Effects of Differing Item Parameters on Closed-Interval DIF Statistics

Zachary S. Feinstein
University of Minnesota

The closed-interval signed area (CSA) and closed-interval unsigned area (CUA) statistics were studied by monte carlo simulation to detect differential item functioning when the reference and focal groups had different parameter distributions. When the pseudo-guessing parameter was varied, the CSA was better able to detect moderate to large differences between the groups than the CUA. However, the effect of the pseudo-guessing parameter varied depending on item discriminations. *Index terms: closed-interval measures, differential item functioning, item response theory, monte carlo simulation, signed area measures, unsigned area measures.*

In item response theory (IRT), differential item functioning (DIF) is the term used to describe test items that have different probabilities of a correct (keyed) response, given the same trait level for two groups of examinees—usually referred to as the reference and focal groups (Holland & Wainer, 1993). The reference group is usually the majority group in a population (a base group) and the focal group represents a population minority group (usually the focus of a DIF study) (Angoff, 1993). DIF has been assessed by different methods. Comparison of a reference group to a focal group, when controlling for trait level, is accomplished by conventional (observed score) and latent trait procedures. Controlling for trait level differentiates DIF from impact (Dorans & Holland, 1993), in which overall group differences are examined without conditioning on trait level.

Methods for Identifying DIF

Non-IRT methods. Millsap & Everson (1993) provided a thorough review of non-IRT (observed score) methods for identifying DIF, as well as IRT-based (unobserved score) methods. Non-IRT methods have many weaknesses, even though the MH statistic is quite popular (Angoff, 1993) and statistically powerful (Dorans & Holland, 1993). The transformed item difficulty method (or delta plot method) does not take item discrimination or the propensity to guess an item correctly into account (Angoff). Conventional methods typically use an internal matching criterion (Angoff; Dorans & Holland, 1993), such as a person's total score. Zieky (1993) indicated that using the total score results in a possible biased test because it includes DIF items in the matching criterion.

IRT methods. IRT methods for DIF detection (Bock, 1993; Millsap & Everson, 1993; Thissen, Steinberg, & Wainer, 1993) differ from conventional methods because in IRT local independence of items is assumed (Hambleton & Swaminathan, 1985). Therefore, a total score matching criterion is not used; rather, an estimate of a person's trait level (θ) is used as a matching criterion. One requirement for detecting DIF using IRT methods is that the test should be unidimensional (Cole, 1993). Use of these methods alleviates problems with the circularity of using a conventional matching criterion, such as sample specificity. In the three-parameter logistic model (3PLM; Hambleton & Swaminathan, 1985), the item response function (IRF) is described by the difficulty (b), discrimination (a), and pseudo-guessing parameters (c; probability that a group member of infinitely low θ will correctly answer the item). IRT DIF methods examine the differences in the parameter

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 19, No. 2, June 1995, pp. 131–142

0146-6216/95/020131-12$1.85

estimates between the reference and focal groups (Lord, 1980); however, *c*s usually are held constant across groups (Cohen, Kim, & Subkoviak, 1991; Lord; Osterlind, 1983).

Uniform DIF occurs when the differences between IRFs are constant across θ; nonuniform DIF occurs when these differences are not constant across all θs (Millsap & Everson, 1993). IRFs can be compared using area DIF detection methods (Wainer, 1993) to indicate whether no DIF, uniform DIF, or nonuniform DIF exists. DIF may be conceptualized as the signed or unsigned area between two IRFs (Cohen et al., 1991; Kim & Cohen, 1991, 1992). A signed-area DIF statistic may indicate no DIF, yet DIF could be apparent when examined graphically because positive and negative area differences between two group IRFs may cancel each other (Shealy & Stout, 1993a). This occurs only when there is nonuniform DIF and group IRFs cross symmetrically (Cohen et al.); therefore, an unsigned index appears to be more appropriate to detect DIF. Linn (1993) noted that graphical methods of exhibiting DIF, such as those that compare IRFs, are beneficial to test developers because of their interpretive ease.

The area between the reference and focal group IRFs can be examined using either an exact area or a closed-interval method (Raju, 1988). Exact area methods use integral calculus to find the area between two IRFs. Integration methods may give exact estimates for the area between two IRFs, but the area between two IRFs cannot be computed for the 3PLM, because varying the *c*s causes the area estimates to go toward infinity (Raju). Kim & Cohen (1991) discussed methods of computing the area between two IRFs when *a*, *b*, and *c* vary. For closed-interval methods, the area underneath the IRF between two arbitrarily selected points on the θ continuum is computed by moving along the θ continuum. The area below one group's IRF is subtracted from the area below another group's IRF. Raju claimed that closed-interval methods are biased; however, Millsap & Everson (1993) discussed how "closed-interval" area measures may be as efficient as "unbounded" (or exact) area measures, and they called for more investigations of the bounded area difference as a method of DIF detection.

Purpose

This study examined the closed-interval signed area (CSA) and closed-interval unsigned area (CUA) statistics for DIF detection. Millsap & Everson (1993) found no evidence of superiority of one statistic over the other, but suggested investigating the methods to determine the conditions under which each statistic would perform the best. The purpose of this study was to examine how the closed-interval DIF detection methods, CSA and CUA, operated under various conditions.

This study differs from other empirical investigations that investigated parameter estimation when the majority of items were not biased (i.e., only a few items exhibited DIF; Miller & Oshima, 1992). This study examined the effects of pervasive DIF, or DIF that occurred for all items on a test (Shealy & Stout, 1993b). That is, the item parameters differed between the reference and focal groups because the focal group reacted to a "nuisance trait" (i.e., a confounding dimension for an item in a multidimensional framework of examining DIF), but the reference group did not (Donoghue, Holland, & Thayer, 1993; Millsap & Everson, 1993; Shealy & Stout, 1993b). Examples of such a pervasive nuisance trait may be examinee groups reacting in a different way to a set of instructions for a test, or a test not measuring a particular construct it purports to measure. A test comprised of many DIF items indicates "test bias" (Camilli, 1993; Shealy & Stout, 1993b) or "differential test functioning." However, there is not much empirical evidence showing a direct connection between DIF and differential test functioning (Camilli). The focal group was always designed to be at a disadvantage in this study, and it should be noted that multidimensional effects may cancel each other with real data (Shealy & Stout, 1993b).

Method

IRF Parameters

The CSAs and CUAs between two IRFs were examined as a function of varying item parameters that would reflect group differences. Hambleton & Swaminathan (1985) defined the 3PLM IRF as:

$$P_i(\theta) = c_i + (1 - c_i)\left\{1 + \exp\left[-Da_i(\theta - b_i)\right]\right\}^{-1}, \quad (1)$$

where

$P_i(\theta)$ is the probability of a correct response given a person's θ level;

i indexes an item, $i = 1, \ldots, n$;

a, b, and c are the item parameters (defined above); and

D is a scaling constant equivalent to 1.7.

The two IRFs represented the reference and focal groups, respectively. Both the CSA and the CUA statistics were computed along the q continuum from -3.5 to 3.5.

The a, b, and c distributions were varied for the reference and focal group IRFs. All item parameters were generated to be normally distributed, and they simulated a 75-item power test with no missing data. Each of 24 different combinations of item parameter values were simulated to comprise a test (see Table 1). Baseline levels were simulated for b and c to have no bias (Miller & Oshima, 1992). One of the purposes of varying the a parameter was because DIF detection is largely dependent on a (Donoghue et al., 1993). Highly discriminating items tend to be identified as DIF items more often than less discriminating items (Burton & Burton, 1993; Linn, 1993), sometimes being identified as a cause for DIF over and above when b differences between groups exhibit DIF (Rudner, Getson, & Knight, 1980), because a differences shift the IRF of one of the groups. However, Miller and Oshima found that differences between as for different groups do not account for much accumulated difference between test response functions. b and c parameter effects were examined within levels of a. Interactions between all parameter effects also were investigated.

Two different levels of a were used. Level A1 had an intended mean of $a = .75$ [intended standard deviation (SD) = .1]. Level A2 had an intended mean of $a = 1.50$ (intended SD = .2), to simulate a test that was able to discriminate well between low- and high-scoring examinees. Conditions 1–12 simulated Level A1, and Conditions 13–24 simulated Level A2, as shown in Table 1.

Different θ distributions for each group are common in DIF studies (Donoghue et al., 1993; Shealy & Stout, 1993a). Different θ distributions were simulated here as differences between the item difficulties. They represented no difference, moderate difference, and large difference in the θ levels between the reference and focal groups. This study addressed the b parameter as the b/θ parameter because both θ and b are on an additive scale (Hambleton & Swaminathan, 1985). The procedure of varying bs instead of θs controlled for the errors that might normally occur with linking and equating for calibrating examinees onto a common θ scale (Linn & Levine, 1981).

All b distributions had intended SDs = 1.0. In Level B1, both the reference group and the focal group were generated to have an intended mean of $b = 0.0$ (the baseline level for the b effect). For the next three levels of b, the reference groups' intended bs were generated to be less than the focal groups' intended bs. Level B2 for the b effect was simulated in accordance with what Miller & Oshima (1992) termed a "moderate difference"—the reference group was generated to have an intended mean of $b = -.35$, and the focal group was generated to have an intended mean of $b = 0.0$. Level B3 simulated a large difference (Osterlind, 1983)—the reference group was generated to have an intended mean of $b = -.70$ and the focal group was generated to have an intended mean of $b = .20$. Conditions 1–4 and 13–16 simulated Level B1, Conditions 5–8 and 17–20 simulated Level B2, and Conditions 9–12 and 21–24 simulated Level B3, as shown in Table 1.

The cs were varied to simulate a lower propensity to guess a multiple-choice item's answer correctly in the focal group and a higher propensity to guess in the reference group. This hypothesis is not documented in previous literature (Ironson & Subkoviak, 1979). DIF detection could depend on how the varying cs can shift the location of the IRF to be significantly different from another group's IRF so as to result in DIF, and this may affect detection due to other parameter effects (Donoghue et al., 1993).

Table 1

Intended Means, Actual Means, and Actual SDs for the *a*s, *b*s, and *c*s for the Reference and Focal Group for Each Condition

	Reference			Focal				Reference			Focal		
Condition	*a*	*b*	*c*	*a*	*b*	*c*	Condition	*a*	*b*	*c*	*a*	*b*	*c*
Condition 1							Condition 13						
Intended Mean	.75	0.00	.25	.75	0.00	.25	Intended Mean	1.50	0.00	.25	1.50	0.00	.25
Actual Mean	.76	−.09	.25	.74	−.04	.24	Actual Mean	1.53	.13	.25	1.48	−.12	.25
Intended SD	.10	1.00	.05	.10	1.00	.05	Intended SD	.20	1.00	.05	.20	1.00	.05
Actual SD	.09	.91	.06	.10	1.08	.05	Actual SD	.20	.97	.05	.19	.97	.05
Condition 2							Condition 14						
Intended Mean	.75	0.00	.25	.75	0.00	.20	Intended Mean	1.50	0.00	.25	1.50	0.00	.20
Actual Mean	.74	−.08	.25	.76	.06	.20	Actual Mean	1.51	.11	.25	1.49	.18	.20
Intended SD	.10	1.00	.05	.10	1.00	.05	Intended SD	.20	1.00	.05	.20	1.00	.05
Actual SD	.09	1.01	.05	.10	1.01	.05	Actual SD	.19	1.02	.05	.19	.95	.05
Condition 3							Condition 15						
Intended Mean	.75	0.00	.30	.75	0.00	.25	Intended Mean	1.50	0.00	.30	1.50	0.00	.25
Actual Mean	.75	.34	.30	.76	−.20	.24	Actual Mean	1.52	0.00	.30	1.49	.13	.25
Intended SD	.10	1.00	.05	.10	1.00	.05	Intended SD	.20	1.00	.05	.20	1.00	.05
Actual SD	.11	1.09	.04	.10	1.00	.05	Actual SD	.18	.96	.06	.20	1.01	.06
Condition 4							Condition 16						
Intended Mean	.75	0.00	.35	.75	0.00	.25	Intended Mean	1.50	0.00	.35	1.50	0.00	.25
Actual Mean	.73	.12	.35	.76	.07	.25	Actual Mean	1.52	−.04	.35	1.48	−.04	.26
Intended SD	.10	1.00	.05	.10	1.00	.05	Intended SD	.20	1.00	.05	.20	1.00	.05
Actual SD	.10	.96	.05	.09	.97	.06	Actual SD	.21	1.07	.05	.19	.82	.05
Condition 5							Condition 17						
Intended Mean	.75	−.35	.25	.75	0.00	.25	Intended Mean	1.50	−.35	.25	1.50	0.00	.25
Actual Mean	.74	−.43	.25	.73	.01	.24	Actual Mean	1.53	−.36	.26	1.47	−.08	.26
Intended SD	.10	1.00	.05	.10	1.00	.05	Intended SD	.20	1.00	.05	.20	1.00	.05
Actual SD	.09	1.03	.05	.10	.97	.05	Actual SD	.19	1.09	.04	.20	1.01	.05
Condition 6							Condition 18						
Intended Mean	.75	−.35	.25	.75	0.00	.20	Intended Mean	1.50	−.35	.25	1.50	0.00	.20
Actual Mean	.73	−.20	.25	.74	.01	.19	Actual Mean	1.49	−.46	.25	1.48	−.03	.19
Intended SD	.10	1.00	.05	.10	1.00	.05	Intended SD	.20	1.00	.05	.20	1.00	.05
Actual SD	.11	.87	.05	.10	.87	.05	Actual SD	.17	.87	.05	.19	.91	.05
Condition 7							Condition 19						
Intended Mean	.75	−.35	.30	.75	0.00	.25	Intended Mean	1.50	−.35	.30	1.50	0.00	.25
Actual Mean	.77	−.37	.30	.76	−.11	.25	Actual Mean	1.51	−.50	.30	1.51	.16	.24
Intended SD	.10	1.00	.05	.10	1.00	.05	Intended SD	.20	1.00	.05	.20	1.00	.05
Actual SD	.10	1.03	.05	.08	.83	.06	Actual SD	.21	.87	.05	.19	.90	.05
Condition 8							Condition 20						
Intended Mean	.75	−.35	.35	.75	0.00	.25	Intended Mean	1.50	−.35	.35	1.50	0.00	.25
Actual Mean	.74	−.43	.35	.74	.26	.25	Actual Mean	1.50	−.26	.36	1.48	−.10	.25
Intended SD	.10	1.00	.05	.10	1.00	.05	Intended SD	.20	1.00	.05	.20	1.00	.05
Actual SD	.12	.95	.05	.11	1.03	.06	Actual SD	.25	.98	.06	.19	.98	.05
Condition 9							Condition 21						
Intended Mean	.75	−.70	.25	.75	.20	.25	Intended Mean	1.50	−.70	.25	1.50	.20	.25
Actual Mean	.76	−.82	.24	.77	.21	.25	Actual Mean	1.50	−.52	.25	1.51	.08	.25
Intended SD	.10	1.00	.05	.10	1.00	.05	Intended SD	.20	1.00	.05	.20	1.00	.05
Actual SD	.09	.97	.05	.09	.92	.05	Actual SD	.17	.84	.05	.20	1.10	.05
Condition 10							Condition 22						
Intended Mean	.75	−.70	.25	.75	.20	.20	Intended Mean	1.50	−.70	.25	1.50	.20	.20
Actual Mean	.75	−.79	.25	.74	.36	.20	Actual Mean	1.47	−.73	.26	1.50	.28	.20
Intended SD	.10	1.00	.05	.10	1.00	.05	Intended SD	.20	1.00	.05	.20	1.00	.05
Actual SD	.10	1.03	.06	.09	1.04	.05	Actual SD	.20	.96	.05	.19	.99	.06
Condition 11							Condition 23						
Intended Mean	.75	−.70	.30	.75	.20	.25	Intended Mean	1.50	−.70	.30	1.50	.20	.25
Actual Mean	.76	−.62	.30	.74	.21	.25	Actual Mean	1.47	−.54	.31	1.47	.31	.25
Intended SD	.10	1.00	.05	.10	1.00	.05	Intended SD	.20	1.00	.05	.20	1.00	.05
Actual SD	.10	.93	.05	.11	.91	.04	Actual SD	.20	.81	.06	.20	1.10	.05
Condition 12							Condition 24						
Intended Mean	.75	−.70	.35	.75	.20	.25	Intended Mean	1.50	−.70	.35	1.50	.20	.25
Actual Mean	.76	−.69	.35	.74	.27	.25	Actual Mean	1.50	−.72	.35	1.48	.16	.26
Intended SD	.10	1.00	.05	.10	1.00	.05	Intended SD	.20	1.00	.05	.20	1.00	.05
Actual SD	.10	.96	.05	.11	.94	.05	Actual SD	.23	.96	.05	.23	.97	.05

Four levels of c were simulated. One level simulated no difference, two levels simulated small differences, and one level simulated a large difference between the cs for the reference and focal groups. All items were simulated as four-option multiple-choice items (intended SD = .05). The baseline level (Level C1) was generated to have an intended mean c = .25 for both groups. Level C2 had a reference group intended mean c = .25 with a focal group intended mean c = .20. Level C3 was generated to have a reference group intended mean c = .30 with a focal group intended mean of c = .25. Level C4 was generated to have a reference group intended mean of c = .35 with a focal group intended mean of c = .25.

1,500 simulated examinees were generated. 1,000 of the examinees represented the reference group; 500 of the examinees represented the focal group. Random numbers were generated by an algorithm obtained from *Numerical Recipes in C* (Press, Flannery, Teukoisky, & Vetterling, 1992). Examinee true θs were generated to have the same mean and SD for the reference and focal groups (Raju, 1988) across all conditions. θs had uniform distributions with an intended mean of θ = 0.0 (intended SD = 1.0) with a range from –3.0 to +3.0.

Table 1 shows the intended means, the actual means, the intended SDs, and the actual SDs of the generated parameters based on each of the 75 items for each level of each effect. Actual (generated) parameters were closer to the intended (true) parameters when the magnitude of the SDs was lower for the distribution of generated parameters. All parameters were used to compute the CSAs and CUAs between the groups, using Kim & Cohen's (1992) IRTDIF program.

Closed-Interval Area Statistics

Kim & Cohen (1991, 1992) defined the area between two points on the θ continuum as

$$S_i(\theta_L,\theta_U) = c_i(\theta_U-\theta_L)+\frac{1-c_i}{Da_i}\ln\left\{\frac{1+\exp\left[Da_i(\theta_U-b_i)\right]}{1+\exp\left[Da_i(\theta_L-b_i)\right]}\right\}, \quad (2)$$

where

S_i is the area between the two θ points,
θ_L is the lower bound θ, and
θ_U is the upper bound θ.

The CSA measure is centered around 0, with any deviation indicating DIF. Kim & Cohen (1991, 1992) defined the CSA statistic as

$$\text{CSA} = S_R(\theta_L,\theta_U) - S_F(\theta_L,\theta_U), \quad (3)$$

where S_R is the reference group area, and S_F is the focal group area. The CUA 0 point indicates no DIF as well. Kim & Cohen (1991, 1992) defined the CUA statistic as

$$\text{CUA} = \left|S_R(\theta_L,\theta_U) - S_F(\theta_L,\theta_U)\right|. \quad (4)$$

When using the CUA method, intensive computation is needed to compute the area between IRFs when the IRFs cross at one or two points (Kim & Cohen, 1991, 1992). When the IRFs cross at one point, Kim & Cohen (1991, 1992) defined the formula for the CUA statistic as

$$\text{CUA} = \left|S_R(\theta_L,\theta_X) - S_F(\theta_L,\theta_X)\right| + \left|S_R(\theta_X,\theta_U) - S_F(\theta_X,\theta_U)\right|, \quad (5)$$

where θ_X is the value of θ at which the IRFs cross.

The Newton-Raphson method can be used to compute the point(s) where the two IRFs cross (Kim & Cohen, 1991). However, a trapezoidal approximation also may be used as a substitute for the Newton-

Raphson algorithm (S. H. Kim, personal communication, July 1, 1993), or when it is difficult to locate where on the θ continuum the IRFs cross (Kim & Cohen, 1992). The formulas for the trapezoidal method as implemented by IRTDIF (Kim & Cohen, 1992) are

$$\mathrm{CUA} = \int_{\theta_L}^{\theta_U} \left| P_R(\theta) - P_F(\theta) \right| d\theta, \tag{6}$$

and its discrete approximation,

$$\mathrm{CUA} = \sum_{j=1}^{n} \left| P_R(\theta) - P_F(\theta) \right| \Delta\theta + \frac{1}{2}\left[\left| P_R(\theta_L) - P_F(\theta_L) \right| - \left| P_R(\theta_U) - P_F(\theta_U) \right| \right] \Delta\theta. \tag{7}$$

The CSA and CUA were the dependent variables in their respective analyses.

Significance Testing of Simple Effects and Contrasts

The 24 cells of the design, with 75 observations (items) per cell, were analyzed by a three-way ANOVA. Table 2 shows the observed means and SDs of CSA and CUA for each of the 24 different conditions. Significance testing by traditional ANOVA procedures was used because of the arbitrariness of the significance of DIF detection by IRT methods, except when variance/covariance matrices are available for the item parameters (Kim & Cohen, 1992; Lord, 1980).

Table 2
Means and SDs of CSA and CUA for Each Condition

	CSA		CUA	
Condition	Mean	SD	Mean	SD
1	.081	1.050	.894	.586
2	.297	.988	.908	.559
3	−.177	1.064	.909	.626
4	.329	.977	.934	.546
5	.350	.996	.902	.567
6	.340	1.018	.950	.532
7	.348	.867	.772	.557
8	.772	.930	1.030	.665
9	.692	1.020	1.014	.727
10	.993	1.012	1.176	.803
11	.742	.807	.933	.601
12	.973	.886	1.124	.712
13	−.193	.977	.885	.547
14	.231	1.022	.991	.554
15	.255	1.006	.894	.609
16	.340	.928	.966	.508
17	.215	1.021	.896	.603
18	.507	1.043	1.010	.617
19	.699	.754	.897	.540
20	.468	1.134	1.034	.743
21	.444	.973	.928	.601
22	.993	1.056	1.225	.810
23	.811	.996	1.128	.685
24	.915	.899	1.091	.715

A multiple comparison procedure was used to identify which levels of the *b*s and *c*s were different from each other within their respective significant parameter effects. The Scheffé test (Howell, 1987; Klockars & Sax, 1986) was selected because it is an a posteriori significance testing procedure; it was used after

main effects and interactions were examined for significance. The comparisons were nonorthogonal because more comparisons were made than the number of levels minus 1. The lack of orthogonality was counteracted by examining combinatory effects of the levels that were interesting, as well as all pairwise comparisons. Not all significant main effects can be interpreted with the Scheffé comparisons, unless all combinatory comparisons are examined. This was not done for the purposes of this study. For *b*, interesting combinatory comparisons were comparing Level B1 to the combination of Levels B2 and B3. For *c*, there were three nonpairwise comparisons—Level C1 was compared to a combination of Levels C2, C3, and C4; Level C1 was compared to Levels C2 and C3; and Level C4 was compared to Levels C2 and C3. A significance level of $\alpha = .05$ was used for all Scheffé tests because it is designed to control for Type I error.

The study also examined whether the *a*s caused DIF in the dependent variables, had a moderating effect, or had no significant role in DIF detection. A simple effects design that examined the *a*s within each of the *b* and *c* effects was used to investigate whether there were any interactions at particular levels of the *b* and/or *c* effects, which included the *a* condition. Simple effects were to be used if the *b* and *c* effects interacted with each other as well. For the purposes of examining the *a*s, significant simple effects might indicate where DIF detection may be a function of high or low *a* test items. A significance level equal to .05 divided by the number of levels of both effects that interacted was specified, a priori, to reduce the number of Type I errors that might arise because of the multiple significance tests for each dependent variable.

Results

CSA

Results of the ANOVA for CSA (Table 3) show significant effects for *b*, *c*, and the $a \times c$ interaction. The means of the CSA statistic for all levels of *b* were .1456, .4624, and .8206, respectively for Levels B1 to B3, and are presented in Table 4. All pairwise Scheffé mean CSA comparisons between levels of *b* (Comparisons 1, 2, and 3), and Scheffé comparisons using a combination of Levels B2 and B3 compared against B1 (Comparison 4), were statistically significant, as shown in Table 5. The mean CSAs for Level B1 significantly differed from Level B2, Level B1 differed from Level B3, and Level B2 significantly differed from Level B3. Also, the no-DIF baseline level (Level B1) was significantly different from the combination of the two other levels of the *b*s (Levels B2 and B3). No combination of levels created any single homogeneous subset of CSA means that were equal to each other.

The CSA means for all levels of *c* are shown in Table 4. Table 6 shows results of the comparisons of those means. Significant differences were observed between C1 and C2 (Comparison 1), C1 and C4 (Comparison 3), C1 versus the other levels of *c* (Comparison 7), and C1 versus C2 and C3 (Comparison 8).

The significant $a \times c$ interaction was investigated with a simple effects model of the differences between the *a* levels within each of the different *c* levels, as shown in Table 7; a significance level of .05/8 = .00625 was used. Level C3 significantly interacted with the *a* effect. Table 8 shows the means for the interaction between *a* and *c*; Figure 1 displays this interaction. For Level A2, differences in the *c* levels differentiated Level C1 from Levels C2, C3, and C4, and there were no differences between Levels C2, C3, and C4. For Level A1, the same relative pattern held as for Level A2, except there was a decrease in the magnitude of the CSA statistic at Level C3.

CUA

For CUA, Table 3 shows significant main effects for *b* and *c* and no significant interactions. Table 9 shows that for *b*, Level B1 was significantly different from Level B3 (Comparison 2) and Level B2 was significantly different from Level B3 (Comparison 3). The means for the *b* levels were B1 = .9226, B2 = .9362, and B3 = 1.0773, as shown in Table 4. Level B1 was not significantly different from the combination of Levels B2 and B3.

Table 3
ANOVA of CSA and CUA for Main Effects and Interactions

Source of Variance	*df*	Mean Square	*F*	*p*
CSA				
a	1	.010	.010	[a]
b	2	68.436	71.316	<.001
c	3	11.573	12.060	<.001
a × *b*	2	.337	.351	[a]
a × *c*	3	5.373	5.599	.001
b × *c*	6	1.084	1.130	.342
a × *b* × *c*	6	.934	.973	[a]
Residual	1,776	.960		
Total	1,799	1.059		
CUA				
a	1	.501	1.258	.262
b	2	4.403	11.047	<.001
c	3	2.023	5.077	.002
a × *b*	2	.020	.051	[a]
a × *c*	3	.416	1.044	.372
b × *c*	6	.377	.945	[a]
a × *b* × *c*	6	.180	.451	[a]
Residual	1,776	.399		
Total	1,799	.405		

[a] $F < 1$.

Table 10 shows that for *c*, none of the Scheffé comparisons were significantly different from each other. The means for the *c* levels were .9197, 1.0434, .9220, and 1.0298, respectively for Levels C1 to C4, as shown in Table 4.

Table 4
Means, SDs, and Number of Items (*I*) for Each Level of Each Effect For CSA and CUA

DIF Statistic and Effect/Level	Mean	SD	*I*
CSA			
A1	.4785	1.0239	900
A2	.4739	1.0346	900
B1	.1456	1.0174	600
B2	.4624	.9873	600
B3	.8206	.9703	600
C1	.2650	1.0391	450
C2	.5605	1.0662	450
C3	.4464	.9807	450
C4	.6329	.9933	450
CUA			
A1	.9620	.6336	900
A2	.9954	.6384	900
B1	.9226	.5658	600
B2	.9362	.6089	600
B3	1.0773	.7138	600
C1	.9197	.6062	450
C2	1.0434	.6631	450
C3	.9220	.6106	450
C4	1.0298	.6539	450

Table 5
Scheffé Comparisons Between CSA Means for the *b* Effect (MS Error = .960)

	b Level				
Comparison	B1	B2	B3	*F*	*p*
1	1	−1	0	7.508	.001
2	1	0	−1	35.070	<.001
3	0	1	−1	10.125	<.001
4	2	−1	−1	25.010	<.001

Discussion

CSA

Mean CSA differed as a function of *b* and *c* differences. All *b*/θ levels significantly differed from each other, demonstrating that moderate differences in DIF for all items (Miller & Oshima, 1992) are detectable by the CSA statistic. The means of the CSA for the *b* levels gradually increased in magnitude as the magnitude of differences between groups increased.

Table 6
Scheffé Comparisons Between CSA Means for the *c* Effect (MS Error = .960)

	c Level					
Comparison	C1	C2	C3	C4	*F*	*p*
1	1	−1	0	0	3.285	.020
2	1	0	−1	0	1.266	.285
3	1	0	0	−1	5.063	.002
4	0	1	−1	0	.473	.701
5	0	1	0	−1	.191	.902
6	0	0	1	−1	1.266	.285
7	3	−1	−1	−1	4.485	.004
8	2	−1	−1	0	2.876	.035
9	0	−1	−1	2	.814	.486

c provided varied results for CSA. The moderate between-group difference levels for *c* indicated DIF when the reference group had a mean *c* of .25 and the focal group had a mean *c* of .20, as compared to the baseline level. However, for the moderate difference level in which the reference group had a mean *c* = .30 and the focal group had a mean *c* = .25, there was no significant difference in mean DIF when compared to the baseline level. DIF was detected, compared to the baseline level, when the reference group had a mean *c* = .35 and the focal group had a mean *c* = .25.

Table 7
Simple Effects Between *a* Levels for Each Level of the *c*s

Source of Variance	*df*	Mean Square	*F*	*p*
a at C1	1	5.400	5.624	.018
a at C2	1	.128	.133	[a]
a at C3	1	9.063	9.440	.002
a at C4	1	1.538	1.603	.206
Residual	1,776	.960		

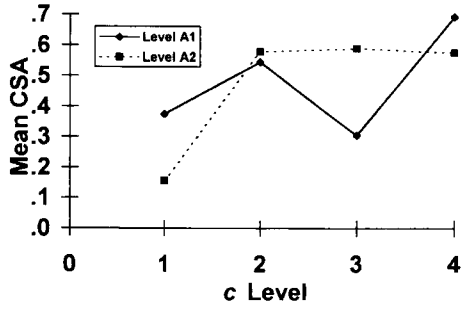
[a] $F < 1$.

Table 8
CSA Means and SDs for Each Level of the *a*s at Each Level of the *c*s, Based on 225 Items

	a Level			
	A1		A2	
c Level	Mean	SD	Mean	SD
C1	.3745	1.0480	.1555	1.0208
C2	.5436	1.0513	.5773	1.0830
C3	.3045	.9899	.5883	.9527
C4	.6914	.9661	.5745	1.0186

Previous DIF studies have demonstrated that highly discriminating items are identified more often as DIF items (Burton & Burton, 1993; Linn, 1993); however, in this study the less discriminating items differed more from the baseline condition for DIF when $c = .25$ for the reference group and $c = .20$ for the focal group. Higher levels of *c* apparently shifted the IRFs enough to result in higher values of CSA, depending on the level of the *a* parameter.

Figure 1
Mean of CSA as a Function of *c* for Levels A1 and A2



CUA

The CUA statistic was able to detect significant differences in the items based on differences in *b* across all items. There was no significant difference between the mean CUA when there were moderate differences between the *b*s, compared with the baseline; however, there was a significant difference in the mean CUA from the baseline for the high *b* condition. The follow-up comparisons also showed that there was a difference between the moderate *b* and high *b* conditions. Large differences between the groups' *b* parameters resulted in larger magnitudes of the CUA statistic.

The *c* levels were significantly different from each other in the main effects model; however, using the

Table 9
Scheffé Comparisons Between CUA Means for the *b* Effect (MS Error = .399)

	b Level				
Comparison	B1	B2	B3	*F*	*p*
1	1	−1	0	.075	.928
2	1	0	−1	4.812	.008
3	0	1	−1	3.684	.025
4	2	−1	−1	2.030	.132

Table 10
Scheffé Comparisons Between CUA Means for the *c* Effect (MS Error = .399)

	c Level					
Comparison	C1	C2	C3	C4	*F*	*p*
1	1	−1	0	0	1.353	.255
2	1	0	−1	0	0.000	.392
3	1	0	0	−1	1.137	.333
4	0	1	−1	0	1.353	.255
5	0	1	0	−1	.009	.999
6	0	0	1	−1	1.137	.333
7	3	−1	−1	−1	.829	.478
8	2	−1	−1	0	.451	.717
9	0	−1	−1	2	.313	.816

Scheffé follow-up comparisons, no differences were found between the CUA means on a priori comparisons. Perhaps additional contrasts would indicate where there were significant differences in the *c* condition.

Higher or lower discriminating items did not affect mean values of CUA. This result conflicts with prior research that suggests that DIF is more easily detectable for higher discriminating items (Donoghue et al., 1993).

Conclusions

The primary result of this study was the different behavior of CUA and CSA as a function of the *a*, *b*, and *c* parameters. The differences between CSA and CUA should not be confused with differences due to differing *a*s between groups (Millsap & Everson, 1993). CSA appeared to be a better indicator of differences between groups than CUA when b/θ or *c* differences are expected. However, the mean CSA differed depending on *a*.

When pervasive bias is suspected, CSA should be used to index DIF. Future research should also examine why CUA provides such varied magnitudes when *c* varies. One interesting note is that CUA is correlated with Lord's χ^2 test of significance (Millsap & Everson, 1993); however, Lord's test does not allow the *c*s to vary.

The discrimination of a set of items should be investigated further to assess how it affects the identification of DIF items. The intricacies of examining how *a* interacts with the *c* and *b* parameters also need to be examined.

A possible follow-up to this investigation should be to examine how DIF detection is affected by the introduction of equating error. Experimental manipulation of pervasive DIF would be informative as well. Standard errors surrounding each IRF should be examined under circumstances similar to those of this design (Linn & Levine, 1981). Real empirical parameter estimation procedures also should be examined to determine if they moderate or exaggerate measures of DIF for the closed-interval methods.

References

Angoff, W. H. (1993). Perspectives on differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale NJ: Erlbaum.

Bock, R. D. (1993). Different DIFs: Comment on the papers read by Neil Dorans and David Thissen. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 115–122). Hillsdale NJ: Erlbaum.

Burton, E., & Burton, N. W. (1993). The effects of item screening on test scores and test characteristics. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 321–335). Hillsdale NJ: Erlbaum.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), *Differential item function-*

ing (pp. 397–413). Hillsdale NJ: Erlbaum.

Cohen, A. S., Kim, S.-H., & Subkoviak, M. J. (1991). Influence of prior distributions on detection of DIF. *Journal of Educational Measurement, 28,* 49–59.

Cole, N. S. (1993). History and development of DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 25–29). Hillsdale NJ: Erlbaum.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A monte-carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137–166). Hillsdale NJ: Erlbaum.

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: Erlbaum.

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

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning.* Hillsdale NJ: Erlbaum.

Howell, D. C. (1987). *Statistical methods for psychology* (2nd ed.). Boston: PWS-KENT.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement, 16,* 209–225.

Kim, S.-H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement, 15,* 269–278.

Kim, S.-H., & Cohen, A. S. (1992). IRTDIF: A computer program for IRT differential item functioning analysis. *Applied Psychological Measurement, 16,* 158.

Klockars, A. J., & Sax, G. (1986). *Multiple comparisons.* Newbury Park CA: Sage.

Linn, R. L. (1993). The use of DIF statistics: A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349–364). Hillsdale NJ: Erlbaum.

Linn, R. L., & Levine, M. V. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5,* 159–173.

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

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

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

Osterlind, S. J. (1983). *Test item bias.* Beverly Hills: Sage.

Press, W. H., Flannery, B. P., Teukoisky, S. A., & Vetterling. W. T. (1992). *Numerical recipes in C: The art of scientific computing* (2nd ed.). Cambridge, England: Cambridge University Press.

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

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). A monte carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement, 17,* 1–10.

Shealy, R., & Stout, W. (1993a). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159–194.

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

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale NJ: Erlbaum.

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123–135). Hillsdale NJ: Erlbaum.

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: Erlbaum.

Acknowledgments

Thanks to Jane Schleisman for helping create the tables, Mike Finger for help with multiple comparison procedures, John DeWitt for help with random number generators, Seock-Ho Kim for guidance, and two anonymous reviewers whose comments helped clarify the purpose of this paper.

Author's Address

Send requests for reprints or further information to Zachary S. Feinstein, N650 Elliott Hall, Department of Psychology, University of Minnesota, Minneapolis MN 55455, U.S.A. Internet: fein0001@gold.tc.umn.edu.