A Comparison of Two Area Measures for Detecting Differential Item Functioning

Seock-Ho Kim and Allan S. Cohen
University of Wisconsin

The area between two item response functions is often used as a measure of differential item functioning under item response theory. This area can be measured over either an open interval (i.e., exact) or closed interval. Formulas are presented for computing the closed-interval signed and unsigned areas. Exact and closed-interval measures were estimated on data from a test with embedded items intentionally constructed to favor one group over another. No real differences in detection of these items were found between exact and closed-interval methods. *Index terms: BILOG, closed interval, differential item functioning, item response functions, open interval, signed area, unsigned area.*

The closed-interval area between two item response functions (IRFs) from two different groups has been used in a number of studies as a measure of differential item functioning (DIF) (Ironson & Subkoviak, 1979; Linn, Levine, Hastings, & Wardrop, 1981; McCauley & Mendoza, 1985; Rudner, Getson, & Knight, 1980; Shepard, Camilli, & Averill, 1981; Shepard, Camilli, & Williams, 1984; 1985). The use of the closed interval assumes that the limits of the interval define the region of the θ scale that is of interest; generally, this is the portion with the most examinees. It is typically not of interest to examine differences in area at the extremes of the scale. Consequently, there seems to be little reason for setting the limits of the closed interval much beyond ± 4 on the θ scale.

Raju (1988) suggested, however, that the particular set of limits is often arbitrary and that another set of limits might be selected that would be finite, yet different from the first, and for which the area would also be different. Such arbitrariness, Raju noted, should be removed by integrating θ over the entire scale to obtain a measure of the exact area between the two IRFs. Although this argument seems compelling, Raju illustrated the differences between the exact and the closed-interval areas for only a single item. If the exact method is in fact superior, then it should detect DIF better than the closed-interval method over all the items in a test. This paper presents a comparison of the exact and the closed-interval area measures on a set of actual test data. The data contained a set of items that were intentionally constructed to favor one group of examinees over the other group in the sample.

Signed and Unsigned Areas Over a Closed Interval

Raju (1988) presented general equations for two exact area measures—the exact signed area (ESA) and the exact unsigned area (EUA)—between two IRFs for the one- (1PM), two- (2PM), and three-parameter (3PM) models. Rudner (1977) and Rudner et al. (1980) described a method of summing the differences between two IRFs over successive intervals of width .005 between two finite points on the θ scale. Although Rudner's summing method is conceptually simple and easy to implement for a computer program, computing area in this way can take substantial computing time. The closed-interval area between two IRFs can also be obtained from an integration method that results in more

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 15, No. 3, September 1991, pp. 269-278

0146-6216/91/030269-10$1.75

accurate estimates and in considerable reduction of computing time. A special case of this method has been noted by Shepard et al. (1981), but no details of the method were given.

For completeness, therefore, the general equations are presented below for the closed-interval area measures—the closed-interval signed area (CSA) and the closed-interval unsigned area (CUA)—between two IRFs for the 1PM, 2PM, and 3PM. These formulas are easily contrasted with the exact area formulas given by Raju.

The IRF of the 3PM for an item is given by

$$P(\theta) = c + (1 - c)P^*(\theta) \quad , \tag{1}$$

where

$$P^*(\theta) = \{1 + \exp[-Da(\theta - b)]\}^{-1} \quad , \tag{2}$$

a, b, and c are parameters characterizing the item, and D is a scaling constant, usually set to 1 or 1.7.

The area under the IRF between two finite points $[\theta_1, \theta_2]$ on the θ scale can be written as

$$S(\theta_1,\theta_2) = \int_{\theta_1}^{\theta_2} P(\theta)d\theta = c(\theta_2 - \theta_1) + (1 - c)(Da)^{-1}\ln\left\{\frac{\{1 + \exp[Da(\theta_2 - b)]\}}{\{1 + \exp[Da(\theta_1 - b)]\}}\right\} \quad . \tag{3}$$

In a DIF study, there are two sets of item parameters for each item: (a_R, b_R, c_R) from the reference group and (a_F, b_F, c_F) from the focal group. The CSA and CUA between two IRFs in the interval $[\theta_1, \theta_2]$ can take the following general forms:

$$\text{CSA} = \int_{\theta_1}^{\theta_2} [P_R(\theta) - P_F(\theta)]d\theta = S_R(\theta_1,\theta_2) - S_F(\theta_1,\theta_2) \tag{4}$$

and

$$\text{CUA} = \int_{\theta_1}^{\theta_2} |P_R(\theta) - P_F(\theta)|d\theta \quad . \tag{5}$$

One-Parameter (Rasch) Model

The 1PM (Wright, 1977) is the special case of the 3PM when $D = 1$, $a_R = a_F = 1$, and $c_R = c_F = 0$ for all items. Consequently, there is no point on the θ scale where the two IRFs cross each other.

The CSA and CUA between two IRFs are defined by

$$\text{CSA} = S_R(\theta_1,\theta_2) - S_F(\theta_1,\theta_2) = \ln\left\{\frac{[1 + \exp(\theta_2 - b_R)][1 + \exp(\theta_1 - b_F)]}{[1 + \exp(\theta_1 - b_R)][1 + \exp(\theta_2 - b_F)]}\right\} \tag{6}$$

and

$$\text{CUA} = |\text{CSA}(6)| \quad , \tag{7}$$

where CSA(6) denotes the CSA as defined in Equation 6.

Two-Parameter Model

When the 2PM is used to estimate item parameters, $c_R = c_F = 0$, there are two cases: *Case I* in which $a_R = a_F = a$ and *Case II* in which $a_R \neq a_F$. For either case, the CSA can be expressed as

$$\text{CSA} = S_R(\theta_1,\theta_2) - S_F(\theta_1,\theta_2) = \ln\left\{\frac{\{1 + \exp[Da_R(\theta_2 - b_R)]\}^{1/Da_R}\{1 + \exp[Da_F(\theta_1 - b_F)]\}^{1/Da_F}}{\{1 + \exp[Da_R(\theta_1 - b_R)]\}^{1/Da_R}\{1 + \exp[Da_F(\theta_2 - b_F)]\}^{1/Da_F}}\right\} \quad . \tag{8}$$

Case I. There is no finite intersection point on the θ scale because $a_R = a_F = a$. Thus, the CSA can be simplified as

$$\text{CSA} = (Da)^{-1}\ln\left\{\frac{\{1 + \exp[Da(\theta_2 - b_R)]\}\{1 + \exp[Da(\theta_1 - b_F)]\}}{\{1 + \exp[Da(\theta_1 - b_R)]\}\{1 + \exp[Da(\theta_2 - b_F)]\}}\right\} . \quad (9)$$

The CUA between two IRFs has the form

$$\text{CUA} = |\text{CSA}(9)| . \quad (10)$$

Case II. Because $a_R \neq a_F$, the crossing point of the two IRFs is

$$\theta_x = \frac{a_R b_R - a_F b_F}{a_R - a_F} . \quad (11)$$

When θ_x is located outside the interval $[\theta_1, \theta_2]$, the CUA is defined as

$$\text{CUA} = |\text{CSA}(8)| . \quad (12)$$

When θ_x is found within the interval $[\theta_1, \theta_2]$, the CUA is defined as

$$\begin{aligned}\text{CUA} &= |S_R(\theta_1,\theta_x) - S_F(\theta_1,\theta_x)| + |S_R(\theta_x,\theta_2) - S_F(\theta_x,\theta_2)| \\ &= \left|\ln\left\{\frac{1 + \exp[Da_R(\theta_x - b_R)]\}^{1/Da_R}\{1 + \exp[Da_F(\theta_1 - b_F)]\}^{1/Da_F}}{\{1 + \exp[Da_R(\theta_1 - b_R)]\}^{1/Da_R}\{1 + \exp[Da_F(\theta_x - b_F)]\}^{1/Da_F}}\right\}\right| \\ &+ \left|\ln\left\{\frac{1 + \exp[Da_R(\theta_2 - b_R)]\}^{1/Da_R}\{1 + \exp[Da_F(\theta_x - b_F)]\}^{1/Da_F}}{\{1 + \exp[Da_R(\theta_x - b_R)]\}^{1/Da_R}\{1 + \exp[Da_F(\theta_2 - b_F)]\}^{1/Da_F}}\right\}\right| . \end{aligned} \quad (13)$$

Three-Parameter Model

When the 3PM is used to estimate item parameters, there are four possible cases: *Case I* in which $c_R = c_F = c$ and $a_R = a_F = a$; *Case II* in which $c_R = c_F = c$ and $a_R \neq a_F$; *Case III* when $c_R \neq c_F$ and $a_R = a_F = a$; and *Case IV* in which $c_R \neq c_F$ and $a_R \neq a_F$. *Case I* and *Case II* are essentially the same as for the 2PM.

CSA is obtained in the same way for all four cases and can be expressed as

$$\begin{aligned}\text{CSA} &= S_R(\theta_1,\theta_2) - S_F(\theta_1,\theta_2) \\ &= (c_R - c_F)(\theta_2 - \theta_1) + \ln\left\{\frac{\{1 + \exp[Da_R(\theta_2 - b_R)]\}^{(1-c_R)/Da_R}\{1 + \exp[Da_F(\theta_1 - b_F)]\}^{(1-c_F)/Da_F}}{\{1 + \exp[Da_R(\theta_1 - b_R)]\}^{(1-c_R)/Da_R}\{1 + \exp[Da_F(\theta_2 - b_F)]\}^{(1-c_F)/Da_F}}\right\} . \end{aligned} \quad (14)$$

Case I. Because $c_R = c_F = c$ and $a_R = a_F = a$, there is no finite point on the θ scale where the two IRFs cross each other. The CSA can be simplified as

$$\text{CSA} = (1 - c)(Da)^{-1}\ln\left\{\frac{\{1 + \exp[Da(\theta_2 - b_R)]\}\{1 + \exp[Da(\theta_1 - b_F)]\}}{\{1 + \exp[Da(\theta_1 - b_R)]\}\{1 + \exp[Da(\theta_2 - b_F)]\}}\right\} . \quad (15)$$

The CUA between the two IRFs has the form

$$\text{CUA} = |\text{CSA}(15)| . \quad (16)$$

Case II. In this case the crossing point of the two IRFs can be obtained from Equation 11. When θ_x is located outside the interval $[\theta_1, \theta_2]$, CUA is defined as

$$\text{CUA} = |\text{CSA}(14)| \quad . \tag{17}$$

When θ_x is located inside the interval, CUA is defined as

$$\begin{aligned}\text{CUA} &= |S_R(\theta_1,\theta_x) - S_F(\theta_1,\theta_x)| + |S_R(\theta_x,\theta_2) - S_F(\theta_x,\theta_2)| \\ &= (1-c)\left|\ln\left\{\frac{\{1+\exp[Da_R(\theta_x-b_R)]\}^{1/Da_R}\{1+\exp[Da_F(\theta_1-b_F)]\}^{1/Da_F}}{\{1+\exp[Da_R(\theta_1-b_R)]\}^{1/Da_R}\{1+\exp[Da_F(\theta_x-b_F)]\}^{1/Da_F}}\right\}\right| \\ &+ (1-c)\left|\ln\left\{\frac{\{1+\exp[Da_R(\theta_2-b_R)]\}^{1/Da_R}\{1+\exp[Da_F(\theta_x-b_F)]\}^{1/Da_F}}{\{1+\exp[Da_R(\theta_x-b_R)]\}^{1/Da_R}\{1+\exp[Da_F(\theta_2-b_F)]\}^{1/Da_F}}\right\}\right| \quad .\end{aligned} \tag{18}$$

Case III. Because $a_R = a_F = a$, the crossing point of the two IRFs is

$$\theta_x = (a)^{-1}\ln\left\{\frac{c_R - c_F}{[(1-c_R)/\exp(ab_F)] - [(1-c_F)/\exp(ab_R)]}\right\} \quad . \tag{19}$$

When θ_x is located outside the interval $[\theta_1, \theta_2]$, CUA is defined as

$$\text{CUA} = |\text{CSA}(14)| \quad . \tag{20}$$

If θ_x does not exist, that is, either

$$\frac{c_R - c_F}{[(1-c_R)/\exp(ab_F)] - [(1-c_F)/\exp(ab_R)]} \leq 0 \quad , \tag{21}$$

$c_R > c_F$, and $b_R \leq b_F$; or $c_R < c_F$ and $b_R \geq b_F$; then there is no finite crossing point on the θ scale. In these special situations, CUA is defined as

$$\text{CUA} = |\text{CSA}(14)| \quad . \tag{22}$$

When θ_x exists and is located inside the interval $[\theta_1, \theta_2]$, CUA is defined as:

$$\begin{aligned}\text{CUA} &= |S_R(\theta_1,\theta_x) - S_F(\theta_1,\theta_x)| + |S_R(\theta_x,\theta_2) - S_F(\theta_x,\theta_2)| \\ &= \left|(c_R-c_F)(\theta_x-\theta_1) + \ln\left\{\frac{\{1+\exp[Da(\theta_x-b_R)]\}^{(1-c_R)/Da}\{1+\exp[Da(\theta_1-b_F)]\}^{(1-c_F)/Da}}{\{1+\exp[Da(\theta_1-b_R)]\}^{(1-c_R)/Da}\{1+\exp[Da(\theta_x-b_F)]\}^{(1-c_F)/Da}}\right\}\right| \\ &+ \left|(c_R-c_F)(\theta_2-\theta_x) + \ln\left\{\frac{\{1+\exp[Da(\theta_2-b_R)]\}^{(1-c_R)/Da}\{1+\exp[Da(\theta_x-b_F)]\}^{(1-c_F)/Da}}{\{1+\exp[Da(\theta_x-b_R)]\}^{(1-c_R)/Da}\{1+\exp[Da(\theta_2-b_F)]\}^{(1-c_F)/Da}}\right\}\right| \quad .\end{aligned} \tag{23}$$

Case IV. In this case the Newton-Raphson method of approximation (Burden & Faires, 1985) is used to find the crossing points. There may be either 0, 1, or 2 crossing points on the θ scale.

For the Newton-Raphson procedure, let

$$f(\theta) = P_R(\theta) - P_F(\theta) \quad , \tag{24}$$

then

$$f'(\theta) = \frac{df(\theta)}{d\theta} = (1-c_R)Da_R P_R^*(1-P_R^*) - (1-c_F)Da_F P_F^*(1-P_F^*) \quad , \tag{25}$$

where P^* is defined in Equation 2. By giving an initial value θ_0, the crossing point θ_x can be

approximated using

$$\theta_{n+1} = \theta_n - \frac{f(\theta_n)}{f'(\theta_n)} \quad , \tag{26}$$

where $f'(\theta_n)$ is the derivative of $f(\theta)$ at θ_n.

When θ_xs are located outside the interval $[\theta_1, \theta_2]$, the CUA can be written as

$$\text{CUA} = |\text{CSA}(14)| \quad . \tag{27}$$

When there is only one crossing point, θ_x, within the interval, CUA is defined as

$$\begin{aligned}\text{CUA} &= |S_R(\theta_1,\theta_x) - S_F(\theta_1,\theta_x)| + |S_R(\theta_x,\theta_2) - S_F(\theta_x,\theta_2)| \\ &= \left| (c_R - c_F)(\theta_x - \theta_1) + \ln\left\{\frac{\{1+\exp[Da_R(\theta_x - b_R)]\}^{(1-c_R)/Da_R}\{1+\exp[Da_F(\theta_1 - b_F)]\}^{(1-c_F)/Da_F}}{\{1+\exp[Da_R(\theta_1 - b_R)]\}^{(1-c_R)/Da_R}\{1+\exp[Da_F(\theta_x - b_F)]\}^{(1-c_F)/Da_F}}\right\}\right| \\ &+ \left| (c_R - c_F)(\theta_2 - \theta_x) + \ln\left\{\frac{\{1+\exp[Da_R(\theta_2 - b_R)]\}^{(1-c_R)/Da_R}\{1+\exp[Da_F(\theta_x - b_F)]\}^{(1-c_F)/Da_F}}{\{1+\exp[Da_R(\theta_x - b_R)]\}^{(1-c_R)/Da_R}\{1+\exp[Da_F(\theta_2 - b_F)]\}^{(1-c_F)/Da_R}}\right\}\right| \quad . \end{aligned} \tag{28}$$

If there are two crossing points, θ_{x_1} and θ_{x_2} ($\theta_{x_1} < \theta_{x_2}$) within the interval $[\theta_1, \theta_2]$, the CUA is

$$\begin{aligned}\text{CUA} &= |S_R(\theta_1,\theta_{x_1}) - S_F(\theta_1,\theta_{x_1})| + |S_R(\theta_{x_1},\theta_{x_2}) - S_F(\theta_{x_1},\theta_{x_2})| + |S_R(\theta_{x_2},\theta_2) - S_F(\theta_{x_2},\theta_2)| \\ &= \left| (c_R - c_F)(\theta_{x_1} - \theta_1) + \ln\left\{\frac{\{1+\exp[Da_R(\theta_{x_1} - b_R)]\}^{(1-c_R)/Da_R}\{1+\exp[Da_F(\theta_1 - b_F)]\}^{(1-c_F)/Da_F}}{\{1+\exp[Da_R(\theta_1 - b_R)]\}^{(1-c_R)/Da_R}\{1+\exp[Da_F(\theta_{x_1} - b_F)]\}^{(1-c_F)/Da_F}}\right\}\right| \\ &+ \left| (c_R - c_F)(\theta_{x_2} - \theta_{x_1}) + \ln\left\{\frac{\{1+\exp[Da_R(\theta_{x_2} - b_R)]\}^{(1-c_R)/Da_R}\{1+\exp[Da_F(\theta_{x_1} - b_F)]\}^{(1-c_F)/Da_F}}{\{1+\exp[Da_R(\theta_{x_1} - b_R)]\}^{(1-c_R)/Da_R}\{1+\exp[Da_F(\theta_{x_2} - b_F)]\}^{(1-c_F)/Da_F}}\right\}\right| \\ &+ \left| (c_R - c_F)(\theta_2 - \theta_{x_2}) + \ln\left\{\frac{\{1+\exp[Da_R(\theta_2 - b_R)]\}^{(1-c_R)/Da_R}\{1+\exp[Da_F(\theta_{x_2} - b_F)]\}^{(1-c_F)/Da_F}}{\{1+\exp[Da_R(\theta_2 - b_R)]\}^{(1-c_R)/Da_R}\{1+\exp[Da_F(\theta_{x_2} - b_F)]\}^{(1-c_F)/Da_F}}\right\}\right| . \end{aligned} \tag{29}$$

Method

Data

Data from Subkoviak, Mack, Ironson, and Craig (1984) were reanalyzed for the purposes of this study. The Subkoviak et al. (1984) data consisted of two samples: 1,008 Black and 1,021 White college students. Because of limitations on the microcomputer version of BILOG (Mislevy & Bock, 1986), two random subsets were selected consisting of 1,000 examinees each from the White and Black groups, respectively.

The instrument was a 50-item vocabulary test with four choices for each item. Examinees were asked to select an option that had the same meaning as the stem. In the test, 40 of the items contained standard American English vocabulary words and were drawn from the Verbal Section of the College Qualification Test (Psychological Corporation, 1956). The remaining 10 items were intentionally constructed to be differentially easier for Black examinees. One item with this intentional DIF was inserted randomly into each consecutive block of five items on the test.

Estimation of Item Parameters

Selection of an item response theory (IRT) model. Application of an IRT model requires the data

to be unidimensional. Using Reckase's (1979) suggested criterion, the contribution of the first component from the unrotated solution of a principal components analysis using tetrachoric correlations indicated sufficient unidimensionality for use of an IRT model for the White data (26%) and was marginal for the Black data (16%).

Most work in detection of DIF under IRT has focused on use of the 2PM or the 3PM. Difficulties, such as estimation of the c parameter of the 3PM, have been noted (Baker, 1986; Shepard et al., 1981). To some extent, difficulty in estimation of c can be averted by using a modification of the 3PM in which c is fixed for all items. As Raju (1988) suggested, this version of the 3PM is more appropriate for estimates of exact areas, because when $c_F \neq c_R$ the exact area is infinite. For purposes of this study, both the 3PM and the 3PM with a fixed or equal c parameter (3PM-c) were used.

Estimation of a common c. Estimation of item parameters was done using PC-BILOG (Mislevy & Bock, 1986). PC-BILOG implements a marginal Bayesian estimation procedure—marginal maximum a posteriori estimation (Bock & Aitkin, 1981). A common or fixed c was estimated in the following way: First, a sample of 1,000 cases (500 Black and 500 White examinees) was randomly drawn from the Black and White datasets, and parameters were estimated for the 3PM. The average c for this dataset was .23. Next, the 3PM-c with a value of $c = .23$ was estimated in each dataset separately.

Area Measures

Under IRT, in the absence of DIF, the IRFs computed from each group of examinees will be identical and the area between the two IRFs, after being placed on the same metric, will be 0. In this study, all item parameters for both groups were placed on the same scale using the test characteristic curve method (Stocking & Lord, 1983) as implemented in program EQUATE (Baker, 1990).

Four area measures were computed from the sets of equated item parameters: CSA(14); CUA(17) or CUA(18); ESA—Equation 7 in Raju (1988); and EUA—Equation 8 or Equation 24 in Raju (1988). In the present context, Raju's ESA and EUA can be rewritten as

$$\text{ESA} = (1 - c)(b_F - b_R) \quad , \tag{30}$$

and

$$\text{EUA} = (1 - c)\left| \frac{2(a_F - a_R)}{Da_R a_F} \ln\left\{1 + \exp\left[\frac{Da_R a_F (b_F - b_R)}{a_F - a_R}\right]\right\} - (b_F - b_R)\right| \tag{31}$$

if $a_R \neq a_F$,

or

$$\text{EUA} = (1 - c)|b_F - b_R|, \quad \text{if } a_R = a_F. \tag{32}$$

When calculating signed area measures, the Black group was treated as the reference group and the White group as the focal group. Hence a positive value for both CSA and ESA indicated an item favoring the Black group.

For purposes of analysis, the 10 items with intentional DIF were coded 1, and the 40 standard vocabulary items were coded 0. Point-biserial correlations were then computed between the 0-1 coding of the items and each of the four area measures to determine the success of each area measure in detecting the items with intentional DIF. In addition, intercorrelations among the area measures were computed as an indication of the similarity between area measures.

Error Rates in Detection of DIF

Error rates in detection of a priori DIF in the experimentally manipulated items were also estimated. First, it was assumed that each area measure was normally distributed, and a mean and standard deviation for that area measure was estimated based on the 40 nonmanipulated items. Next, two critical values were selected using one-tailed .05 and .01 levels of significance. Area measures greater than the critical value were identified as DIF items. Classification of items in this way resulted in two types of errors: false negatives and false positives. False negatives occurred when items containing a priori DIF had an area measure smaller than the critical value. Items containing no a priori DIF but identified as having an area measure larger than the critical value were considered false positives. Clearly, with $p = .05$, more false positive identifications are likely to occur than with $p = .01$.

Results

Detection of A Priori DIF

Point-biserial correlations of a priori DIF indices with each of the four area measures are given in Table 1. Correlations with a priori DIF for signed area measures were essentially the same for both the closed ($r = .745$) and exact ($r = .743$) areas. A similar result occurred for unsigned areas ($r = .795$ and $r = .807$, respectively). This suggests that there was probably no difference between the exact and closed-interval areas with respect to detection of DIF in these items. Correlations between a priori DIF and the unsigned area measures were slightly larger than for the signed measures, suggesting that the unsigned area measure, whether exact or closed interval, was more sensitive to detection of the a priori DIF.

Relationships Among Area Measures

The correlation between the closed-interval measures (given in Table 1) was lower ($r = .861$) than between the exact measures ($r = .932$). This suggests that exact measures were more alike than were closed-interval measures. There was substantial agreement, however, between closed-interval and exact measures for both signed area measures ($r = .985$) and unsigned area measures ($r = .966$). This indicated that the two methods, exact and closed interval, provided very similar information.

Error Rates in Detection of DIF

Table 2 presents error rates in detection of DIF for each area measure. In terms of type of error,

Table 1
Correlations of A Priori DIF Indices and Area Measures

Area Measures	A Priori DIF[a]	3PM-*c*				3PM
		CSA	CUA	ESA	EUA	CSA
3PM-*c*						
CSA	.745					
CUA	.795	.861				
ESA	.743	.985	.882			
EUA	.807	.895	.966	.932		
3PM						
CSA	.870	.932	.870	.914	.885	
CUA	.756	.851	.976	.869	.933	.842

[a]Correlations with a priori DIF are point-biserials.

Table 2
False Positive (FP) and False Negative (FN)
Classification Errors in the Detection of DIF

Significance Level	3PM-*c*								3PM			
	CSA		CUA		ESA		EUA		CSA		CUA	
	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
.05	2	2	2	2	2	2	3	1	2	0	4	2
.01	1	2	1	2	1	2	1	2	0	0	2	2

about the same numbers of false positive errors were made whether closed-interval or exact methods were used. Seven of the 10 items with a priori DIF were detected correctly by all four area measures. In addition, 9 of the 10 items were correctly detected by the combined use of the signed and unsigned closed-interval area measures. All 10 of the items were detected by the combined use of the signed and unsigned exact areas.

Five items were falsely identified by one of the four measures. Of these items, four were identified by either the signed or the unsigned version of both the exact and closed-interval area measures. The closed-interval method was slightly less prone to making an error of this type.

Comparison of 3PM and 3PM-*c* Area Measures

The exact area method is relatively easy to estimate in comparison to the closed-interval method. As Raju (1988) has noted, the choice of the limits on the closed interval is arbitrary. However, a major disadvantage of the exact area measure is that it cannot be estimated when $c_F \neq c_R$. Under these conditions, if 3PM item parameter estimates are to be used, the 3PM-*c* usually needs to be used. A disadvantage of the 3PM-*c* is that it requires two calibration runs: the first to estimate the common *c* and the second to estimate the 3PM-*c*. The 3PM, however, only requires a single run and, therefore, is somewhat more convenient to use. In view of these advantages and disadvantages, it was of interest to determine whether differences occurred in detection of a priori DIF or in classification of items as functioning differentially, when the closed-interval measure was estimated for the 3PM as compared to the exact or closed-interval measures for the 3PM-*c*.

Detection of a priori DIF. Point-biserial correlations between the index of a priori DIF and the area measure are given in Table 1. The correlation of .870 with a priori DIF for CSA for the 3PM (3PM-CSA) is slightly higher than similar correlations for the other area measures. This would suggest that the 3PM-CSA is slightly more sensitive than any of the other measures. The correlation for CUA ($r = .756$), however, is at the low end of the values reported for the other four area measures.

Relationships among 3PM and 3PM-c area measures. The correlations in Table 1 indicate that the signed area measures are more similar to each other than they are to the unsigned measures. Similarly, the unsigned measures are more strongly related to one another than they are to the signed measures. The magnitudes of the correlations in Table 1, however, indicate a relatively strong set of relationships among all the area measures, whether based on the 3PM or 3PM-*c* (range = .842 to .985). This suggests that much of the same information might be available regardless of whether area measures were obtained from 3PM or 3PM-*c* item parameter estimates.

Comparison of classification error rates. The error rates for the closed-interval measures are given in Table 2 along with those for the measures based on the 3PM-*c*. The rates for the 3PM-CSA were lower than for any of the other measures. No errors were made by 3PM-CSA at the .01 level and two false positive errors were made at the .05 level. These results indicate a relative superiority for the 3PM-CSA over all other measures used in this study. The 3PM-CUA error rate, however, was slightly

higher than the other area measures, particularly at the .05 level.

With respect to the specific items that were misclassified, no overlap occurred in the false positives made by the 3PM-CSA and 3PM-CUA. Consequently, combined use of the two closed-interval measures at the .05 level would have resulted in making a total of six false positive errors. False negative errors are generally more serious in DIF studies. Therefore, a test developer would want to use a more conservative level of significance in order to decrease the rate of false negative errors. As might be anticipated, and as can be seen from the data in Table 2, the more conservative the level of significance, the greater the number of false positive errors, regardless of which area measure is used.

Discussion

An important benefit from the present study is that the comparisons of these area measures were done in the context of real, experimentally manipulated data as opposed to either nonmanipulated or simulated data. Studies that seek to examine the presence of DIF in nonmanipulated datasets do not necessarily provide the best results because the presence of DIF is only inferred after the data are analyzed. Data simulations can provide an important means of testing hypotheses about the behavior of statistics. However, results based on simulated data suffer from the fact that the simulation data are really only generalizable to data that meet the conditions under which the data were generated (Lord, 1980).

The results of this study were encouraging because they indicated that detection of a priori DIF is good using any of the methods in this study. The best detection occurred with the closed-interval signed method based on 3PM parameter estimates. This was evident in the correlations between the area measures and the a priori DIF indices, and also in the classification errors.

For purposes of this study, identification of non-DIF items as functioning differentially was labeled a false positive. In actuality, these items may have been functioning differentially; they were only errors in the sense that the items had not been manipulated experimentally to contain a priori DIF. Error rates were relatively low for both exact and closed-interval methods, suggesting that both methods were sensitive to this kind of DIF. The number of false negative errors decreased as the critical value became more conservative. By the same token, the number of false positives increased as the criterion became more conservative.

The results of the study indicate little difference in detection of DIF between the two methods for computing area. Choice of one method or the other, therefore, seems to be a matter of reasonable indifference. The computational simplicity of the exact method seems to give it an advantage over the slightly more complicated computations required for the closed-interval methods. Further, the exact method does not introduce an arbitrariness into the area estimation problem. One difficulty with the exact method, however, occurs with the 3PM when unequal c parameters are present. Under such conditions, the exact method yields infinite areas. If a three-parameter model is to be used with the exact method, then a common c must first be estimated before items can be calibrated with the 3PM-c. This is really only a minor nuisance given the speed of modern computers. However, because the 3PM does not require an arbitrary assumption that all the items have an equal lower asymptote as does the 3PM-c, and assuming the model itself is appropriate for the data, in view of the relative similarity in the results from both 3PM and 3PM-c procedures, the closed-interval method may be the method of choice.

References

Baker, F. B. (1986, April). *Two parameter: The forgotten model.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco CA.

Baker, F. B. (1990). *EQUATE: Computer program for equating two metrics in item response theory* [Computer program]. Madison WI: University of Wisconsin, Laboratory of Experimental Design.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Burden, R. L., & Faires, J. D. (1985). *Numerical analysis* (3rd ed). Boston MA: Prindle, Weber, Schmidt.

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

Linn, R. L., Levine, M. V., & Hastings, C. N., & Wardrop, J. L. (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.

McCauley, C. D., & Mendoza, J. (1985). A simulation study of item bias using a two-parameter item response model. *Applied Psychological Measurement, 9,* 389–400.

Mislevy, R. J., & Bock, R. D. (1986). *PC-BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville IN: Scientific Software.

Psychological Corporation. (1956). *College Qualification Test.* New York: Author.

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

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230.

Rudner, L. M. (1977, April). *An approach to biased item identification using latent trait measurement theory.* Paper presented at the annual meeting of the American Educational Research Association, New York.

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.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6,* 317–375.

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9,* 93–128.

Shepard, L., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22,* 77–105.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Craig, R. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. *Journal of Educational Measurement, 21,* 49–58.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97–116.

Acknowledgments

The authors thank M. J. Subkoviak for permission to use his data for this study.

Author's Address

Send requests for reprints or further information to Seock-Ho Kim, Testing and Evaluation Services, University of Wisconsin, 1025 W. Johnson St., Madison WI 53706, U.S.A.