Differential Item Functioning Detection Across Two Methods of Defining Group Comparisons: Pairwise and Composite Group Comparisons

Educational and Psychological
Measurement

2015, Vol. 75(4) 648–676

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164414549764

epm.sagepub.com

Halil Ibrahim Sari¹ and Anne Corinne Huggins¹

Abstract

This study compares two methods of defining groups for the detection of differential item functioning (DIF): (a) pairwise comparisons and (b) composite group comparisons. We aim to emphasize and empirically support the notion that the choice of pairwise versus composite group definitions in DIF is a reflection of how one defines fairness in DIF studies. In this study, a simulation was conducted based on data from a 60-item ACT Mathematics test (ACT; Hanson & Béguin). The unsigned area measure method (Raju) was used as the DIF detection method. An application to operational data was also completed in the study, as well as a comparison of observed Type I error rates and false discovery rates across the two methods of defining groups. Results indicate that the amount of flagged DIF or interpretations about DIF in all conditions were not the same across the two methods, and there may be some benefits to using composite group approaches. The results are discussed in connection to differing definitions of fairness. Recommendations for practice are made.

Keywords

differential item functioning, pairwise, composite group, fairness

¹University of Florida, Gainesville, FL, USA

Corresponding Author:

Anne Corinne Huggins, College of Education, University of Florida, P.O. Box 117049, Gainesville, FL 32611-7049, USA.

Email: ahuggins@coe.ufl.edu

Goal and Importance of the Study

Differential item functioning (DIF) methods compare item statistics across individuals with the same latent ability levels but belonging to different groups to make a conditional comparison of the performance of the groups (Holland & Wainer, 1993; Osterlind & Everson, 2009; Penfield & Camilli, 2007). Comparisons across two groups are often done by comparing the groups directly to each other, which can be called a pairwise comparison. In this approach, if a variable categorizes examinees into three groups (e.g., Grades 6-8), two groups are selected as the focal groups (e.g., Grades 6 and 7), one group is selected as the reference group (e.g., Grade 8), and then each focal group is directly and independently compared with the reference group (e.g., Grade 6 compared with Grade 8, and Grade 7 compared with Grade 8). Although the pairwise approach is a typical way to examine DIF and extensively used, no studies to date have inquired whether or not this approach measures fairness as defined by the *Standards* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). Group comparison choices in DIF have never been directly related to fairness as a lack of bias in the *Standards* (1999), and it may be that the pairwise approach does not always provide us with information about a true lack of fairness.

A less common approach of composite group comparisons was introduced by Ellis and Kimmel (1992). In this approach, each of the groups is considered a focal group (e.g., Grade 6, Grade 7, or Grade 8) and a composite group that includes all examinees is used as the reference group. Each focal group is compared with this composite group (e.g., Grade 6 compared with composite, Grade 7 compared with composite, and Grade 8 compared with composite). It appears that Ellis and Kimmel (1992) have been the only researchers who have applied this type of DIF analysis in a published research study. This is despite the fact that a composite group approach is commonly used when defining groups in a score equity assessment of equating (Dorans & Holland, 2000) and that there may be several benefits associated with using this composite group approach in DIF studies.

Researchers and practitioners tend to use pairwise comparisons rather than composite group comparisons in DIF analysis, but this study questions that practice for three main reasons: (a) pairwise and composite group approaches address two different types of fairness and yet this has not been discussed in the literature and, in practice, the approaches are rarely connected to fairness definitions prior to implementation; (b) the two approaches may result in different conclusions drawn about the presence of DIF and yet those different conclusions have never been demonstrated through a comparative analysis; and (c) pairwise and composite group approaches have different relative advantages that have not been discussed in the literature on DIF. This study addresses these three concerns and in doing so provides researchers and practitioners with a framework for choosing the most appropriate method for defining groups in their DIF analyses.

We aim to emphasize and empirically support the notion that the choice of pairwise versus composite group definitions in DIF analysis is important. It reflects how

we define fairness and influences the interpretations that will be made from the DIF analysis, making the choice worthy of careful consideration prior to investigating DIF.

Introduction to Item Response Theory and DIF Detection With Area Measures

While there are several models for dichotomous (i.e., binary) test data, the two-parameter logistic (2PL) model (Birnbaum, 1968) is the focus of this study and defines the conditional probability of a correct response to an item as

$$P(X_j = 1|\theta) = \frac{\exp[a_j(\theta - b_j)]}{1 + \exp[a_j(\theta - b_j)]}, \quad (1)$$

where $P(X_j = 1|\theta)$ is the probability of a correct response on item j , θ is the latent trait, a_j is an item discrimination parameter, and b_j is an item difficulty parameter.

Area measure methods estimate DIF as the area between two item characteristic curves (ICCs), one for each of the two groups being compared (Raju, 1988). Raju (1988) defined the unsigned area (UA) as

$$UA = \int_{-\infty}^{\infty} |P(X = 1|\theta, G = R) - P(X = 1|\theta, G = F)|d\theta, \quad (2)$$

where G equals group, R is the reference group, and F is the focal group.

The magnitude of DIF serves as an effect size and is defined as the area between the two ICCs. In addition to computing the size of the UA , one can determine whether this area is statistically significant different from zero, by dividing the area estimate by its standard error to obtain z -statistics.

When detecting DIF with area measures in a pairwise comparison, group-level ICCs are compared directly with each other via Equation (2). For example, an ICC for females and an ICC for males are directly compared. In a composite group comparison, ICCs calibrated from each group of participants are compared with ICCs calibrated using the composite group, which are the ICCs used in operational practice. These definitions show a significant difference between pairwise and composite comparisons. The ICCs being compared depend on the method chosen, and so results of the methods should differ and, therefore, there may be some benefits to using one method over another.

Definition of Fairness in Pairwise and Composite Group Comparisons

The main goal of both pairwise and composite group DIF comparisons is to examine test items for fairness as a lack bias. However, it can be argued that the two

Figure 1. Differential item functioning under pairwise and composite group comparisons for two- and three-group conditions.

Note. ICC = item characteristic curve.

approaches are evaluating different types of lack of bias that align with different definitions of fairness. As can be understood from Figure 1, in pairwise comparisons fairness is achieved if the ICC for one group on an item is the same relative to other groups. In composite group comparisons, fairness is achieved if the ICC for one group on an item is the same relative to the function used in operational practice. Within this framework, a question is begged: “How do we define fairness as a lack of item bias?”

The *Standards* (AERA, APA, & NCME, 1999) provides definitions of fairness that are to permeate the field of educational measurement and drive decisions about how fairness is defined. The *Standards* (AERA, APA, & NCME, 1999) states, “[Fairness as lack of bias] is said to arise when deficiencies in a test itself or the manner in which it is used result in different meanings for scores earned by members of different identifiable subgroups” (p. 74). According to this definition, it can be argued that we are concerned with the scores that students “earn” based on the operational ICCs. This aligns with the definition of composite group comparisons. Furthermore, the *Standards* state that many decisions, especially in high-stake tests, such as pass/fail or admit/reject, are based on the full population of examinees taking the test (AERA, APA, & NCME, 1999). These statements about fairness suggest that student success is determined by test scores resulting from calibrations that include all examinees, and therefore issues of fairness as a lack of bias should relate to test responses that are compared with a composite group rather than to a single, reference group of individuals. Thus, the language of the *Standards* supports the composite group approach to DIF analysis, even if it does not directly refute the pairwise approach.

Potential Advantages to Using a Composite Group Approach in DIF Studies

There are several potential advantages to using composite group approaches over pairwise approaches. First, composite group comparisons can more easily allow for a fine-tuned definition of groups based on more than one grouping variable (as mentioned by J. Liu & Dorans, 2013). For example, if it is found that Hispanic students are disadvantaged by some items, this disadvantage may not hold over different gender groups. One could easily compare Hispanic females with the composite group of all examinees, yet in a pairwise comparison one is left wondering who an appropriate reference group is for Hispanic females. Second, it can be argued that it is problematic to define one particular group as a reference to which all other groups are to be compared (APA, 2009). For example, choosing Caucasian examinees as a reference for all other examinees has an underlying value statement about fairness that is not always readily supported outside a particular person's value perspective. Composite group comparisons overcome this problem as the reference group always consists of all examinees rather than a chosen group. Third, pairwise comparisons ignore operational item parameters even though examinees receive test scores based on these critical parameters. Composite group comparisons overcome this problem. Fourth, the composite group approach allows for a separate DIF estimate for each group. For example, we can talk about fairness as a lack of bias for Grade 6 examinees without having to refer to Grade 7, Grade 8, or any other reference group. This makes it easier on practitioners to determine which groups might have bias problems in their reported scores. In pairwise comparisons, particularly when there are four or more groups, one has to look through many pairs of results to try to figure out the overall nature of group differences. Not only can it be difficult to determine this nature, but one is also left with several results for each group. Of course one can select a reference group to minimize the comparisons, but the sacrifice is that the overall nature of group differences in the item is lost because all differences are relative to a single reference group. When using composite group approaches a single DIF effect is estimated for each group and directly answers the question: "Is the group different from the overall item parameters used for estimating reported scores?" Fifth, running multiple DIF tests results in alpha inflation if adjustments are not made (Penfield, 2001). When a variable groups examinees into four or more groups, the number of pairwise comparisons needed to complete a DIF analysis on this variable is greater than the number of composite group comparisons. Although composite group comparisons cannot overcome the problem of alpha inflation, they can reduce it relative to pairwise comparisons. The amount of relative reduction in alpha inflation increases as the number of groups being compared increases. We examine observed Type I error rates in our simulation study, as well as the related notion of false discovery rates (FDRs), in order to determine the extent to which inflated alphas and other related concerns have an impact on applications of pairwise and composite group DIF methods.

Potential Disadvantages to Using a Composite Group Approach in DIF Studies

There are some disadvantages of composite group comparisons that can be overcome by using pairwise comparisons. First, pairwise comparisons compare independent groups whereas composite group comparisons have a lack of independence between the groups (Ellis & Kimmel, 1992). Second, there are more DIF detection methods that have been developed for pairwise comparisons, such as Mantel–Haenszel (Holland & Thayer, 1988; Mantel & Haenszel, 1959), Lord's (1980) chi-square, and logistic regression (Swaminathan & Rogers, 1990; Zumbo, 1999). Second, sample sizes of the groups being compared play a major role in composite group comparisons when the DIF indices are not weighted. For example, if the composite group consists of students with severe disabilities (5% of composite group), students with mild disabilities (10% of composite group), and students without disabilities (85% of composite group), then the composite group ICC will be very close to the ICC of the group of students without disabilities. This group will be shown as having no DIF effect by any unweighted index whereas the other groups may or may not display a DIF effect. This is a problem with several indices used in score equity assessment (i.e., examinations of population invariance of equating) as composite group comparisons are often used in those applications of fairness examinations (Dorans, 2004; M. Liu & Holland, 2008). This sample size concern can ultimately be seen as reflecting reality; the operational ICCs used to develop reported scores are based more heavily on groups that constitute a larger portion of the total population of examinees. So when focusing on the question of whether or not a group is different from the ICC used in operational practice (as composite group DIF methods do), the sample size issue mentioned here can be seen as less of a problem and more of a reflection of how reported test scores are derived in practice.

Simulation Study Method

The four research questions of this study are

Research Question 1: Does the nature of true b -parameter differences between groups (i.e., all groups are different from each other vs. a single group is different from all other groups) differentially affect the ability of pairwise and composite group comparisons to detect DIF?

Research Question 2: Does the magnitude of true b -parameter differences between groups differentially affect the ability of pairwise and composite group comparisons to detect DIF?

Research Question 3: Does the number of groups in a DIF analysis differentially affect the ability of pairwise and composite group comparisons to detect DIF?

Research Question 4: Are the two methods of defining groups associated with different rates of observed Type I errors and false discovery?

The 2PL model (Birnbaum, 1968) was used for data generation in R Version 2.15.1 (R Development Core Team, 2013). The item parameters used in this simulation study were based on estimated item parameters from the 1997 ACT Mathematics test (Hanson & Béguin, 2002). Aligned with Hanson and Béguin (2002), the distributions of the true item parameters of 60 dichotomous items were used to generate item response data for 60 dichotomous items used in this study. Specifically, the difficulty parameters were selected from a distribution of $N(0.11, 1.11)$. The discrimination parameters were generated from a random uniform distribution and ranged from $\min(a) = 0.42$ to $\max(a) = 1.88$. True ability parameters (θ) were randomly sampled from a normal distribution of $N(0, 1)$, and 100 replications were performed for each simulation condition.

Three Factors That Defined Study Design Conditions

Number of Groups. Each data set was generated with respect to three, four, or five groups that were to be compared in the DIF analysis. A sample size of 500 examinees was created within all groups, which is adequate sample size for estimating the 2PL model (Birnbaum, 1968) and for ensuring appropriate power for UA DIF methods (see, Holland & Wainer, 1993; Kim & Cohen, 1995).

Magnitude of True b -Parameter Differences. Each condition of the study had a single test item in which either small, moderate, or large true b -parameter differences were introduced into the true group-level item difficulty parameters. The size of true b -parameter differences was closely determined according to the ETS classifications of pairwise DIF effects (Zieky, 1993; Zwick, 2012). This classification places items into three categories (i.e., A items, B items, and C items). Based on the ETS classification scheme, the magnitude of differences in b parameters across the reference and the focal groups are defined as follows: $|(b_F - b_R)| < 0.43$ represents small DIF (A items), $0.43 \leq |(b_F - b_R)| \leq 0.64$ represents moderate DIF (B items), and $|(b_F - b_R)| > .64$ represents large DIF (C items).

However, not all researchers follow this guidance when specifying magnitudes of b -parameter differences. For example, Shepard, Camilli, and Williams (1985) used a difference of .20 and .35 in the b parameter to manipulate small and moderate group parameter differences, respectively. Moreover, Hidalgo and Lopez-Pina (2004) used a difference of .3, .6, and 1.00 in the b parameters to represent moderate and large group parameter differences, respectively. In this study, a difference of either $b_F - b_R = 0.3$, $b_F - b_R = 0.6$, or $b_F - b_R = 0.9$ was introduced between b parameters to represent small, moderate, and large group parameter differences, respectively.

Furthermore, the magnitude of true b -parameter differences was manipulated so as to be the maximum amount of b -parameter differences between any of the pairs of 3, 4, or 5 groups. Thus, the magnitude of the true b -parameter differences is more aptly stated as defining conditions of small or less DIF, moderate or less DIF, or large or

less DIF between any two pairs of groups. One can refer to Table 1 to better understand this definition of magnitude of DIF.

Nature of Group Differences in b Parameters. The left side of Figure 1 visually displays group and composite ICCs. As can be inferred from the left side of Figure 1, the amount of flagged DIF is expected to be smaller in some composite group comparisons than in some pairwise comparisons. However, this will not always be the case. The right side of Figure 1 displays an example of a type of composite and pairwise comparison in which the relative magnitude of the comparisons would vary. In other words, when there are more than two groups, the nature of the group differences in b parameters is not always consistent. Thus, to take this situation into account two factors were created and named as “all groups differ in b parameters” and “one group differs in b parameter” (see Table 1). For all data sets classified as “all groups differ in b parameters” the magnitudes of small, moderate, or large true b -parameter differences were spread among the groups. For all data sets classified as “one group differs in b parameters,” the magnitude of small, moderate, or large b -parameter differences were only added to the last group, and the remaining groups were specified as having the same true b parameters.

Data Analysis

The test data were analyzed with the 2PL model (Birnbaum, 1968). Raju's (1988) UA method was used for all DIF analyses under all possible pairwise and composite group defining approaches. All model and DIF estimation was completed with the difR package (Magis, Beland, Tuerlinckx, & De Boeck, 2010) in R version 2.15 (R Development Core Team, 2013).

Each research question in the study focuses on comparing the effect size and statistical significance of DIF detected between the two approaches of group definitions. We followed Jodoin and Gierl (2001) recommendation for the UA measure effect sizes to be able to categorize DIF effects as small, moderate, or large. Based on their classification scheme, the magnitude of differences in DIF effects sizes across the reference and the focal groups are defined as follows: $UA \leq .4$ represents small DIF, $.4 < UA \leq .6$ represents moderate DIF, and $UA > .6$ represents large DIF. Percentage of DIF was recorded as the number of times (out of 100 iterations) that an item exhibited statistically significant DIF at $\alpha = .01$ (corresponding z value is ± 2.576). Both the average effect size for each condition and the percentage of statistically significant DIF for each condition were compared across the pairwise and composite group approaches.

Simulation Study Results

Results of the Conditions Classified as All Groups Differ in b Parameters

Results of effect sizes and the statistical significance when all groups differ in b parameters are provided in Table 2. Based on the three group results under the small true

Table 1. True Item Difficulty Parameters Across the Groups.

All groups differ in b parameters				One group differs in b parameters		
	Small	Moderate	Large	Small	Moderate	Large
Three groups	G1	$b_1 = b^* - .15$	$b_1 = b^* - .3$	$b_1 = b^* - .45$	$b_1 = b^*$	$b_1 = b^*$
	G2	$b_2 = b^*$	$b_2 = b^*$	$b_2 = b^*$	$b_2 = b^*$	$b_2 = b^*$
	G3	$b_3 = b^* + .15$	$b_3 = b^* + .3$	$b_3 = b^* + .45$	$b_3 = b^* + .3$	$b_3 = b^* + .6$
Four groups	G1	$b_1 = b^* - .15$	$b_1 = b^* - .3$	$b_1 = b^* - .45$	$b_1 = b^*$	$b_1 = b^*$
	G2	$b_2 = b^* - .05$	$b_2 = b^* - .1$	$b_2 = b^* - .15$	$b_2 = b^*$	$b_2 = b^*$
	G3	$b_3 = b^* + .05$	$b_3 = b^* + .1$	$b_3 = b^* + .15$	$b_3 = b^*$	$b_3 = b^*$
Five groups	G4	$b_4 = b^* + .15$	$b_4 = b^* + .3$	$b_4 = b^* + .45$	$b_4 = b^* + .3$	$b_4 = b^* + .6$
	G1	$b_1 = b^* - .15$	$b_1 = b^* - .3$	$b_1 = b^* - .45$	$b_1 = b^*$	$b_1 = b^*$
	G2	$b_2 = b^* - .075$	$b_2 = b^* - .15$	$b_2 = b^* - .225$	$b_2 = b^*$	$b_2 = b^*$
	G3	$b_3 = b^*$	$b_3 = b^*$	$b_3 = b^*$	$b_3 = b^*$	$b_3 = b^*$
	G4	$b_4 = b^* + .075$	$b_4 = b^* + .15$	$b_4 = b^* + .225$	$b_4 = b^*$	$b_4 = b^*$
G5	$b_5 = b^* + .15$	$b_5 = b^* + .3$	$b_5 = b^* + .45$	$b_5 = b^* + .3$	$b_5 = b^* + .6$	$b_5 = b^* + .9$

Note. b^* = the true item difficulty parameter that was sampled for the particular condition.

Table 2. Results of the All Conditions Classified as All Groups Differ in *b* Parameters.

	Comparison (pairwise or composite)	Small <i>b</i> -parameter difference		Moderate <i>b</i> -parameter difference		Large <i>b</i> -parameter difference	
		Effect size	Percentage of statistical significance	Effect size	Percentage of statistical significance	Effect size	Percentage of statistical significance
Three groups	G1 vs. G2	0.26	7	0.36	35	0.50	100
	G1 vs. G3	0.35	32	0.61	100	0.99	100
	G2 vs. G3	0.25	5	0.38	39	0.48	100
	G1 vs. C	0.18	23	0.34	46	0.43	60
	G2 vs. C	0.09	1	0.10	1	0.10	2
Four groups	G3 vs. C	0.18	24	0.35	48	0.42	59
	G1 vs. G2	0.15	4	0.20	24	0.33	35
	G1 vs. G3	0.26	9	0.30	30	0.64	100
	G1 vs. G4	0.34	32	0.61	100	0.99	100
	G2 vs. G3	0.14	3	0.22	30	0.32	32
	G2 vs. G4	0.24	12	0.32	35	0.65	100
	G3 vs. G4	0.15	7	0.22	26	0.34	38
	G1 vs. C	0.17	20	0.32	55	0.48	100
	G2 vs. C	0.09	2	0.11	7	0.18	37
	G3 vs. C	0.08	2	0.11	5	0.17	33
Five groups	G4 vs. C	0.17	19	0.31	51	0.47	100
	G1 vs. G2	0.12	2	0.19	4	0.26	8
	G1 vs. G3	0.22	4	0.40	36	0.50	100
	G1 vs. G4	0.26	4	0.50	67	0.77	100

(continued)

Table 2. (continued)

Comparison (pairwise or composite)	Small <i>b</i> -parameter difference		Moderate <i>b</i> -parameter difference		Large <i>b</i> -parameter difference	
	Effect size	Percentage of statistical significance	Effect size	Percentage of statistical significance	Effect size	Percentage of statistical significance
G1 vs. G5	0.34	30	0.60	100	1.01	100
G2 vs. G3	0.14	3	0.20	6	0.26	9
G2 vs. G4	0.21	3	0.39	41	0.52	100
G2 vs. G5	0.27	5	0.54	72	0.76	100
G3 vs. G4	0.13	8	0.19	6	0.26	8
G3 vs. G5	0.20	4	0.38	35	0.52	100
G4 vs. G5	0.13	5	0.19	8	0.25	8
G1 vs. C	0.17	37	0.34	51	0.51	100
G2 vs. C	0.11	7	0.19	26	0.26	29
G3 vs. C	0.08	3	0.09	2	0.10	7
G4 vs. C	0.10	8	0.18	24	0.26	30
G5 vs. C	0.18	40	0.35	50	0.50	100

b-parameter difference condition, both pairwise and composite group comparisons indicated that each comparison displayed a small or negligible amount of DIF. As the true *b*-parameter differences increased to moderate in size, the pairwise comparison of Group 1 with Group 3 was large in size and all other pairwise and composite group comparisons were small. As the true *b*-parameter differences increased to large in size, all pairwise comparisons showed DIF effect sizes that were moderate or large whereas composite group comparisons consistently displayed that only Groups 1 and 3 had moderate DIF effects but Group 2 had a small, negligible DIF effect. Percentage of statistical significance aligned with these effect size findings.

Based on the four-group results under the condition of small true *b*-parameter differences, both pairwise and composite group comparisons found similar DIF effects that were always small. However, as the true *b*-parameter differences increased to moderate, the pairwise approach showed the pairwise comparison of Group 1 with Group 4 was large in size and all other pairwise comparisons were small. When the differences in true *b* parameter were large, three pairwise comparisons that involved some of the combinations of Groups 1 to 4 were flagged as having large DIF effects. On the other hand, composite group comparisons always showed all groups as having small, negligible DIF effects when true *b*-parameter difference were small or moderate. When true *b*-parameter differences were large, two composite comparisons had moderate effect sizes. Percentage of statistical significance aligned with these effect size findings.

Based on the five-group results, under the conditions of small true *b*-parameter differences, both pairwise and composite group comparisons showed similar DIF effects that were always small or negligible DIF, and percentage of statistical significances aligned with these effect size findings. However, as the true *b*-parameter differences increased to moderate, pairwise comparisons displayed four comparisons as having moderate DIF effects, in which all groups were present in at least one of the comparisons. Similarly, as the true *b*-parameter differences increased to large magnitudes, the pairwise approach showed six DIF effect sizes of moderate or large magnitude, and again each group was involved in at least one of these comparisons. However, under the condition of moderate and large true *b*-parameter differences, composite group comparisons always showed Groups 1 and 5 as having moderate DIF effect sizes. Other groups were never flagged.

Results of the All Conditions Classified as One Group Differs in b Parameters

Results of effect sizes and statistical significance for all conditions where only one group differed in true *b* parameters are provided in Table 3. Based on the three-group results under all true *b*-parameter magnitudes, pairwise comparisons consistently indicated that Group 3 had the expected small, moderate, or large DIF effects relative to other groups. Results pertaining to the percentage of statistical significance aligned with the effect size results. Recall that in these conditions, the last group's *b* parameter was specified as variant to the other groups' *b* parameters, which were

Table 3. Results of the All Conditions Classified as One Group Differs in *b* Parameters.

	Comparison (pairwise or composite)	Small <i>b</i> -parameter difference		Moderate <i>b</i> -parameter difference		Large <i>b</i> -parameter difference	
		Effect size	Percentage of statistical significance	Effect size	Percentage of statistical significance	Effect size	Percentage of statistical significance
Three groups	G1 vs. G2	0.09	2	0.10	2	0.10	1
	G1 vs. G3	0.35	34	0.68	100	1.00	100
	G2 vs. G3	0.35	36	0.67	100	0.99	100
	G1 vs. C	0.12	5	0.23	28	0.36	77
	G2 vs. C	0.12	4	0.22	25	0.37	74
Four groups	G3 vs. C	0.12	4	0.44	100	0.62	100
	G1 vs. G2	0.13	2	0.10	2	0.11	3
	G1 vs. G3	0.15	1	0.09	2	0.10	2
	G1 vs. G4	0.32	38	0.63	100	0.95	100
	G2 vs. G3	0.15	2	0.15	3	0.09	1
	G2 vs. G4	0.33	41	0.64	100	0.96	100
	G3 vs. G4	0.33	39	0.63	100	0.95	100
	G1 vs. C	0.11	4	0.16	7	0.24	29
	G2 vs. C	0.12	5	0.18	8	0.23	24
	G3 vs. C	0.12	3	0.17	7	0.25	29
Five groups	G4 vs. C	0.20	10	0.45	70	0.65	100
	G1 vs. G2	0.09	1	0.10	4	0.10	2
	G1 vs. G3	0.10	3	0.11	5	0.09	1
	G1 vs. G4	0.09	2	0.09	4	0.09	1

(continued)

Table 3. (continued)

Comparison (pairwise or composite)	Small <i>b</i> -parameter difference		Moderate <i>b</i> -parameter difference		Large <i>b</i> -parameter difference	
	Effect size	Percentage of statistical significance	Effect size	Percentage of statistical significance	Effect size	Percentage of statistical significance
G1 vs. G5	0.33	42	0.66	100	0.97	100
G2 vs. G3	0.09	2	0.10	4	0.09	2
G2 vs. G4	0.10	3	0.10	3	0.10	3
G2 vs. G5	0.34	45	0.66	100	0.98	100
G3 vs. G4	0.10	2	0.10	5	0.09	2
G3 vs. G5	0.33	41	0.66	100	0.96	100
G4 vs. G5	0.34	42	0.65	100	0.97	100
G1 vs. C	0.17	15	0.20	22	0.34	43
G2 vs. C	0.18	17	0.21	24	0.36	46
G3 vs. C	0.17	17	0.22	26	0.35	46
G4 vs. C	0.16	15	0.21	21	0.34	44
G5 vs. C	0.25	30	0.52	96	1.67	100

	<u>Fail to Reject Null</u>	<u>Reject Null</u>	<u>Total</u>
True Nulls	A	B	x_0
False Nulls	C	D	$x - x_0$
Total	$x - E$	E	x

Figure 2. Statistical conclusions when testing x null hypotheses.

invariant to each other. Under the small, moderate, and large true b -parameter difference conditions, the highest composite group DIF effects were always found in comparisons of Group 3 with the composite group.

Based on the four-group results, under the small true b -parameter difference condition both pairwise and composite group comparisons always displayed small or negligible DIF effects. However, when the true b -parameter differences increased to moderate or large, the pairwise comparisons consistently found large DIF effects in comparisons that involved Group 4 and more data sets showed statistically significant effects for this group. Similarly, under the moderate and large true b -parameter difference conditions, composite group comparisons always found moderate or large DIF effects in only comparisons of Group 4 with the composite group and more data sets showed a statistically significant effect for this group. Results for the five-group case showed the same pattern as the four-group case.

Type I Error and False Discovery Rates Study Method

Observed Type I error and FDRs are of interest in many recent DIF studies (see, Cauffman & MacIntosh, 2006; Thissen, Steinberg, & Kuang, 2002; Woods, 2009). In Figure 2, A , B , C , and D are unobserved random variables, E is an observed random variable, and x is the number of all hypotheses. As discussed in Benjamini and Hochberg (1995), observed Type I error rate is defined as the proportion of falsely rejected null hypotheses. Using notation from Figure 2, this is calculated as B/x for all items that have a true null condition (i.e., no DIF). Observed FDR is defined as the proportion of erroneously rejected null hypotheses, or Q . Using notation from Figure 2, the observed FDR can be calculated as $Q = B/E$. In our study, we calculated both Type I errors and false discoveries for the large DIF condition in data sets classified as “all groups differ in b parameters” (i.e., column 5 of Table 1). We recalibrated said condition with 1,000 iterations.

For Type I error rate, we calculated the total observed Type I errors for 59 non-DIF items within each of the three-group, four-group, and five-group cases. Within

each group case, there were multiple pairwise comparisons. For example, within the four-group case there were six pairwise comparisons. We collapsed across all comparisons to obtain, for example, a single observed Type I error rate for pairwise comparisons in the four-group case. The same was true with composite comparisons. For example, the five-group case has five composite group comparisons and we collapsed across them to obtain a single observed Type I error rate for the five-group composite comparison case. We collapsed comparisons within pairwise and composite approaches because we wanted to obtain the cumulative observed Type I error rate for each non-DIF item when one must complete multiple comparisons to fully examine that item.

The aforementioned Type I error rate calculations for pairwise assumed that one would examine *all* possible pairwise comparisons (e.g., in the three-group case, comparing Group 1 with Group 2, Group 1 with Group 3, and Group 2 with Group 3). However as mentioned previously, practitioners often do not look at all possible comparisons but rather they select a single reference group (e.g., in the three-group case, selecting Group 3 as a reference and comparing Group 1 with Group 3 and Group 2 with Group 3). This would be associated with a different observed Type I error rate as there are a different number of comparisons. In Table 4, we display the observed Type I error rates for this type of “One Reference Group” pairwise approach.

When calculating the FDRs (Benjamini & Hochberg, 1995), we had to use not only data from the 59 non-DIF items but also data from the single DIF item. It was critical to control for the magnitude of DIF in this single DIF item. Therefore, we only looked at certain groups within the large DIF condition in data sets classified as “all groups differ in b parameters” (i.e., Column 5 of Table 1). Specifically, for the pairwise portion of the three-group case we examined Group 1 with Group 2 and Group 2 with Group 3 (omitting the comparison of Group 1 with Group 3), and for the composite portion of the three-group case we examined Group 1 with composite and Group 3 with composite. These selected pairwise and composite comparisons had the same magnitude of true b -parameter differences, thereby controlling for effect size in the calculation of FDRs. A similar process was completed for the five-group case, but the four-group case was omitted as the nature of true b -parameter differences did not allow for appropriate effect size control.

Each of the 59 non-DIF items has an FDR that was calculated such that B came from the non-DIF item and D came from the DIF item (i.e., Item 1). For example, Item 7’s FDR was calculated such that B came from Item 7 results and D came from Item 1 results. Each of the 59 non-DIF items had multiple FDRs associated with it within pairwise, due to the multiple pairwise comparisons within that item. We selected the maximum FDR from all those comparisons, providing a single pairwise FDR for each non-DIF item. The same maximum selection process was used for composite group comparisons, providing a single composite FDR for each non-DIF item. Selecting the maximum ensured that we were examining results that were the worst FDR situation possible.

Table 4. Collapsed Type I Error Rates for Non-DIF Items.

Item	All possible comparisons						One reference group		
	3 G		4 G		5 G		3 G	4 G	5 G
	P	C	P	C	P	C	P	P	P
2	.022	.009	.050	.045	.067	.021	.013	.047	.032
3	.018	.008	.116	.023	.217	.038	.010	.053	.086
4	.066	.014	.126	.040	.057	.033	.048	.065	.026
5	.027	.009	.136	.038	.298	.032	.017	.076	.117
6	.029	.009	.038	.037	.258	.032	.023	.007	.097
7	.060	.016	.064	.017	.287	.031	.042	.034	.103
8	.057	.017	.149	.030	.161	.033	.031	.078	.006
9	.022	.007	.092	.023	.128	.036	.014	.041	.045
10	.037	.007	.103	.020	.124	.040	.026	.050	.062
11	.074	.011	.111	.048	.195	.031	.051	.049	.074
12	.055	.011	.041	.036	.084	.027	.031	.042	.035
13	.043	.013	.073	.049	.286	.024	.027	.070	.113
14	.063	.014	.132	.036	.321	.032	.036	.072	.133
15	.070	.015	.089	.032	.097	.036	.041	.046	.059
16	.058	.011	.058	.038	.298	.034	.038	.035	.125
17	.070	.012	.094	.033	.227	.038	.045	.047	.082
18	.017	.012	.093	.038	.161	.032	.013	.051	.069
19	.079	.015	.138	.056	.227	.022	.058	.073	.090
20	.019	.009	.112	.037	.182	.026	.012	.047	.050
21	.040	.012	.094	.035	.116	.049	.022	.046	.098
22	.035	.016	.076	.052	.096	.037	.019	.058	.059
23	.045	.012	.071	.033	.043	.032	.029	.040	.020
24	.086	.016	.091	.040	.255	.031	.056	.047	.102
25	.018	.007	.147	.037	.158	.044	.012	.065	.047
26	.055	.010	.119	.026	.075	.034	.036	.058	.042
27	.071	.017	.081	.043	.148	.028	.040	.058	.029
28	.042	.012	.041	.048	.041	.029	.026	.050	.044
29	.073	.011	.147	.021	.137	.023	.046	.075	.015
30	.014	.007	.136	.029	.165	.021	.007	.060	.026
31	.014	.011	.112	.038	.151	.034	.006	.052	.021
32	.068	.009	.067	.044	.067	.028	.013	.054	.015
33	.012	.013	.027	.063	.202	.021	.010	.009	.083
34	.029	.009	.049	.024	.296	.037	.019	.027	.122
35	.066	.013	.085	.031	.214	.022	.020	.044	.093
36	.012	.008	.105	.047	.176	.033	.005	.049	.029
37	.006	.004	.082	.044	.262	.043	.005	.047	.109
38	.042	.010	.093	.044	.230	.043	.041	.060	.087
39	.065	.017	.112	.044	.162	.035	.048	.049	.061
40	.067	.011	.092	.039	.296	.023	.044	.059	.119
41	.052	.014	.099	.036	.208	.039	.031	.041	.086
42	.054	.012	.087	.031	.270	.030	.034	.055	.110
43	.051	.010	.096	.033	.256	.033	.029	.049	.107
44	.077	.012	.079	.021	.180	.034	.047	.018	.070
45	.011	.007	.055	.019	.114	.039	.009	.040	.040
46	.057	.012	.123	.050	.250	.031	.040	.058	.104
47	.028	.010	.084	.044	.200	.020	.020	.048	.080

(continued)

Table 4. (continued)

Item	All possible comparisons						One reference group		
	3 G		4 G		5 G		3 G	4 G	5 G
	P	C	P	C	P	C	P	P	P
48	.065	.015	.124	.026	.178	.039	.050	.062	.016
49	.041	.006	.099	.038	.280	.045	.032	.065	.127
50	.034	.014	.101	.031	.180	.028	.020	.041	.018
51	.025	.009	.054	.026	.281	.032	.013	.031	.133
52	.031	.010	.040	.045	.059	.043	.024	.041	.011
53	.025	.011	.055	.044	.074	.024	.014	.052	.041
54	.047	.009	.100	.046	.148	.029	.034	.049	.053
55	.023	.008	.112	.048	.283	.026	.017	.066	.116
56	.076	.013	.110	.015	.196	.039	.046	.046	.080
57	.033	.009	.080	.034	.088	.032	.022	.051	.049
58	.057	.012	.067	.051	.289	.023	.031	.061	.111
59	.025	.008	.038	.036	.322	.028	.011	.037	.141
60	.027	.010	.049	.035	.099	.013	.021	.022	.029

Note. P = pairwise; C = composite; G = groups; DIF = differential item functioning. All bold and italic numbers represent unique results that are discussed in text.

Results of Type I Error and False Discovery Rates

Results of Type I errors for 59 non-DIF items across pairwise and composite group comparisons are given in Table 4. All bold and italic numbers represent unique results that are discussed in the remainder of this paragraph. First, we will discuss results when pairwise comparisons are based on all possible comparisons (left side of Table 4). In the three-group case, all items except Item 33 displayed a higher Type I error rate for pairwise as compared with composite comparisons. In the four-group case, all items except Item 52 displayed a higher Type I error rate for pairwise as compared with composite comparisons. In the five-group case, all items without exception displayed a higher Type I error rate for pairwise as compared with composite comparisons. Next, we compared the pairwise and composite methods using pairwise results that originated from the utilization of only one reference group (right side of Table 4). In the three-group case, all items except Item 36 displayed a higher Type I error rate for pairwise as compared with composite comparisons. In the four-group case, all items except Items 3, 6, 33, 52, and 60 displayed a higher Type I error rate for pairwise as compared with composite comparisons. In the five-group case, 10 items out of 59 (i.e., Items 4, 8, 23, 29, 31, 32, 36, 48, 50, and 52) displayed a higher Type I error rate for composite as compared with pairwise comparisons.

The results of FDRs are given in Table 5. All bold and italic numbers represent unique results that are discussed in the remainder of this paragraph. In the three-group case, all but two items had pairwise FDRs that were higher than composite group

Table 5. Maximum False Discovery Rates.

Item	Three groups		Five groups	
	Pairwise	Composite	Pairwise	Composite
2	.009	.004	.013	.005
3	.008	.003	.024	.011
4	.032	.010	.005	.003
5	.015	.004	.032	.005
6	.017	.004	.027	.008
7	.031	.009	.029	.008
8	.038	.009	.002	.006
9	.008	.004	.002	.008
10	.015	.006	.001	.013
11	.032	.006	.026	.004
12	.024	.007	.011	.005
13	.016	.007	.028	.004
14	.027	.007	.036	.011
15	.031	.005	.003	.009
16	.026	.007	.038	.007
17	.033	.006	.036	.009
18	.006	.007	.018	.008
19	.040	.010	.024	.005
20	.007	.004	.008	.006
21	.018	.005	.004	.009
22	.007	.006	.008	.006
23	.015	.005	.005	.003
24	.032	.009	.036	.006
25	.011	.006	.010	.006
26	.013	.009	.015	.006
27	.031	.007	.033	.005
28	.020	.005	.026	.005
29	.027	.009	.037	.004
30	.006	.004	.007	.006
31	.010	.004	.009	.008
32	.031	.007	.037	.006
33	.006	.005	.020	.007
34	.019	.004	.032	.006
35	.022	.007	.024	.005
36	.007	.004	.009	.013
37	.007	.002	.029	.010
38	.018	.009	.023	.010
39	.019	.010	.020	.006
40	.033	.006	.034	.005
41	.030	.006	.031	.010
42	.020	.007	.033	.004
43	.030	.003	.032	.006
44	.028	.009	.028	.006
45	.002	.003	.013	.008
46	.031	.009	.032	.006

(continued)

Table 5. (continued)

Item	Three groups		Five groups	
	Pairwise	Composite	Pairwise	Composite
47	.020	.006	.026	.003
48	.029	.007	.009	.008
49	.029	.003	.030	.008
50	.020	.009	.021	.005
51	.012	.006	.031	.007
52	.017	.006	.023	.012
53	.011	.004	.020	.006
54	.016	.006	.018	.003
55	.010	.004	.032	.009
56	.011	.009	.024	.004
57	.007	.004	.008	.006
58	.028	.006	.033	.004
59	.017	.004	.044	.007
60	.015	.004	.018	.005
Mean	.190	.006	.021	.006
Max	.040	.010	.044	.013

Note. All bold and italic numbers represent unique results that are discussed in text.

FDRs. In the five-group case, all items had FDRs that were higher in pairwise comparisons than composite group comparisons, except Items 8, 9, 10, 15, 21, and 36. At the bottom of Table 5, it can be seen that the mean FDR across items is lower for composite as compared with pairwise in both three- and five-group cases. The same pattern is seen when we selected the maximum FDR across items.

Operational Data Study

In order to demonstrate the findings across pairwise and composite group comparisons, an operational data set was also analyzed. The data set used in this study was a subset of the data (ST²L) described and analyzed in Huggins, Ritzhaupt, and Dawson (2014). The subset of the original data we used consisted of dichotomous responses of 4,682 middle school students from 13 Florida school districts to 16 items measuring digital citizenship of students. There was no missing data on these items for these students. The data had the students' grade (e.g., G6, G7, and G8), race (e.g., R1 = Black, R2 = Hispanic, R3 = White, R4 = Others), and age (e.g., A11, A12, A13, A14, and A15) information. Thus, in this study the students were divided into grade, racial, and age categories to be able to create three, four, and five groups, respectively. It should be noted that we selected a subset of items that displayed some DIF, while the full ST²L scale showed much less DIF in the total 114 items.

The test was assessed for dimensionality in Mplus (Muthén & Muthén, 1998-2012), and results indicated unidimensionality: comparative fit index (CFI) = 0.97;

Table 6. Three-Group Case Based on Grade (Grades 6, 7, and 8).

Item	Pairwise comparisons			Composite group comparisons		
	G6 vs. G7	G6 vs. G8	G7 vs. G8	G6 vs. C	G7 vs. C	G8 vs. C
1	0.57	0.39	0.45	0.21	0.26	0.40
3	0.73	0.41	0.94	0.35	0.81	0.93
4	1.80	0.48	2.08	1.14	1.94	0.77
8	0.17	2.22	1.82	1.64	1.44	1.03
9	1.94	3.91*	1.73	3.82*	0.95	1.18
10	3.54*	3.22	1.87	3.31*	1.80	0.49
11	2.24	1.88	2.88	2.08	2.50	1.19
12	1.71	1.06	2.58	0.90	2.62	0.47
15	3.32*	0.99	2.57	1.16	3.08	0.46
16	0.97	1.53	2.12	0.89	1.80	1.08
17	1.97	2.41	0.87	1.49	0.56	3.32*
19	0.08	1.40	1.39	0.22	0.13	1.85
20	1.11	2.85	1.57	2.38	0.93	1.18
21	1.27	0.47	1.20	0.81	0.89	0.86
23	2.45	2.53	2.18	2.79	1.67	1.03
25	2.57	1.84	1.34	1.91	2.00	0.46

*Significant at .01 alpha level.

Tucker–Lewis index (TLI) = 0.97; root mean square error of approximation (RMSEA) = .03; $\chi^2(90) = 531.00$, $p < .001$. Then, pairwise and composite group comparisons were conducted with the three-group case, the four-group case, and the five-group case under the same conditions in simulations (e.g., 2PL model, R software, difR package).

Results for pairwise and composite group comparisons when the students were categorized according to grade are given in Table 6. We considered any DIF that was statistically significant to be of concern. In both pairwise and composite group comparisons, three items displayed DIF with respect to at least one of the comparisons, but not the same three items. In pairwise comparisons, some comparisons between G6 and G7 as well as G6 and G8 were flagged for DIF. In composite group comparisons, two items had DIF for G6 and one item had DIF for G8.

Results for pairwise and composite group comparisons when the students were categorized according to race are given in Table 7. Both pairwise and composite group comparisons showed statistically significant DIF problems across multiple items. In pairwise comparisons, several items had significant DIF effects across comparisons made between the Groups R1 and R2, Groups R2 and R3, and Groups R3 and R4. However, in composite group findings all but one item with DIF concerns were associated with R2. In fact, R2 was flagged as having 11 items with DIF, signifying a major concern with this group that was not as obvious with pairwise findings.

Results for pairwise and composite group comparisons when the students were categorized based on age are given in Table 8. Statistically significant DIF was found

Table 7. Four-Group Case Based on Race (Races 1-4).

Item	Pairwise comparisons						Composite group comparisons											
	R1 vs. R2			R1 vs. R3			R1 vs. R4			R2 vs. R3			R2 vs. R4			R3 vs. R4		
	R1 vs. R2	R1 vs. R3	R1 vs. R4	R2 vs. R3	R2 vs. R4	R3 vs. R4	R1 vs. C	R2 vs. C	R3 vs. C	R4 vs. C	R1 vs. C	R2 vs. C	R3 vs. C	R4 vs. C	R1 vs. C	R2 vs. C	R3 vs. C	R4 vs. C
1	0.26	0.32	0.15	0.20	0.12	0.08	0.38	0.17	0.20	0.02								
3	2.03	0.85	0.60	3.28*	0.91	0.89	1.50	3.30*	0.49	0.69								
4	2.81	2.08	2.99	4.14*	0.72	4.41*	1.92	3.52*	1.42	3.83*								
8	4.60*	0.27	0.65	5.80*	1.95	0.70	0.82	6.18*	1.02	0.31								
9	1.66	0.07	0.78	1.66	1.37	0.80	0.85	1.62	0.93	0.83								
10	3.45*	2.03	0.94	4.46*	1.26	1.67	2.22	4.05*	1.27	1.09								
11	4.32*	2.29	2.12	7.05*	1.54	3.26	1.66	6.66*	1.64	2.48								
12	3.15	1.36	1.85	4.14*	0.83	2.34	1.34	3.79*	1.15	1.83								
15	4.26*	2.83	2.89	3.68*	4.49	2.52	2.08	3.80*	0.77	2.37								
16	1.22	0.69	0.28	2.34	0.75	0.29	0.16	1.96	0.85	0.24								
17	1.10	1.15	0.97	0.42	0.51	0.47	0.89	0.66	0.48	0.64								
19	4.68*	1.19	2.03	6.29*	1.25	2.98	0.43	5.92*	1.87	2.06								
20	3.06	4.08*	2.92	3.79*	1.21	0.75	3.15	3.33*	1.50	1.25								
21	2.55	0.94	1.77	3.93*	1.12	2.08	1.11	3.67*	1.02	1.57								
23	1.82	1.64	0.78	2.68	1.68	0.82	1.38	2.63	0.53	0.82								
25	2.27	2.62	0.82	4.08*	1.08	1.56	1.92	3.49*	1.52	0.89								

*Significant at .01 alpha level.

Table 8. Five-Group Case Based on Age (Ages 11-15).

Item	Pairwise comparisons composite group comparisons															
	A11 vs. A12	A11 vs. A13	A11 vs. A14	A11 vs. A15	A12 vs. A13	A12 vs. A14	A12 vs. A15	A13 vs. A14	A13 vs. A15	A14 vs. A15	A11 vs. A12 vs. A15	A11 vs. A12 vs. A14	A11 vs. A13 vs. A15	A12 vs. A13 vs. A15	A12 vs. A14 vs. A15	A13 vs. A14 vs. A15
	vs. C	vs. C	vs. C	vs. C	vs. C	vs. C	vs. C	vs. C	vs. C	vs. C	vs. C	vs. C	vs. C	vs. C	vs. C	vs. C
1	0.70	0.82	0.32	0.08	0.26	0.56	0.01	0.62	0.04	0.07	0.98	0.25	0.06	0.80	0.04	0.04
3	0.71	0.80	0.74	1.86	0.14	1.18	3.02	1.20	5.03*	4.13*	0.72	0.09	0.39	1.36	6.57*	6.57*
4	3.50*	2.49	2.00	4.93*	2.27	2.77	6.72*	0.86	8.24*	7.16*	2.65	2.91	0.05	1.07	12.16*	12.16*
8	2.04	0.45	0.76	3.96	2.90	2.41	4.89*	0.79	7.08*	6.38*	0.77	3.26	0.59	0.52	10.21*	10.21*
9	1.90	1.62	1.38	4.05*	2.79	2.49	3.75*	0.29	6.04*	5.67*	1.86	3.05	0.64	0.31	7.77*	7.77*
10	2.61	2.62	3.48*	5.01*	2.55	2.74	6.42*	0.95	8.75*	7.63*	2.11	2.72	1.22	1.04	13.21*	13.21*
11	2.15	1.89	1.20	5.17*	3.81*	3.94*	7.79*	1.67	11.25*	8.44*	1.66	4.20*	1.16	1.79	14.87*	14.87*
12	2.49	1.12	0.91	5.25*	2.24	3.24	6.46*	1.25	8.21*	7.91*	1.26	2.85	0.07	1.52	13.56*	13.56*
15	2.03	2.88	1.29	5.43*	1.36	1.25	6.83*	2.63	6.63*	7.86*	2.08	0.39	2.27	1.19	12.91*	12.91*
16	1.00	0.47	0.39	1.65	1.68	1.71	1.90	0.98	3.23	2.59	0.47	2.04	0.09	0.97	3.67*	3.67*
17	1.26	2.27	2.10	4.41*	3.45*	3.26	5.38*	0.54	7.11*	6.88*	1.52	3.41*	1.40	1.18	11.37*	11.37*
19	3.48*	2.51	2.35	4.91*	2.61	2.95	6.75*	0.48	8.17*	7.65*	2.68	3.15	0.49	1.06	13.41*	13.41*
20	0.81	1.32	1.65	4.33*	2.60	2.28	5.48*	1.85	7.79*	6.00*	1.06	2.56	0.75	1.48	10.24*	10.24*
21	1.05	0.55	1.83	1.88	1.24	1.71	1.41	0.90	2.04	2.26	0.53	1.63	0.26	0.83	2.20	2.20
23	3.14	2.80	1.95	4.86*	4.50*	2.67	5.83*	2.45	9.67*	7.08*	2.47	4.40*	1.89	1.06	12.24*	12.24*
25	2.33	2.52	2.16	4.93*	1.58	1.40	6.59*	0.39	8.24*	7.07*	2.50	1.80	0.41	0.15	12.20*	12.20*

*Significant at .01 alpha level.

in many pairwise comparisons with most comparisons with group A15 being problematic. Composite group comparisons clearly showed that two groups (A12 and A15) had several DIF effects, but similar to the pairwise approach, composite group comparisons showed group A15 as having the most DIF effects of concern.

Discussion and Conclusion

The purpose of the study was to examine the impact of two methods of defining group comparisons in DIF detection. The results showed that pairwise and composite method results did not differ much in interpretation when only one group differed from other groups in true b -parameter differences. Under pairwise methods, every pair that included the group of concern was flagged, so a practitioner would easily be able to interpret that the last group had problematic concerns for fairness. Under the composite group methods, the group of concern was flagged through its group-specific DIF effect. So even though the DIF estimates from the methods are different because they are comparing different groups, a researcher or practitioner would draw the same conclusions regardless of the choice between pairwise and composite. This, of course, is assuming that a Type I error or false discovery was not committed, and it should be noted that we found the chances of committing one of these errors are higher in pairwise comparisons as compared with composite group comparisons.

In the case in which moderate to large true b -parameter differences were spread among the groups, different conclusions were often drawn between the pairwise and composite methods. For example, Table 2 shows that when the true b -parameter differences were large across three groups pairwise methods indicated some problems for all the pairs which would lead a practitioner to interpret that all groups have some concerns. If one was concerned with group-to-group invariance, this would be an appropriate interpretation of DIF. However, Table 2 also shows that when true b -parameter differences were large across three groups, composite methods indicated that Group 1 and Group 3 had DIF concerns, but not Group 2. The interpretation is different as compared with pairwise, but it is more appropriate if one is concerned with group item parameters being invariant to the operational item parameters. This is just one of many examples of when true b -parameter differences resulted in different interpretations between pairwise and composite approaches. In these cases, it is critical to make an informed decision about which method to use. Researchers do not know the nature of true b -parameter group differences in their observed data indicating that they do not have a priori knowledge as to whether or not pairwise and composite group methods will result in similar or dissimilar DIF interpretations. Therefore, it is always important to make an informed choice between pairwise and composite methods.

This study also showed that one reason to consider using composite approaches is because of the ease in interpretation. For example, the right side of Table 2 (Columns 7 and 8) shows that pairwise comparisons indicated problematic DIF for Groups 1 to 5 in various pairwise sets. However, if one looks at pairs without flagged problems

of DIF, these include Groups 1 to 5. In other words, all five groups are invariant to some groups and not invariant to others on different items. So how does one decide where the problem lies? Which groups are being disadvantaged? A researcher could reduce the number of pairwise comparisons by using a single reference group; however, one will only be able to interpret DIF between the reference group and focal groups and will not be able to interpret DIF problems among the other groups that are selected as focal groups. If these issues are deemed problematic in a particular application of DIF detection, it is better to use a composite comparison approach which assigns each group with its own DIF effect. Composite comparisons make the group and direction of advantage very clear. Looking at the effect sizes and the percentages of statistical significances on the right side of Table 2 (columns 7 and 8), for example, indicates that Group 1 is advantaged (i.e., smaller b estimate than composite b estimate) and Group 5 is disadvantaged (i.e., larger b estimate than composite b estimate). Particularly when there are many groups, this can be an advantage of composite group comparisons. The application with operational data displayed the same advantage for the composite group comparisons.

Composite Type I error rates were lower in the vast majority of situations. However, this advantage of composite method was more obvious when all possible pairwise comparisons are made, as opposed to when only one reference group is chosen to conduct pairwise comparisons. But again for emphasis, in all cases the vast majority of results showed lower Type I error rates for composite method as compared with pairwise methods. This occurs because of a combination of two factors: (a) the relatively lower number of comparisons made in composite methods as compared with pairwise, and (b) the larger sample size used in composite methods as compared with pairwise. When all pairwise comparisons are made, both factors play a role in the lower Type I error rate for composite methods. When one reference group is chosen for pairwise comparisons, factor b only plays a role in the lower Type I error rate for composite methods.

For FDRs, with a few exceptions, composite method was superior to pairwise method. This was true at the item and aggregate level. Similar to Type I error rates, the reasons for these FDR results are due to some combination of two factors, those of less number of comparisons in composite methods (i.e., less accumulated alpha) and larger sample size in composite methods. These factors result in more correct decisions being made in the hypothesis testing process, which is associated with less false discoveries.

In the simulation study, a discrepancy was sometimes found between the effect sizes and percentage of statistical significance of DIF when looking across pairwise and composite comparisons. For example, in Table 2 (three-group condition under small b -parameter differences), a $UA = .25$ effect size was detected in the pairwise comparison of Group 2 with Group 3 and a $UA = .18$ effect size was detected in the comparison of Group 1 with the composite group, yet the smaller effect size had a noticeably larger percentage of statistically significant DIF. The same thing was seen in the four- and five-group conditions under small true b -parameter difference (e.g.,

pairwise of Group 1 vs. Group 3 and the composite group comparison of Group 1 vs. composite). These discrepancies between the magnitude of effect size and percentage of statistical significance were most likely due to sample size differences, as composite comparisons by nature include a larger sample size than pairwise, assuming more than two groups of examinees are under investigation. Larger sample sizes lead to more power, and the Type I error rate investigation we conducted ensures that the higher rejection rate is not coincided with a problematic rate of Type I errors. Indeed, observed Type I error rates are lower under composite than pairwise methods, so more frequent (and accurate) rejection of the null under composite group comparisons can be seen as an additional benefit of the method over pairwise methods.

In sum, this study showed that pairwise and composite methods do not always provide the same results and there are several implications of this finding. First, it is important for practitioners to determine how one defines fairness as a lack of bias on their assessment. If the practitioner wants to know if two groups are invariant to each other, then they should use pairwise methods. If the practitioner is interested in whether or not the operational item parameters are invariant to individual group parameters, they should use the composite group approach. It can be argued that the *Standards* (1999) favor the latter interpretation of fairness a lack of bias. Second, if a researcher is using the pairwise approach as is commonplace but is unable to interpret which individual groups have DIF concerns, she or he may want to consider running an additional DIF analysis using the composite group approach. This will allow for ease of interpretation and connect the DIF concerns directly to the item parameters used for developing reported scores. It seems that there are multiple instances in which it may be appropriate to use both methods in conjunction to obtain a full picture of possible item bias problems. Third, the composite group approach has a clear advantage in terms of Type I error rates and FDRs, and this must be kept in mind when choosing a method and/or using both methods conjointly.

This study does not argue against pairwise comparisons, as the composite and pairwise methods are different rather than correct/incorrect. Rather, this study simply makes the case that the two methods can provide different interpretations of DIF results, that practitioners can decide which to use based on definitions of fairness that are relevant to their purpose of DIF analysis, and that the uncommon approach of composite group comparisons has several advantages and should deserve more consideration in future DIF studies.

Limitations and Further Research

DIF in the simulation portion of the study was introduced by changing the item difficulty parameters of only one item. Often times, more than one item in a test will display some DIF concerns. It is strongly recommended that this study be replicated with multiple items containing DIF as well as unequal group ratios. As mentioned previously, when conducting composite group comparisons, weighting groups may be necessary as is seen in some applications of detecting equating invariance. Last,

we created a composite group in such a way that it included all groups, which introduces dependency between the focal and the composite groups (Kanje, 2007). Future studies may want to test DIF between a single group and a modified composite group in which all examinees are included with the exception of those from the single group being tested.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- ACT, Inc. (1997). *ACT Assessment technical manual*. Iowa City, IA: Author.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- American Psychological Association. (2009). *The publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57, 289-300.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.
- Cauffman, E., & ManIntosh, R. (2006). A Rasch differential item functioning analysis of the Massachusetts Youth Screening instrument. *Educational and Psychological Measurement*, 66, 502-521.
- 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. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37, 281-306.
- Ellis, B., & Kimmel, H. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology*, 77, 177-184.
- Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, 26, 3-24.
- Hidalgo, M. D., & Lopez-Pina, J. (2004). Differential item functioning detection and effect-size: A comparison between LR and MH procedures for detecting differential item functioning. *Educational and Psychological Measurement*, 64, 903-915.
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

- Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.
- Huggins, A. C., Ritzhaupt, A., & Dawson, K. (2014, April). *Validation of the student tool for technology literacy*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.
- Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.
- Kanjee, A. (2007). Using logistic regression to detect bias when multiple groups are tested. *South African Journal of Psychology, 37*, 47-61.
- Kim, S. H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood test on detection of differential item functioning. *Applied Measurement in Education, 8*, 291-312.
- Liu, J., & Dorans, J. N. (2013). Assessing a critical aspect of construct continuity when test specifications change or test forms deviate from specifications. *Educational Measurement: Issues and Practice, 32*(1), 15-22.
- Liu, M., & Holland, P. W. (2008). Exploring population sensitivity of linking functions across three law school admissions test administrations. *Applied Psychological Measurement, 32*, 27-44.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*, 847-862.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.
- Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.
- Osterlind, S., & Everson, H. (2009). *Differential item functioning*. Thousand Oaks, CA: Sage.
- Penfield, R. D. (2001). Assessing differential item functioning across multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education, 14*, 235-259.
- Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 125-167). Amsterdam, Netherlands: Elsevier.
- R Development Core Team. (2013). *R: A language and environment for statistical computing, reference index* (Version 2.2.1). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org>
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.
- Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*, 77-105.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.
- Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27*, 77-83.

- Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, 33, 42-57.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum.
- Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (RR-12-08). Princeton, NJ: Educational Testing Service.
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.