g$ and $1 \leq m \leq M$, with $Y_{nm} = 1$ for a correct response and $Y_{nm} = 0$ for an incorrect response. The total score of the same person is $T_n = \sum_{m=1}^{M} Y_{nm}$. The average total score for group g on a DIF-free form is equal to $\bar{T}_{g,f}$. Similarly, the average total score on a DIF form is $\bar{T}_{g,d}$. With IRT ability estimates $\hat{\theta}_n$ for examinees in group g, $1 \leq n \leq N_g$, the average abilities based on a DIF-free form and a DIF form are $\bar{\hat{\theta}}_{g,f}$ and $\bar{\hat{\theta}}_{g,d}$, respectively.

Method

Data Generation

One base form with 60 DIF-free items was simulated with parameters coming from the following distributions: $a \sim N(1.22, .7)$, with $a \geq .3$; $b \sim N(0, .72)$; and $c \sim N(.2, .06)$, with $0 \leq c \leq .6$. They appear in Table 1. The base form was fixed throughout the study, and forms with various amounts of DIF were composed using base form items. On the other hand, ability θ was distributed $N(\mu_g, 1)$. Group ability difference was introduced through μ_g. The case in which $\mu_R = \mu_F = 0$ assumed no group ability difference ($d_t = 0$), whereas the case where $\mu_R = .25$ and $\mu_F = -.25$ considered the presence of group ability difference (of size $d_t = .5$).

Design

Four factors varied in generating item responses to the two groups: test length ($M = 30$ or 60), percentage of DIF items per form (5%, 10%, or 20%), sample sizes of reference groups and focal groups (N_R/N_F = 500/250, 1,000/500, or 1,500/1,500), and

Table 1
Item Parameters for the Base Form

Item number	*a*	*b*	*c*	Item number	*a*	*b*	*c*
1	0.941	0.210	0.236	31	0.300	−0.660	0.288
2	0.323	−0.605	0.284	32	0.904	−1.116	0.271
3	1.369	−0.423	0.096	33	0.801	−0.969	0.228
4	0.383	−0.624	0.216	34	0.733	−0.441	0.217
5	0.300	−0.643	0.116	35	0.300	−0.912	0.269
6	0.680	−0.166	0.199	36	0.824	−1.173	0.282
7	0.544	−0.227	0.108	37	1.833	0.058	0.141
8	0.300	−0.467	0.185	38	0.300	0.182	0.218
9	1.513	−0.159	0.163	39	1.592	0.648	0.232
10	0.520	−0.374	0.331	40	0.465	0.365	0.123
11	1.358	−0.956	0.178	41	1.204	−0.477	0.143
12	1.487	0.759	0.187	42	1.438	0.295	0.225
13	1.304	−0.415	0.215	43	0.300	−0.228	0.127
14	2.087	0.437	0.151	44	1.349	1.148	0.187
15	1.139	0.044	0.132	45	0.300	−0.408	0.234
16	1.434	0.706	0.218	46	0.696	−0.350	0.225
17	1.368	−0.633	0.269	47	1.371	0.448	0.072
18	1.236	0.895	0.215	48	0.300	−1.102	0.248
19	0.300	0.557	0.229	49	1.290	−0.034	0.260
20	1.619	−0.680	0.281	50	0.993	1.036	0.203
21	1.503	0.208	0.025	51	1.798	−0.451	0.147
22	0.684	−0.575	0.166	52	1.401	0.247	0.085
23	0.995	−1.050	0.195	53	1.148	0.573	0.172
24	1.024	0.671	0.223	54	1.583	0.353	0.201
25	1.122	0.131	0.336	55	0.704	−1.089	0.211
26	0.862	0.349	0.161	56	1.614	1.372	0.239
27	2.481	0.565	0.187	57	2.063	0.764	0.246
28	0.939	0.391	0.183	58	1.307	−0.121	0.153
29	1.033	0.166	0.281	59	1.054	1.131	0.193
30	2.339	−0.381	0.147	60	1.510	1.085	0.259

Note. The true item parameters for the 30-item test are the first 30 items.

types of items (free-response or multiple-choice items). These four factors were crossed to yield 36 conditions. It should be noted that the three percentages of DIF items per form were chosen to illustrate how and to what extent the measurement consequences may be affected as the percentage increases. Although it is quite unlikely for testing programs to allow 10% or 20% of DIF items to remain in the tests after they do item screening, our results provide an upper bound for the impact.

A M-item DIF-free form contained the first M items in the base form. The number of DIF items K, each corresponding to a specific combination of test length and percentage of DIF items per form, was equal to 1 (rounding down from 1.5), 3, or 6 for a 30-item test and 3, 6, or 12 for a 60-item test. Table 2 summarizes the six sets of DIF items based on M and K. A M-item DIF form was created by substituting K DIF items for the last K items in a M-item DIF-free form. The item number in Table 2 indicates the position where a DIF item was inserted to replace a DIF-free item in a DIF-free form. In this study, DIF was introduced through the b parameter, so there were two true difficulties for an item: b_R for reference groups and b_F for focal groups. Take $M = 30$ and $K = 1$, for example. This corresponds to a 30-item DIF form whose first 29 items are Items 1–29 in Table 1, and the 30th item has parameters $a = 2.339$, $b_R = -.556$, $b_F = -.206$, and $c = .147$, as indicated in Table 2. A DIF-free form composed of Items 1–30 in Table 1 is said to be *paired* with this DIF form if the following three conditions are present: the forms have identical first 29 items, their 30th items have the same a and c parameters, and the difficulty parameters of their 30th items satisfy $b = (b_R + b_F)/2 = -.381$, where b is the difficulty parameter of the 30th DIF-free item. The same rule was applied to other combinations of M and K to produce other paired (DIF and DIF-free) forms for comparison in section 2.3. Notice that $b_R < b_F$ for all cases in Table 2 because DIF was designed to favor the reference groups.

The distance between b_R and b_F was chosen so that each DIF item exhibited large enough DIF (i.e., B DIF or C DIF). To determine this, the NAEP NDIF program (Kulick, 2000) was used to analyze 100 sets of responses generated from a two-parameter logistic (2PL) model with θ distributed $N(0, 1)$ and sample sizes $N_R/N_F = 1,500/1,500$. The program conducts DIF analysis based on both Mantel-Haenszel and standardization

Table 2
Parameters of Differential Item Functioning (DIF) Items

		Item					Category	Category
M	K	number	a	b_R	b_F	c	B	C
30	1	30	2.339	−0.556	−0.206	0.147	99	1
30	3	28	0.939	0.141	0.641	0.183	86	14
		29	1.033	−0.084	0.416	0.281	80	20
		30	2.339	−0.556	−0.206	0.147	75	25
30	6	25	1.122	−0.094	0.356	0.336	29	71
		26	0.862	0.099	0.599	0.161	67	32
		27	2.482	0.390	0.740	0.187	3	97
		28	0.939	0.141	0.641	0.183	46	54
		29	1.033	−0.084	0.416	0.281	18	82
		30	2.339	−0.556	−0.206	0.147	2	98
60	3	58	1.307	−0.371	0.129	0.153	91	9
		59	1.054	0.832	1.432	0.193	91	9
		60	1.510	0.835	1.335	0.259	91	9
60	6	55	0.704	−1.389	−0.789	0.211	81	18
		56	1.614	1.122	1.622	0.239	48	52
		57	2.063	0.514	1.014	0.246	48	52
		58	1.307	−0.371	0.129	0.153	48	52
		59	1.054	0.832	1.432	0.193	48	52
		60	1.510	0.835	1.335	0.259	48	52
60	12	49	1.290	−0.284	0.216	0.260	0	100
		50	0.993	0.736	1.336	0.203	8	92
		51	1.799	−0.701	−0.201	0.147	0	100
		52	1.401	−0.003	0.497	0.086	1	99
		53	1.148	0.273	0.873	0.172	0	100
		54	1.583	0.053	0.653	0.201	0	100
		55	0.704	−1.389	−0.789	0.211	70	26
		56	1.614	1.122	1.622	0.239	8	92
		57	2.063	0.514	1.014	0.246	0	100
		58	1.307	−0.371	0.129	0.153	1	99
		59	1.054	0.832	1.432	0.193	2	98
		60	1.510	0.835	1.335	0.259	3	97

procedures. Its output includes the ETS categorization scheme. Running DIF analysis once using one set of responses yielded one DIF classification for each item; such an analysis was repeated 100 times so that 100 DIF classifications were available for each item. Table 2 includes two columns for the number of times each selected item was identified as displaying B DIF or C DIF among the 100 analyses. It appears that those selected items were almost 100% diagnosed as displaying DIF to the amount equivalent to category B plus category C.

The chosen sample sizes led to *combined examinee groups* of 750, 1,500, and 3,000. In a regular testing scenario, a combined examinee group receives one test form. An additional assumption can be made in a simulated testing scenario: The combined examinee group receives not only a DIF form but also the paired DIF-free form. The measurement consequences obtained from the DIF and DIF-free forms can then be compared.

The difference between free-response items and multiple-choice items was reflected by the underlying probability models for a correct response. It was assumed that a free-response item cannot be answered correctly by guessing, but such a possibility exists for a multiple-choice item. In addition, a test form either solely comprised free-response items or multiple-choice items. Let a_m, b_m, and c_m be the item parameters of item m. This study used the slope and location formulation of the 2PL model, given by

$$P(Y_{nm} = 1|\theta_n) = \frac{1}{1 + \exp[-1.7a_m(\theta_n - b_m)]},$$

to generate responses to forms with free-response items. Analogously, the 3PL model,

$$P(Y_{nm} = 1|\theta_n) = c_m + \frac{1 - c_m}{1 + \exp[-1.7a_m(\theta_n - b_m)]},$$

was used to produce responses to forms with multiple-choice items. Once Y_{nm} were obtained, they were regarded as observed responses in real tests from unknown probability models.

One hundred data sets (replications) were generated for each of the 36 conditions. For each condition, the item parameters and the ability distributions for both the reference and focal groups were fixed. However, a new set of θ was generated from the ability distributions at each replication.

Evaluation

For each condition, the total score of each examinee was computed and further summarized within either group. Two types of comparisons can be conducted with simulated data. First, one can compare the average total scores of the reference group and the focal group on each test form. This is referred to as *group difference* and was defined as $D_{f,\mathrm{RF}} = \bar{T}_{R,f} - \bar{T}_{F,f}$ for DIF-free forms and $D_{d,\mathrm{RF}} = \bar{T}_{R,d} - \bar{T}_{F,d}$ for DIF forms. Second, either group received one DIF form and the paired DIF-free form. It is feasible to assess the difference in average total scores between the two forms, which determines the impact of DIF on total score, if the group received a DIF form rather than the DIF-free form. This compares the *form difference* through $D_{g,fd} = \bar{T}_{g,f} - \bar{T}_{g,d}$ for $g = R$ or F.

These comparisons can be analogously defined with respect to IRT θ estimates. The maximum likelihood estimate (MLE) and expected a posteriori (EAP) estimate for θ were implemented using the expectation-maximization algorithm. It is known that the 2PL model can provide more stable parameter estimation than the 3PL model. Thus the 2PL model was employed to model the item responses for both types of items, regardless of the true underlying probability models. The group differences in terms of θ were evaluated by $D'_{f,\mathrm{RF}} = \bar{\hat{\theta}}_{R,f} - \bar{\hat{\theta}}_{F,f}$ for DIF-free forms and $D'_{d,\mathrm{RF}} = \bar{\hat{\theta}}_{R,d} - \bar{\hat{\theta}}_{F,d}$ for DIF forms. The $D'_{g,fd} = \bar{\hat{\theta}}_{g,f} - \bar{\hat{\theta}}_{g,d}$ defined the form difference for group g, $g = R$ or F.

Cronbach's alpha was computed to estimate the reliability of each test form using responses of each combined examinee group. Recall that each combined examinee group received one DIF form and the paired DIF-free form. Thus a reliability estimate can be obtained for either form based on the same examinees. If there is a noticeable difference between these two estimates, it is due to the existence of DIF items.

All the measures discussed were computed once using each data set and were averaged across 100 replications. Variability across replications was also examined. In addition, two-sided t tests were performed for each condition to see if the observed differences were significantly nonzero. The test statistics followed a t distribution with 99 degrees of freedom. Significance level was set to be .05. The results are described in the following section.

Results

Total Score Differences: Reference Group Versus Focal Group

Table 3 summarizes the total score difference between groups. The results for test forms with free-response items or multiple-choice items are labeled 2PL or 3PL, respectively. Take the 2PL results, for example: Ideally, there should be no total score difference, except for random noise, if both groups received the same DIF-free form, given no group ability difference. This is verified since $D_{f,\mathrm{RF}}$ are small and around zero for most conditions, and the two-sided t tests for the hypothesis that $D_{f,\mathrm{RF}} = 0$ are nonsignificant for all conditions given no group ability difference. The group differences on DIF forms, $D_{d,\mathrm{RF}}$, are consistently positive. The observation coincides with our intention to introduce DIF to favor the reference groups. In addition, $D_{d,\mathrm{RF}}$ increases as K increases. DIF forms containing the same number of DIF items have similar group differences, regardless of the actual test length or the actual percentage of DIF items. For example, the value of $D_{d,\mathrm{RF}}$ for 30-item forms with 20% of DIF items is approximately equal to that for 60-item forms with 10% of DIF items since all forms have six DIF items; the approximation is better for larger sample sizes. In general, longer tests and/or higher percentages of DIF items per form tend to yield larger group differences on DIF forms. The two-sided t tests for the hypothesis that $D_{d,\mathrm{RF}} = 0$ demonstrate that the group differences on DIF forms are significant for all conditions, except for two cases where the examinees with sample sizes, $N_R/N_F = 500/250$, received forms only consisting of 5% of DIF items. The results for 3PL reveal similar phenomena, but the group differences are uniformly less remarkable in size than the corresponding 2PL results. The variability across replications is smaller for 3PL as well; the corresponding t tests are still significant for most conditions.

The presence of group ability difference contributes directly to $D_{f,\mathrm{RF}}$ but less directly to $D_{d,\mathrm{RF}}$. Consider the following example: When $N_R/N_F = 1,500/1{,}500$, $M = 60$, and $K = 12$, $D_{f,\mathrm{RF}} = 6.816$ when there is group ability difference, whereas $D_{f,\mathrm{RF}} = -.044$ when there is no group ability difference. The difference between these two numbers, $6.816 - (-.044) = 6.86$, suggests the effect due to group ability difference, and $-.044$ reflects

Table 3

Ability Comparison of New and Old Groups

				2PL		3PL	
N_R	N_F	M	K	$D_{f,\mathrm{RF}}$	$D_{d,\mathrm{RF}}$	$D_{f,\mathrm{RF}}$	$D_{d,\mathrm{RF}}$
Group ability difference = 0							
500	250	30	1	−0.031	0.103	0.002	0.091
500	250	30	3	0.029	0.393	−0.019	0.328
500	250	30	6	−0.057	0.690	−0.079	0.545
500	250	60	3	−0.223	0.183	−0.165	0.134
500	250	60	6	0.135	0.854	0.124	0.659
500	250	60	12	−0.004	1.678	0.022	1.335
1000	500	30	1	0.014	0.141	0.014	0.101
1000	500	30	3	−0.048	0.368	−0.005	0.294
1000	500	30	6	0.015	0.759	0.009	0.629
1000	500	60	3	−0.094	0.299	−0.063	0.228
1000	500	60	6	−0.037	0.706	−0.018	0.566
1000	500	60	12	−0.061	1.615	−0.047	1.302
1500	1500	30	1	0.007	0.111	−0.012	0.104
1500	1500	30	3	0.014	0.421	0.023	0.310
1500	1500	30	6	−0.025	0.764	−0.016	0.585
1500	1500	60	3	0.038	0.424	0.030	0.338
1500	1500	60	6	0.007	0.739	0.014	0.585
1500	1500	60	12	−0.044	1.653	−0.019	1.317
Group ability difference = .5							
500	250	30	1	3.614	3.731	2.896	3.004
500	250	30	3	3.713	4.140	2.971	3.308
500	250	30	6	3.689	4.412	2.984	3.554
500	250	60	3	7.155	7.545	5.785	6.134
500	250	60	6	7.240	7.921	5.780	6.420
500	250	60	12	7.106	8.718	5.703	7.063
1000	500	30	1	3.668	3.806	2.965	3.058
1000	500	30	3	3.692	4.071	2.975	3.277
1000	500	30	6	3.759	4.493	3.023	3.602
1000	500	60	3	7.041	7.419	5.651	5.980
1000	500	60	6	7.127	7.843	5.783	6.297
1000	500	60	12	7.272	8.934	5.845	7.157
1500	1500	30	1	3.625	3.734	2.918	3.025
1500	1500	30	3	3.626	4.000	2.927	3.204
1500	1500	30	6	3.643	4.386	2.938	3.520
1500	1500	60	3	7.087	7.465	5.685	6.015
1500	1500	60	6	6.982	7.683	5.627	6.208
1500	1500	60	12	6.816	8.456	5.510	6.809

Note. 2PL = two-parameter logistic; 3PL = three-parameter logistic.

the amount of random noise. Under the same condition with group ability difference, $D_{d,\mathrm{RF}} = 8.456$, among which 6.816 is mainly attributable to the group ability difference, and $D_{d,\mathrm{RF}} - D_{f,\mathrm{RF}} = 8.456 - 6.816 = 1.64$ reflects the effect of administering DIF forms rather than the DIF-free forms, with adjustment on the group ability difference. Such a DIF effect can be obtained for each condition in Table 3 by subtracting values in the sixth or eighth column ($D_{d,\mathrm{RF}}$) from the corresponding values in the fifth or seventh column ($D_{f,\mathrm{RF}}$) for 2PL or 3PL responses; the results for $N_R/N_F = 1,500/1{,}500$ are shown in Figure 1. Type of response and level of group ability difference are crossed to produce four situations. For each situation, one curve is formed by connecting with straight lines the six discrete points for the six combinations of test length and number of DIF item (at the x-axis), which makes it easy to observe possible patterns. It is clear that the curves for $d_t = 0$ and $d_t = .5$ are very close for either type of response, which suggests that the absence or presence of a group ability difference does not have a material impact on the observed group differences. The curves for 3PL responses are consistently beneath the curves for 2PL responses. Using test forms with 20% of DIF items results in the largest group difference for any test length, types of items, sample size, and group ability difference. The greatest DIF effect is found in the 2PL responses; the magnitude is about 1.7 points on the 0–60 total score scale.

Generally, varying sample size does not have a direct influence on the averaged total score differences. However, larger sample sizes stabilize the variance of total score differences across replications, and a similar pattern remains for different types of responses and for different levels of group ability difference ($d_t = 0$ or .5). As an example, Figure 2 shows the variances of $D_{d,\mathrm{RF}}$ for 2PL responses under each of the six combinations of test length and number of DIF items. The group ability difference is zero. Analogous to Figure 1, three curves are formed to represent three pairs of sample sizes. It is apparent that larger sample sizes correspond to smaller variances under each condition. When the sample sizes are increased from $N_R/N_F = 500/500$ to $N_R/N_F = 1,500/1{,}500$, the proportional reduction in variability is 46% for the 30-item tests and 41% for the 60-item tests.

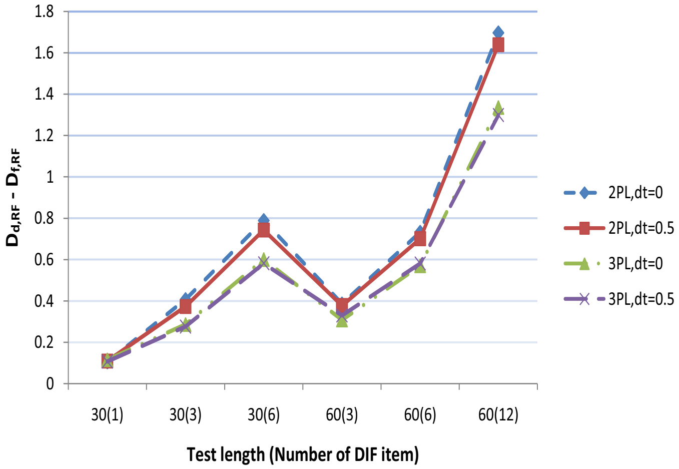

Figure 1. **Effect of administering differential item functioning (DIF) forms rather than DIF-free forms on the total score differences between groups with adjustment on the group ability difference; $N_R/N_F = 1,500/1,500$.**

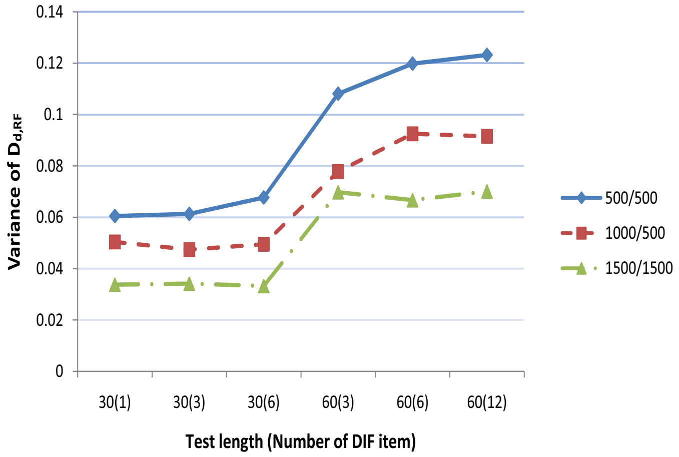

Figure 2. **Variance of the total score differences between groups across replications; DIF forms, two-parameter logistic responses, and $d_t = 0$.**

Total Score Differences: Differential Item Functioning (DIF) Form Versus DIF-Free Form

Table 4 shows the form difference for either group with 2PL or 3PL responses. Group ability difference has no influence on the comparison between test forms. It is reasonable that $D_{R,fd} < 0$ and $D_{F,fd} > 0$ consistently, for DIF forms advantage the reference groups and disadvantage the focal groups by design. The amount of form difference again increases as K increases. Test forms containing the same number of DIF items have similar form differences for either group, regardless of the actual test length or the actual percentage of DIF items. For test forms of the same length, increasing the percentage of DIF items in each form leads to greater form differences. The magnitude of form differences is less substantial when 3PL responses are under consideration. Almost all the differences in Table 4 are significant ($p < .05$). For any test length, types of items, sample size, and level of group ability difference, the greatest form difference is observed when comparing DIF-free forms and DIF forms with 20% of DIF items; the magnitude is less than 0.9 points on the 0–60 total score scale. The increase in sample size corresponds to a decrease in the variability of form difference across replications.

Figure 3 is a graphical illustration of the relationship between the averaged form differences for focal groups ($D_{F,fd}$) and the six combinations of test length and number of DIF item. Types of responses and level of group ability difference are crossed to create four situations in the figure, and one curve is formed to represent each situation. The sample sizes are $N_R/N_F = 1,500/1{,}500$. Apparently, the trend of the curves is the same, showing larger total score differences between forms for greater numbers of DIF items. For each curve, the values of form differences are quite close for test forms with the same number of DIF items (e.g., 30(3) vs. 60(3) and 30(6) vs. 60(6)). For the same type of responses, the form differences tend to be smaller when group ability difference is present. The magnitude of form differences is larger for 2PL responses since the curves for 2PL responses are above those for 3PL responses in almost all conditions.

Overall, the effect for two groups receiving DIF forms (i.e., $D_{d,\mathrm{RF}}$) is greater than the effect of DIF for either group receiving both forms (i.e., $D_{R,fd}$ or $D_{F,fd}$) under each of

Table 4
Total Score Difference Between Forms

				2PL		3PL	
N_R	N_F	M	K	$D_{R,fd}$	$D_{F,fd}$	$D_{R,fd}$	$D_{F,fd}$
Group ability difference = 0							
500	250	30	1	−.069	.064	−.026	.063
500	250	30	3	−.189	.175	−.160	.186
500	250	30	6	−.372	.375	−.321	.303
500	250	60	3	−.208	.198	−.130	.168
500	250	60	6	−.379	.340	−.286	.249
500	250	60	12	−.868	.814	−.726	.587
1000	500	30	1	−.059	.067	−.040	.047
1000	500	30	3	−.208	.207	−.151	.147
1000	500	30	6	−.383	.362	−.298	.322
1000	500	60	3	−.192	.202	−.159	.133
1000	500	60	6	−.366	.377	−.308	.276
1000	500	60	12	−.870	.806	−.710	.640
1500	1500	30	1	−.044	.060	−.046	.070
1500	1500	30	3	−.208	.199	−.141	.145
1500	1500	30	6	−.396	.394	−.308	.293
1500	1500	60	3	−.192	.194	−.152	.156
1500	1500	60	6	−.388	.345	−.289	.282
1500	1500	60	12	−.889	.808	−.679	.657
Group ability difference = .5							
500	250	30	1	−.041	.076	−.039	.069
500	250	30	3	−.205	.222	−.141	.196
500	250	30	6	−.385	.338	−.305	.265
500	250	60	3	−.217	.173	−.174	.175
500	250	60	6	−.414	.267	−.301	.338
500	250	60	12	−.874	.738	−.706	.653
1000	500	30	1	−.051	.086	−.048	.045
1000	500	30	3	−.186	.192	−.160	.142
1000	500	30	6	−.372	.362	−.307	.272
1000	500	60	3	−.219	.159	−.143	.186
1000	500	60	6	−.396	.321	−.294	.220
1000	500	60	12	−.907	.755	−.714	.598
1500	1500	30	1	−.041	.069	−.046	.061
1500	1500	30	3	−.194	.179	−.148	.128
1500	1500	30	6	−.376	.367	−.308	.274
1500	1500	60	3	−.233	.145	−.180	.150
1500	1500	60	6	−.397	.305	−.318	.264
1500	1500	60	12	−.892	.748	−.705	.594

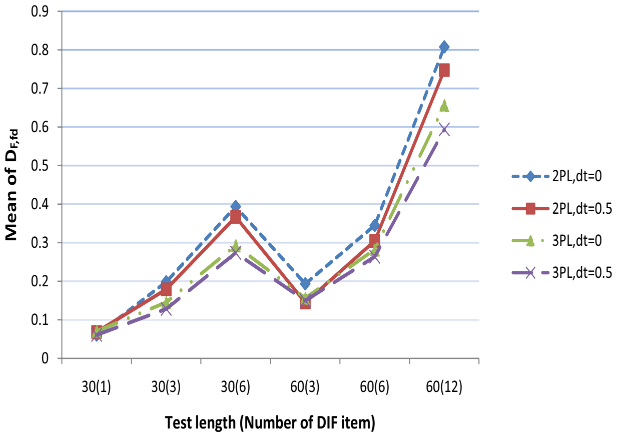

Figure 3. **Mean of the total score differences between forms across replications; focal groups and $N_R/N_F = 1,500/1,500$.**

the conditions. The largest total score difference is observed in the comparison between examinee groups on forms with 20% of DIF items. The greatest group difference is about 1.7 points on the 0–60 total score scale after adjusting the group ability difference.

Differences in θ Estimates: Reference Group Versus Focal Group

The MLEs and EAP estimates led to very similar results, so only results for EAP estimates are reported. Tables 5 and 6 show the group difference and form difference, respectively. The patterns are quite similar to the differences found on total scores, but the values are now on the θ scale. It is worth noting that the observed differences in θ estimates are comparable for 2PL and 3PL responses. As evident in Table 5, the reference groups are advantaged more with respect to $D^{'}_{d,\text{RF}}$ with a larger percentage of DIF items. Without group ability difference, the maximal $D^{'}_{d,\text{RF}}$ is equal to .153 for 3PL responses with $N_R/N_F = 1,500$/1,500, $M = 60$, and $K = 12$. The maximal effect of DIF is also .153. The $D^{'}_{d,\text{RF}}$ is significantly nonzero for all conditions, with the exception of one condition for 3PL with $N_R/N_F = 500/250$, $M = 60$, and $K = 3$. As mentioned earlier, it is less straightforward to interpret the effect of DIF in $D^{'}_{d,\text{RF}}$ as the group ability difference is .5. As an example, take the case of $N_R/N_F = 1,500$/1,500, $M = 60$, $K = 12$, and 3PL responses: The $D'_{d,\mathrm{RF}} = .641$ is a result of group ability difference, random noise, and the effect of DIF. The first two components lead to $D'_{f,\mathrm{RF}} = .511$, so the difference between .641 and .511 reveals the effect of DIF. The θ estimation is improved for a longer test, which is apparent in $D'_{f,\mathrm{RF}}$ when the group ability difference is equal to .5. The variance of group differences is smaller across replications for a larger sample size.

Differences in θ Estimates: DIF Form Versus DIF-Free Form

Table 6 shows the form difference with respect to EAP estimates. The existence of DIF items yields higher EAP estimates for the reference groups (i.e., numbers in Columns 5 and 7 are uniformly negative) and lower EAP estimates for the focal groups (i.e., numbers in Columns 6 and 8 are uniformly positive) for all conditions. All differences are significant at $p < .05$. Under certain conditions, the $D'_{R,fd}$ and $D'_{F,fd}$ are about symmetric around zero. Types of items and group ability difference do not have any impact on form differences. The increase in sample size reduces the variability of form differences across replications. The greatest form difference is .096 under the condition of $N_R/N_F = 1,000/500$, $M = 60$, $K = 12$, and 3PL responses.

Figure 4 is a graphical illustration of the relationship between the averaged form differences for focal groups ($D'_{F,fd}$) and the six combinations of test length and number of DIF items. Analogous to Figure 3, types of responses and level of group ability difference are crossed to produce four situations, and one curve is formed for one situation. The sample sizes are $N_R/N_F = 1,500/1{,}500$. It can be seen that all curves have the same pattern and, for each curve, the averaged $D'_{F,fd}$ increases when the test forms contain a higher percentage of DIF items. As compared to Figure 3, Figure 4 reveals one major difference between $D_{F,fd}$ and $D'_{F,fd}$: Curves for $d_t = 0$ overlap, and so do the curves for $d_t = .5$. This means that type of response is not an influential factor for differences in θ estimates. The form differences appear to be smaller as $d_t = .5$.

In sum, the effect for two groups receiving DIF forms (i.e., $D'_{d,\mathrm{RF}}$) is greater than the effect of DIF for either group receiving both forms (i.e., $D'_{R,fd}$ or $D'_{F,fd}$) under each of the conditions. Again, the greatest score difference in terms of EAP estimates is observed in

Table 5

Difference in Expected A Posteriori (EAP) Estimates Between Groups

				2PL		3PL	
N_R	N_F	M	K	$D'_{f,\text{RF}}$	$D'_{d,\text{RF}}$	$D'_{f,\text{RF}}$	$D'_{d,\text{RF}}$
Group ability difference = 0							
500	250	30	1	−.003	.024	.002	.028
500	250	30	3	.004	.060	.003	.067
500	250	30	6	−.011	.112	−.021	.107
500	250	60	3	−.012	.018	−.013	.016
500	250	60	6	.010	.075	.009	.064
500	250	60	12	−.002	.141	.000	.141
1000	500	30	1	.007	.035	.010	.034
1000	500	30	3	−.009	.058	.000	.061
1000	500	30	6	.002	.124	.006	.137
1000	500	60	3	−.007	.023	−.008	.025
1000	500	60	6	−.002	.066	−.001	.059
1000	500	60	12	−.005	.137	−.007	.138
1500	1500	30	1	.001	.031	.008	.037
1500	1500	30	3	.006	.072	.025	.085
1500	1500	30	6	.000	.129	.005	.130
1500	1500	60	3	.004	.038	.011	.043
1500	1500	60	6	.000	.067	.009	.067
1500	1500	60	12	−.002	.148	.007	.153
Group ability difference = .5							
500	250	30	1	.547	.575	.549	.578
500	250	30	3	.561	.623	.563	.627
500	250	30	6	.557	.671	.569	.677
500	250	60	3	.524	.556	.527	.559
500	250	60	6	.534	.591	.526	.590
500	250	60	12	.523	.651	.520	.652
1000	500	30	1	.553	.585	.561	.585
1000	500	30	3	.558	.617	.560	.622
1000	500	30	6	.565	.677	.571	.681
1000	500	60	3	.519	.550	.516	.549
1000	500	60	6	.528	.588	.531	.579
1000	500	60	12	.538	.671	.535	.661
1500	1500	30	1	.545	.570	.553	.581
1500	1500	30	3	.545	.603	.560	.617
1500	1500	30	6	.547	.665	.560	.675
1500	1500	60	3	.521	.552	.530	.562
1500	1500	60	6	.513	.575	.523	.579
1500	1500	60	12	.503	.637	.511	.641

Table 6
Difference in EAP Estimates Between Forms

				2PL		3PL	
N_R	N_F	M	K	$D'_{R,fd}$	$D'_{F,fd}$	$D'_{R,fd}$	$D'_{F,fd}$
Group ability difference = 0							
500	250	30	1	−.010	.017	−.009	.016
500	250	30	3	−.020	.036	−.022	.042
500	250	30	6	−.044	.078	−.044	.083
500	250	60	3	−.010	.020	−.009	.020
500	250	60	6	−.023	.043	−.017	.038
500	250	60	12	−.049	.094	−.049	.092
1000	500	30	1	−.010	.018	−.008	.017
1000	500	30	3	−.023	.044	−.021	.040
1000	500	30	6	−.044	.078	−.045	.086
1000	500	60	3	−.010	.019	−.011	.022
1000	500	60	6	−.024	.044	−.020	.040
1000	500	60	12	−.049	.093	−.049	.096
1500	1500	30	1	−.015	.015	−.014	.015
1500	1500	30	3	−.034	.033	−.029	.031
1500	1500	30	6	−.064	.065	−.062	.063
1500	1500	60	3	−.017	.017	−.016	.016
1500	1500	60	6	−.034	.033	−.028	.029
1500	1500	60	12	−.075	.075	−.072	.074
Group ability difference = .5							
500	250	30	1	−.011	.017	−.010	.019
500	250	30	3	−.022	.041	−.022	.042
500	250	30	6	−.041	.072	−.040	.067
500	250	60	3	−.012	.021	−.010	.022
500	250	60	6	−.021	.036	−.022	.042
500	250	60	12	−.045	.083	−.045	.087
1000	500	30	1	−.011	.021	−.008	.016
1000	500	30	3	−.020	.039	−.022	.040
1000	500	30	6	−.041	.072	−.040	.071
1000	500	60	3	−.011	.019	−.011	.022
1000	500	60	6	−.022	.039	−.017	.031
1000	500	60	12	−.048	.085	−.044	.081
1500	1500	30	1	−.012	.012	−.014	.014
1500	1500	30	3	−.029	.029	−.028	.029
1500	1500	30	6	−.059	.059	−.058	.057
1500	1500	60	3	−.016	.015	−.016	.016
1500	1500	60	6	−.032	.030	−.028	.028
1500	1500	60	12	−.068	.066	−.065	.065

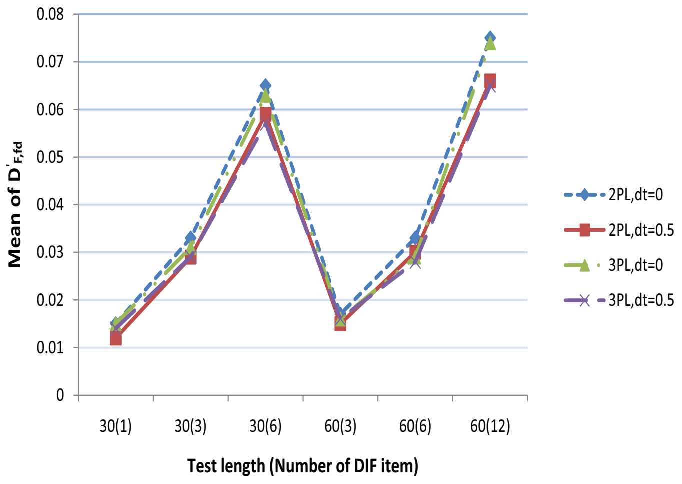

Figure 4. **Mean differences in θ estimates between forms across replications; focal groups and $N_R/N_F = 1,500/1,500$.**

DIF forms between examinee groups; the magnitude is around .15 on the θ scale. Note that DIF has a similar level of impact on θ estimates for both types of items.

Reliability Differences: DIF Form Versus DIF-Free Form

Table 7 displays the mean estimates of reliability for DIF-free forms and DIF forms. The difference in Columns 7 and 10 of Table 7 is defined as the subtraction of the mean estimate for a DIF form from the mean estimate for the paired DIF-free form. The tests are more reliable for 2PL responses than for 3PL responses. Longer tests are more reliable. Changing the percentage of DIF per form does not have a clear influence on test reliability. Sample size does not seem to interact with the mean estimates across replications, but the increase in sample size leads to a decrease in the variability of estimates across replications. The existence of group ability difference leads not only to a 1% increment in reliability but also to more significant differences.

The interesting finding is that the reliability of DIF forms is comparable to the reliability of the paired DIF-free forms when group ability difference is zero, regardless of the number of DIF items. When group ability difference is .5, the reliability estimates for

Table 7
Reliability

				2PL			3PL		
N_R	N_F	M	K	DIF-free	DIF	Difference	DIF-free	DIF	Difference
Group ability difference = 0									
500	250	30	1	.9188	.9181	.0007*	.8685	.8688	−.0003
500	250	30	3	.9193	.9190	.0003	.8699	.8703	−.0004
500	250	30	6	.9188	.9184	.0003	.8692	.8692	.0001
500	250	60	3	.9548	.9548	.0000	.9241	.9240	.0001
500	250	60	6	.9544	.9545	.0000	.9235	.9239	−.0004
500	250	60	12	.9544	.9544	.0000	.9236	.9240	−.0005*
1000	500	30	1	.9180	.9176	.0004	.8691	.8683	.0008*
1000	500	30	3	.9192	.9189	.0003	.8703	.8700	.0003
1000	500	30	6	.9170	.9165	.0004*	.8668	.8670	−.0002
1000	500	60	3	.9547	.9548	.0000	.9240	.9241	−.0001
1000	500	60	6	.9546	.9547	−.0001	.9236	.9241	−.0005*
1000	500	60	12	.9549	.9549	−.0001	.9243	.9248	−.0005*
1500	1500	30	1	.9107	.9104	.0003	.8597	.8596	.0002
1500	1500	30	3	.9090	.9085	.0005*	.8572	.8569	.0003
1500	1500	30	6	.9156	.9149	.0006*	.8656	.8646	.0011*
1500	1500	60	3	.9512	.9511	.0001	.9192	.9193	−.0001
1500	1500	60	6	.9530	.9529	.0001	.9213	.9214	−.0001
1500	1500	60	12	.9494	.9496	−.0002	.9169	.9169	.0000
Group ability difference = .5									
500	250	30	1	.9214	.9214	.0000	.8756	.8749	.0007
500	250	30	3	.9226	.9229	−.0004	.8762	.8770	−.0008
500	250	30	6	.9220	.9225	−.0005	.8753	.8773	−.0021*
500	250	60	3	.9571	.9573	−.0002	.9286	.9292	−.0005*
500	250	60	6	.9570	.9573	−.0003*	.9283	.9294	−.0011*
500	250	60	12	.9568	.9575	−.0007*	.9282	.9298	−.0015*
1000	500	30	1	.9220	.9220	.0000	.8758	.8754	.0004
1000	500	30	3	.9225	.9225	.0000	.8762	.8766	−.0004
1000	500	30	6	.9230	.9234	−.0004	.8771	.8789	−.0019*
1000	500	60	3	.9570	.9573	−.0003*	.9287	.9291	−.0004*
1000	500	60	6	.9570	.9575	−.0005*	.9287	.9295	−.0008*
1000	500	60	12	.9567	.9574	−.0007*	.9280	.9295	−.0014*
1500	1500	30	1	.9217	.9218	−.0002	.8736	.8739	−.0002
1500	1500	30	3	.9223	.9226	−.0004*	.8742	.8748	−.0007*
1500	1500	30	6	.9220	.9226	−.0005*	.8738	.8750	−.0012*
1500	1500	60	3	.9568	.9570	−.0002*	.9271	.9273	−.0002
1500	1500	60	6	.9565	.9568	−.0004*	.9268	.9272	−.0004*
1500	1500	60	12	.9563	.9567	−.0004*	.9263	.9272	−.0009*

$^*p < 0.05$.

DIF forms tend to be greater than those for the paired DIF-free forms because most of the differences are negative. The t tests for difference $= 0$ further suggest that the differences are significant for 22 cases out of the 36 cases shown in the lower half of Table 7. However, the magnitude of the difference is so small (less than .0007 for 2PL responses and .002 for 3PL responses) that the difference may still be negligible in practice. Two observations may be useful to explain the subtle increment in reliability estimates for DIF forms when group ability difference is .5. First, a comparison between the top half and bottom half of Table 7 indicates that the estimated reliability increases for DIF-free (or DIF) forms when the group ability difference is changed from 0 to .5. Second, recall that the group ability difference was introduced through different means for the two ability distributions. DIF was generated through the difficulty parameters (DIF items are easier for the reference groups and more difficult for the focal groups). DIF and group ability difference have the same direction of impact on examinees' performance, so the overall effect may be conceptualized as a result of a combined examinee group with a more drastic group ability difference (e.g., $d_t > .5$) taking the DIF-free forms. According to the first observation, it is not unreasonable to arrive at reliability estimates that are slightly higher for DIF-free forms and $d_t > .5$ than for the same forms and $d_t = .5$. The same logic applies: When $d_t = .5$, the estimated reliability is also likely to be higher for DIF forms than the paired DIF-free forms.

Remark

Another simulation study was carried out, in which DIF was regarded as a result of multidimensionality. Test forms were assumed to measure only one ability, the main ability, but some items intrinsically involved a secondary ability that differed on average for reference and focal groups. The same four factors discussed in section 2.2 were incorporated to produce various conditions. Responses for DIF-free items were simulated from a unidimensional 2PL or 3PL model, and those for DIF items were from a two-dimensional 2PL or 3PL model in which a and b parameters for the main ability were held to be the same as in the paired DIF-free items, but a positive a parameter and zero b parameter were introduced to the secondary ability. Those DIF items favored the reference group because examinees in the reference group had higher secondary abilities on average than examinees in the focal group. The results are comparable to those from the current design. The downside of generating DIF through multidimensional IRT models is that it takes two dimensions of ability and item parameters to jointly determine one item response function. Thus it is less straightforward to control the desired amount of DIF to introduce.

Discussion

In this study, we manipulated factors such as test length, percentage of DIF items per form, sample sizes of reference group and focal group, and types of responses to investigate the impact of DIF on measurement consequences in different testing scenarios. The study suggests that the absence or presence of group ability differences does not have a material effect on the observed impact of DIF items. (Note that DIF analysis procedures usually have more power and less type I error when there is no group ability difference, but in this study, we only focused on the impact of DIF.) Scores in terms of total raw score and θ estimates are affected when there is a certain number of DIF items in the test form; the impact on total scores is more substantial for 2PL responses than for 3PL responses. The difference between groups or forms increases as the number of DIF items increases. The effect for two groups receiving DIF forms is larger than the effect of DIF for either group receiving both forms under each of the conditions. The greatest group difference is less than 2 points on the 0–60 total score scale and .15 on the θ scale. DIF does not appear to distort test reliability.

To see the extent of 2 points on the 0–60 total score scale, consider the following case. The raw (total) score for SAT® Mathematics is roughly on a 0–60 scale. For a raw score difference between .5 and 1, the scale score difference is at most 10 points on the 200–800 scale; a raw score difference between 1 and 2 corresponds to at most 20 points in the middle of the raw score range and at most 40 points at the boundaries (i.e., around 0 or 60). If the same raw-to-scale conversion can be applied to our simulated data, a more evident scale score difference between examinee groups is observable in a test form when more than 10% of the items are DIF items.

Under our design, the DIF forms consist of both moderate (Category B) and large (Category C) DIF items, and at most 20% of the items function differentially in the reference groups and the focal groups by design. One can expand the design in many respects, but some phenomena can already be anticipated based on the magnitude and patterns of the effects observed herein. For example, raising the percentage of DIF items should yield a more significant advantage for the reference group and a disadvantage for the focal group on ability estimation. If the embedded DIF items belong strictly to Category C, the group and form differences will be stronger. When the DIF items in one form favor different groups, or if they display crossing DIF rather than uniform DIF, the impact will be less significant than the results reported in this report.

It is possible to take an analytical approach to examining the consequences of DIF. Item parameters, amount of DIF, examinee abilities, and other factors of interest can be systematically manipulated to understand what is going on with respect to mean score differences and reliability. Take the Rasch model, for example: The only item parameter that varies across items is the item difficulty, and different levels of DIF (quantified by Mantel-Haenszel statistics) can be converted into differences in item difficulties. So some results for the effect of DIF may be derived analytically. Similar techniques may be extended to the 2PL case, which differs from the Rasch case only in the varying discrimination parameters. Further investigation is needed to determine if the analytical approach can be successful.

References

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43–68.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355–368.

ETS. (2009). *ETS fairness review guidelines.* Princeton, NJ: ETS.

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

Kulick, E. (2000). *NDIF/NAEP user's guide* [Computer software]. Princeton, NJ: Author.

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008). *A review of recent developments in differential item functioning* (ETS Statistical Research Rep. No. 08-43). Princeton, NJ: ETS.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C. R. Rao (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 125–167). Amsterdam: North-Holland.

Roever, C. (2005, September 15). *"That's not fair!" Fairness, bias, and differential item functioning in language testing.* Retrieved from http://www2.hawaii.edu/ roever/brownbag.pdf

Shealy, R. T., & Stout, W. F. (1993a). An item response model for test bias and differential test functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–239). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shealy, R. T., & Stout, W. F. (1993b). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 54*, 159–194.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Lawrence Erlbaum Associates.