Detection of Differential Item Functioning in the Graded Response Model

Allan S. Cohen, Seock-Ho Kim, and Frank B. Baker

University of Wisconsin

Methods for detecting differential item functioning (DIF) have been proposed primarily for the item response theory dichotomous response model. Three measures of DIF for the dichotomous response model are extended to include Samejima's graded response model: two measures based on area differences between item true score functions, and a χ^2 statistic for comparing differences in item parameters. An illustrative example is presented. *Index terms: differential item functioning, graded response model, item response theory.*

Item response theory (IRT) methods for detecting differential item functioning (DIF) have been proposed primarily for the dichotomous IRT model in which an item is scored as correct or incorrect. The dichotomous case is the IRT model used for most achievement and ability testing. Fewer methods for detecting DIF for the graded response IRT model have been proposed.

Under IRT, the parameters of the item are assumed to be invariant over samples of examinees drawn from the same population. When invariance does not hold for item parameters, the item is said to be functioning differentially; that is, the item is a *DIF item.* Such items are of concern because they present a potential threat to the validity of the test.

DIF studies under IRT require that item parameters for a set of items be estimated in two groups. The reference group (R) is usually a majority group, and the focal group (F) is usually a minority group. The item parameter estimates must be expressed in the same metric before any comparisons can be made between item response functions (IRFs).

Current methods for DIF detection in the IRT dichotomous model consist of two general approaches. In the first approach, the area between IRFs is estimated and tested using two Z tests proposed by Raju (1990). In the second approach, parameters of IRFs estimated in different samples are compared. Examples of this include the χ^2 described by Lord (1980) and the likelihood ratio test described by Thissen, Steinberg, & Wainer (1988). It is important to differentiate between the detection of DIF and the impact of DIF. DIF is detected either by area or item parameter differences; impact is described by determining who is affected by the DIF. This paper focuses on the detection of DIF.

The transformation or linking of the metric obtained in the focal group to that of the reference group can be affected by the presence of DIF; therefore, the resulting link between the two metrics may be incorrect. If this occurs, errors may result in the detection of DIF (Shepard, Camilli, & Williams, 1984). Recent evidence (Baker & Al-Karni, 1991; Kim & Cohen, 1992) suggests that the test characteristic curve (TCC) method for linking (Stocking & Lord, 1983) may be more accurate for small samples than a mean method (Loyd & Hoover, 1980), a weighted mean and sigma method (Linn, Levine, Hastings, & Wardrop, 1981), or the minimum χ^2 method (Divgi, 1985). Baker (1992) extended the TCC method to the case of the graded response model. The present paper extends Raju's

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 17, No. 4, December 1993, pp. 335-350

0146-6216/93/040335-16$2.05

(1990) and Lord's (1980) DIF detection approaches to Samejima's (1969) graded response model.

The Graded Response Model

Samejima (1969) described a graded response IRT model in which an item has m_j ordered response categories. The examinee is permitted to select only one of the categories. For dichotomous response models, there are two IRFs. The IRF for the correct response, $P_j(\theta)$, under the two-parameter logistic model is defined as

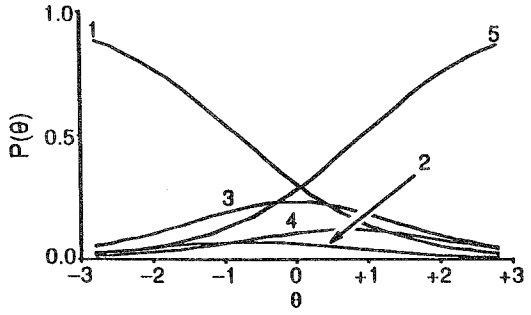
$$P_j(\theta) = \{1 + \exp[-a_j(\theta - b_j)]\}^{-1} , \quad (1)$$

where
a_j is the discrimination parameter for item j,
b_j is the difficulty parameter for item j, and
θ is the trait level parameter.

The IRF for the incorrect response is $Q_j(\theta) = 1 - P_j(\theta)$. Just as there are two IRFs for a dichotomous item, it is possible to specify m_j category response functions (CRFs) for each graded response item, where m_j is the number of response categories for the item. [Thissen, Steinberg, & Mooney (1989) used the term *trace lines* for these functions.] The CRF describes $P_{jk}(\theta)$, which is the probability of response k to item j as a function of θ.

One problem with CRFs is that the functions for a given item do not have the same form (see Figure 1). The CRF for the lowest category, $k = 1$, decreases as θ increases; the CRF for the highest category, $k = 5$, increases as θ increases; and the CRFs for the three intermediate categories, $k = 2$, 3, and 4, increase and then decrease. Because the CRFs do not have a consistent form over the m_j categories, Samejima (1969) defined the boundary response function (BRF) to represent the cumulative probability $P^*_{jk}(\theta)$ of a response above category k.

Figure 1
CRFs for a Five-Category Item



BRFs have a consistent form for a given item and may be characterized by a discrimination parameter a_j, and $m_j - 1$ location parameters b_{jk}. The b_{jk} are ordered, for example, from low ($k = 1$) to high ($k = m_j - 1$). $P_{jk}(\theta)$ is defined in terms of $P^*_{jk}(\theta)$, where $P^*_{jk}(\theta)$ represents the cumulative probability of a response above category k, as follows:
when $k = 1$

$$P_{j1}(\theta) = 1 - P^*_{j1}(\theta) ; \quad (2)$$

when $1 < k < m_j$

$$P_{jk}(\theta) = P^*_{j(k-1)}(\theta) - P^*_{jk}(\theta) \; ; \tag{3}$$

and when $k = m_j$

$$P_{jm_j}(\theta) = P^*_{j(m_j-1)}(\theta) \; . \tag{4}$$

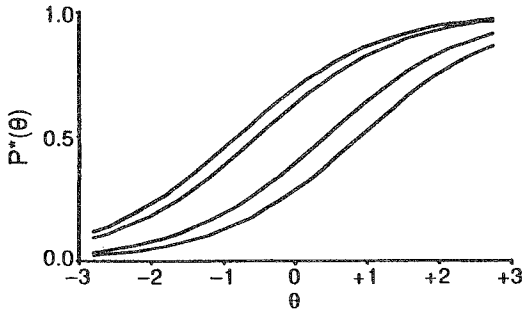
When an item is scored dichotomously and the item score is either 0 (incorrect) or 1 (correct), then $m_j = 2$ and the CRF of the second category, $P_{j2}(\theta)$, is the same as the IRF in Equation 1. Note, further, that $P_{j1}(\theta)$ for this case is expressed in the dichotomous response model as $Q_j(\theta)$.

The logistic form of the BRF is given by

$$P^*_{jk}(\theta) = \{1 + \exp[-a_j(\theta - b_{jk})]\}^{-1} \; . \tag{5}$$

An important assumption of this form of the graded response model is that the reasoning process is homogeneous throughout the set of response categories for an item. Samejima (1969) interpreted this to mean that a_j is constant for all categories in Equation 5. Note that this assumption is specific to the item; it does not mean that the a_j are constant across all items. An example of the four BRFs for a five-category item is given in Figure 2.

Figure 2
BRFs for a Five-Category Item
$a = 1.01$, $b_1 = -.80$, $b_2 = -.52$, $b_3 = .44$, $b_4 = .92$



Item True Score Functions

Baker (1992) defined the true score function for the graded response model as

$$T(\theta) = \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} P_{jk}(\theta) \; , \tag{6}$$

where n is the number of items in the test and u_{jk} is the weight for response category k of item j. (Weights are typically, but not necessarily, taken to be the same as the category values; for example, the weight for category 1 would be 1 and for category 5 it would be 5.)

The true score function in Equation 6 also can be computed for a single item j:

$$T_j(\theta) = \sum_{k=1}^{m_j} u_{jk} P_{jk}(\theta) \; , \tag{7}$$

where $P_{jk}(\theta)$ is defined as in Equations 2–4. This is called the item true score function (ITSF). Note that for the dichotomous case using weights of 0 and 1 for the first and second categories, respectively, the true score function for item j from Equation 7 is the same as the IRF given in Equation 1; that is, $P_j(\theta) = T_j(\theta)$.

DIF in the Graded Response Model

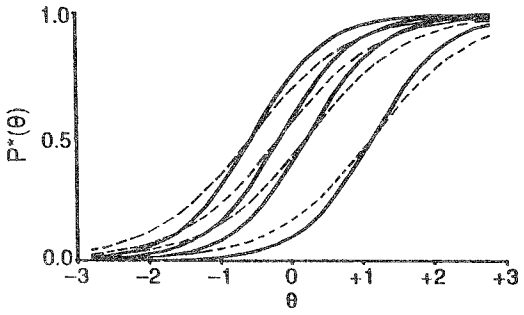
Definition of DIF for ITSFs

An item is considered to be functioning differentially when $T_{jR}(\theta) \neq T_{jF}(\theta)$; that is, when the ITSFs in the reference and focal groups are not equal. Further, the ITSFs from the reference and focal groups are identical when the BRFs for the reference and focal groups are equal, or when the item parameters from the reference and focal groups are equal. These two conditions are essentially equivalent.

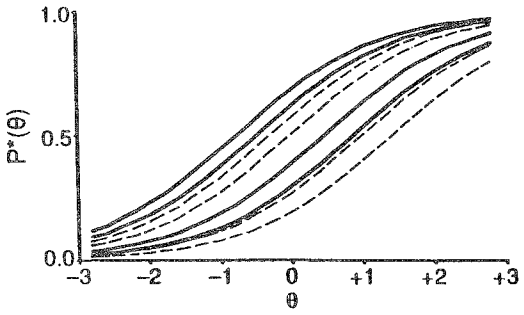
ITSFs may differ in the two-parameter dichotomous response model given in Equation 1 when one of the following occurs: (1) $a_{jR} \neq a_{jF}$ and $b_{jR} = b_{jF}$, (2) $a_{jR} = a_{jF}$ and $b_{jR} \neq b_{jF}$, and (3) $a_{jR} \neq a_{jF}$ and $b_{jR} \neq b_{jF}$. Similar cases may be described for the graded response model. In the graded response model, there is a single a_j and a set of b_{jk}, $\mathbf{b}_j$, for item j. True score functions for item j will differ for this model when one of the following occurs: (1) $a_{jR} \neq a_{jF}$ and $\mathbf{b}_{jR} = \mathbf{b}_{jF}$ (see Figure 3a); (2) $a_{jR} = a_{jF}$ and $\mathbf{b}_{jR} \neq \mathbf{b}_{jF}$ (see Figure 3b); and (3) $a_{jR} \neq a_{jF}$ and $\mathbf{b}_{jR} \neq \mathbf{b}_{jF}$ (see Figure 3c). In Figure 3, solid lines indicate reference group BRFs, and dashed lines indicate focal group BRFs.

Figure 3
DIF for a Five-Category Item

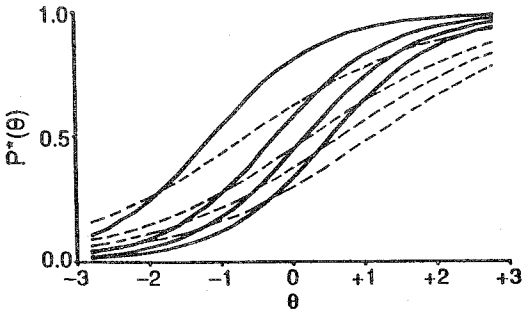
a. Case 1: $a_{jR} \neq a_{jF}$ and $\mathbf{b}_{jR} = \mathbf{b}_{jF}$ ($a_R = 1.94$, $b_{1R} = -.57$, $b_{2R} = -.16$, $b_{3R} = .22$, $b_{4R} = 1.14$; $a_F = 1.44$, $b_{1F} = -.57$, $b_{2F} = -.16$, $b_{3F} = .22$, $b_{4F} = 1.14$)



b. Case 2: $a_{jR} = a_{jF}$ and $\mathbf{b}_{jR} \neq \mathbf{b}_{jF}$ ($a_R = 1.01$, $b_{1R} = -.80$, $b_{2R} = -.52$, $b_{3R} = .44$, $b_{4R} = .92$; $a_F = 1.01$, $b_{1F} = -.30$, $b_{2F} = -.02$, $b_{3F} = .94$, $b_{4F} = 1.42$)



c. Case 3: $a_{jR} \neq a_{jF}$ and $\mathbf{b}_{jR} \neq \mathbf{b}_{jF}$ ($a_R = 1.27$, $b_{1R} = -1.12$, $b_{2R} = -.23$, $b_{3R} = .19$, $b_{4R} = .54$; $a_F = .77$, $b_{1F} = -.62$, $b_{2F} = .27$, $b_{3F} = .69$, $b_{4F} = 1.14$)



DIF Detection Measures

Three DIF detection measures are presented that are based on the definition of DIF given above: two based on areas between ITSFs [the signed area (SA) and unsigned area (UA)] and one based on a comparison of item parameters.

SA between ITSFs. To obtain the SA between ITSFs let

$$\hat{T}_{j\mathrm{R}}(\theta) = \sum_{k=1}^{m_j} u_{jk}\hat{P}_{jk\mathrm{R}}(\theta) \tag{8}$$

be the estimate of the ITSF for item j in the reference group, and let

$$\hat{T}_{j\mathrm{F}}(\theta) = \sum_{k=1}^{m_j} u_{jk}\hat{P}_{jk\mathrm{F}}(\theta) \tag{9}$$

be the estimate of the ITSF in the focal group, where $\hat{P}_{jk\mathrm{R}}(\theta)$ and $\hat{P}_{jk\mathrm{F}}(\theta)$ are the estimates of the CRFs for the reference and focal groups, respectively.

The SA is obtained as

$$\mathrm{SA}_j = \int_{-\infty}^{\infty}[\hat{T}_{j\mathrm{R}}(\theta) - \hat{T}_{j\mathrm{F}}(\theta)]d\theta \ . \tag{10}$$

Substituting Equations 8 and 9 into Equation 10,

$$\mathrm{SA}_j = \int_{-\infty}^{\infty}\left[\sum_{k=1}^{m_j} u_{jk}\hat{P}_{jk\mathrm{R}}(\theta) - \sum_{k=1}^{m_j} u_{jk}\hat{P}_{jk\mathrm{F}}(\theta)\right]d\theta = \int_{-\infty}^{\infty}\sum_{k=1}^{m_j} u_{jk}\left[\hat{P}_{jk\mathrm{R}}(\theta) - \hat{P}_{jk\mathrm{F}}(\theta)\right]d\theta \ . \tag{11}$$

Substituting Equations 2–4 into Equation 11 and expanding individual terms gives

$$\mathrm{SA}_j = \sum_{k=1}^{m_j-1}[u_{j(k+1)} - u_{jk}]\int_{-\infty}^{\infty}[\hat{P}^*_{jk\mathrm{R}}(\theta) - \hat{P}^*_{jk\mathrm{F}}(\theta)]d\theta \ , \tag{12}$$

where $\hat{P}^*_{jk\mathrm{R}}(\theta)$ and $\hat{P}^*_{jk\mathrm{F}}(\theta)$ are the estimates of the BRFs for the reference and focal groups, respectively.

Raju (1988) showed that for the two-parameter model (2PM),

$$\int_{-\infty}^{\infty}[\hat{P}^*_{jk\mathrm{R}}(\theta) - \hat{P}^*_{jk\mathrm{F}}(\theta)]d\theta = \hat{b}_{jk\mathrm{F}} - \hat{b}_{jk\mathrm{R}} \ . \tag{13}$$

Substituting this into Equation 12 gives

$$\mathrm{SA}_j = \sum_{k=1}^{m_j-1}[u_{j(k+1)} - u_{jk}](\hat{b}_{jk\mathrm{F}} - \hat{b}_{jk\mathrm{R}}) \ . \tag{14}$$

Assume further that the asymptotic mean of SA_j is 0 under the null hypothesis that $T_{j\mathrm{R}}(\theta) = T_{j\mathrm{F}}(\theta)$.

The estimated variance of SA_j is defined as

$$\begin{aligned}\mathrm{Var}(\mathrm{SA}_j) = &\sum_{k=1}^{m_j-1}[u_{j(k+1)} - u_{jk}]^2\mathrm{Var}(\hat{b}_{jk\mathrm{F}}) + \sum_{k=1}^{m_j-1}\sum_{l=1}^{m_j-1}[u_{j(k+1)} - u_{jk}][u_{j(l+1)} - u_{jl}]\mathrm{Cov}(\hat{b}_{jk\mathrm{F}}, \hat{b}_{jl\mathrm{F}}) \\ &+ \sum_{k=1}^{m_j-1}[u_{j(k+1)} - u_{jk}]^2\mathrm{Var}(\hat{b}_{jk\mathrm{R}}) + \sum_{k=1}^{m_j-1}\sum_{l=1}^{m_j-1}[u_{j(k+1)} - u_{jk}][u_{j(l+1)} - u_{jl}]\mathrm{Cov}(\hat{b}_{jk\mathrm{R}}, \hat{b}_{jl\mathrm{R}})\end{aligned} \tag{15}$$

for $k \neq l$. The test statistic $Z(\mathrm{SA}_j)$ can be written as

$$Z(\mathrm{SA}_j) = \frac{\mathrm{SA}_j}{[\mathrm{Var}(\mathrm{SA}_j)]^{1/2}} \tag{16}$$

and is based on the assumption that the observed SAs are normally distributed with mean 0 and variance given in Equation 15.

SA for individual BRFs. Recall that the proposed definition of DIF refers to differences in the ITSFs between the reference and focal groups. However, it further states that, if the BRFs for the reference and focal groups are not equal, then the ITSFs cannot be equal. To test the equality of the BRFs, specify the SA between the kth BRFs from the reference and focal groups as

$$\mathrm{SA}_{jk} = [u_{j(k+1)} - u_{jk}](\hat{b}_{jk\mathrm{R}} - \hat{b}_{jk\mathrm{F}}) , \tag{17}$$

where $k = 1, \ldots, m_j - 1$. Under the null hypothesis, this area is assumed to be 0; the variance is given by

$$\mathrm{Var}(\mathrm{SA}_{jk}) = [u_{j(k+1)} - u_{jk}]^2[\mathrm{Var}(\hat{b}_{jk\mathrm{R}}) + \mathrm{Var}(\hat{b}_{jk\mathrm{F}})] . \tag{18}$$

The test statistic for the kth BRF is then

$$Z(\mathrm{SA}_{jk}) = \frac{\hat{b}_{jk\mathrm{F}} - \hat{b}_{jk\mathrm{R}}}{[\mathrm{Var}(\hat{b}_{jk\mathrm{R}}) + \mathrm{Var}(\hat{b}_{jk\mathrm{F}})]^{1/2}} . \tag{19}$$

UA for ITSFs. A second measure of the difference between ITSFs is the UA. The UA can be defined as

$$\mathrm{UA}_j = \int_{-\infty}^{\infty} |\hat{T}_{j\mathrm{R}}(\theta) - \hat{T}_{j\mathrm{F}}(\theta)| \, d\theta = \int_{-\infty}^{\infty} \left| \sum_{k=1}^{m_j} u_{jk}\hat{P}_{jk\mathrm{R}}(\theta) - \sum_{k=1}^{m_j} u_{jk}\hat{P}_{jk\mathrm{F}}(\theta) \right| d\theta . \tag{20}$$

Expressing UA_j in terms of BRFs gives

$$\mathrm{UA}_j = \int_{-\infty}^{\infty} \left| \sum_{k=1}^{m_j - 1} [u_{j(k+1)} - u_{jk}][\hat{P}^*_{jk\mathrm{R}}(\theta) - \hat{P}^*_{jk\mathrm{F}}(\theta)] \right| d\theta . \tag{21}$$

If either $\hat{T}_{j\mathrm{R}}(\theta) \geq \hat{T}_{j\mathrm{F}}(\theta)$ or $\hat{T}_{j\mathrm{R}}(\theta) \leq \hat{T}_{j\mathrm{F}}(\theta)$ for all θ, then

$$\mathrm{UA}_j = |\mathrm{SA}_j| . \tag{22}$$

Assuming that SA_js are normally distributed with mean 0 and variance as given in Equation 15, the expected value of UA_j is

$$\mathrm{E}(\mathrm{UA}_j) = \left[\frac{2}{\pi}\mathrm{Var}(\mathrm{SA}_j)\right]^{1/2} , \tag{23}$$

and the variance of UA_j is

$$\mathrm{Var}(\mathrm{UA}_j) = \mathrm{Var}(\mathrm{SA}_j)\left[1 - \frac{2}{\pi}\right] \tag{24}$$

(Hogg & Craig, 1978; Johnson & Kotz, 1970; see the Appendix for a discussion of the expectation and variance of UA_j.) Note that the assumption of normality for UA_j may not be justified (Raju, 1990). If either $\hat{T}_{j\mathrm{R}}(\theta) \geq \hat{T}_{j\mathrm{F}}(\theta)$ or $\hat{T}_{j\mathrm{R}}(\theta) \leq \hat{T}_{j\mathrm{F}}(\theta)$ for all θ, Equation 16 provides a test of the null hypothesis for the UA.

However, if either condition [$\hat{T}_{j\mathrm{R}}(\theta) \geq \hat{T}_{j\mathrm{F}}(\theta)$ or $\hat{T}_{j\mathrm{R}}(\theta) \leq \hat{T}_{j\mathrm{F}}(\theta)$] does not hold for all θ, UA_j may not have a closed form. In this case, no statistical test is available for the null hypothesis. Even so, it may be of interest to examine the size of UA_j. Therefore, the following approximation may be used

for UA_j: Select two points—θ_L and θ_U, such that $\theta_L < \theta_U$—and divide the range into N intervals. The area UA_j then is approximated as follows by the bounded UA using the trapezoidal approximation (Burden & Faires, 1985):

$$UA_j \approx \sum_{i=1}^{N} \left| \sum_{k=1}^{m_j-1} [u_{j(k+1)} - u_{jk}][\hat{P}^*_{jkR}(\theta_i) - \hat{P}^*_{jkF}(\theta_i)] \right| \Delta\theta + \frac{1}{2} \left| \sum_{k=1}^{m_j-1} [u_{j(k+1)} - u_{jk}][\hat{P}^*_{jkR}(\theta_L) - \hat{P}^*_{jkF}(\theta_L)] \right| \Delta\theta$$

$$- \frac{1}{2} \left| \sum_{k=1}^{m_j-1} [u_{j(k+1)} - u_{jk}][\hat{P}^*_{jkR}(\theta_U) - \hat{P}^*_{jkF}(\theta_U)] \right| \Delta\theta , \tag{25}$$

where $\Delta\theta = (\theta_U - \theta_L)/N$.

UA for individual BRFs. Using an approach similar to that described above for the SA for individual BRFs, the UA between BRFs also can be obtained. The UA between the kth BRFs from the reference and focal groups is

$$UA_{jk} = \int_{-\infty}^{\infty} \left| [u_{j(k+1)} - u_{jk}][\hat{P}^*_{jkR}(\theta) - \hat{P}^*_{jkF}(\theta)] \right| d\theta , \tag{26}$$

where $k = 1, \ldots, m_j - 1$. If it is assumed that the category weights are ordered [e.g., $u_{j(k+1)} > u_{jk}$], then the UA is

$$UA_{jk} = [u_{j(k+1)} - u_{jk}] \int_{-\infty}^{\infty} |\hat{P}^*_{jkR}(\theta) - \hat{P}^*_{jkF}(\theta)| d\theta . \tag{27}$$

According to Raju (1988),

$$\int_{-\infty}^{\infty} |\hat{P}^*_{jkR}(\theta) - \hat{P}^*_{jkF}(\theta)| d\theta = \begin{cases} |\hat{b}_{jkF} - \hat{b}_{jkR}| & \text{if } \hat{a}_{jF} = \hat{a}_{jR} \\ \left| \dfrac{2(\hat{a}_{jF} - \hat{a}_{jR})}{\hat{a}_{jF}\hat{a}_{jR}} \ln [1 + \exp(Y_{jk})] - (\hat{b}_{jkF} - \hat{b}_{jkR}) \right| & \text{otherwise} , \end{cases} \tag{28}$$

where

$$Y_{jk} = \frac{\hat{a}_{jF}\hat{a}_{jR}(\hat{b}_{jkF} - \hat{b}_{jkR})}{\hat{a}_{jF} - \hat{a}_{jR}} . \tag{29}$$

The UA between the kth BRFs from the reference and focal groups then can be expressed as

$$UA_{jk} = \begin{cases} [u_{j(k+1)} - u_{jk}] \; |\hat{b}_{jkF} - \hat{b}_{jkR}| & \text{if } \hat{a}_{jF} = \hat{a}_{jR} \\ [u_{j(k+1)} - u_{jk}] \left| \dfrac{2(\hat{a}_{jF} - \hat{a}_{jR})}{\hat{a}_{jF}\hat{a}_{jR}} \ln [1 + \exp(Y_{jk})] - (\hat{b}_{jkF} - \hat{b}_{jkR}) \right| & \text{otherwise} . \end{cases} \tag{30}$$

When $\hat{a}_{jR} = \hat{a}_{jF}$, the asymptotic mean and variance of the UA of the kth BRFs, under the null hypothesis, are

$$E(UA_{jk}) = [u_{j(k+1)} - u_{jk}] \left\{ \frac{2}{\pi} [\mathrm{Var}(\hat{b}_{jkF}) + \mathrm{Var}(\hat{b}_{jkR})] \right\}^{1/2} \tag{31}$$

and

$$\mathrm{Var}(UA_{jk}) = [u_{j(k+1)} - u_{jk}]^2 \left\{ [\mathrm{Var}(\hat{b}_{jkF}) + \mathrm{Var}(\hat{b}_{jkR})] \left(1 - \frac{2}{\pi}\right) \right\} . \tag{32}$$

Raju (1990) noted that the assumption of normality would not be valid for the UA in this case and suggested instead that the test statistic for $\hat{b}_{jkF} - \hat{b}_{jkR}$ be used. When $\hat{a}_{jF} = \hat{a}_{jR}$, the test statistic is the same as that given in Equation 19.

When $\hat{a}_{jF} \neq \hat{a}_{jR}$, Raju (1990) defined the test statistic as

$$Z(H_{jk}) = \frac{H_{jk}}{[\mathrm{Var}(H_{jk})]^{1/2}}, \tag{33}$$

where

$$H_{jk} = \frac{2(\hat{a}_{jF} - \hat{a}_{jR})}{\hat{a}_{jF}\hat{a}_{jR}} \ln[1 + \exp(Y_{jk})] - (\hat{b}_{jkF} - \hat{b}_{jkR}) \tag{34}$$

and

$$\mathrm{Var}(H_{jk}) = A_{jk}^2\mathrm{Var}(\hat{b}_{jkR}) + B_{jk}^2\mathrm{Var}(\hat{b}_{jkF}) + C_{jk}^2\mathrm{Var}(\hat{a}_{jR}) + D_{jk}^2\mathrm{Var}(\hat{a}_{jF}) + 2A_{jk}C_{jk}\mathrm{Cov}(\hat{a}_{jR}, \hat{b}_{jkR}) + 2B_{jk}D_{jk}\mathrm{Cov}(\hat{a}_{jF}, \hat{b}_{jkF}), \tag{35}$$

where

$$A_{jk} = 1 - \frac{2\exp(Y_{jk})}{1 + \exp(Y_{jk})}, \tag{36}$$

$$B_{jk} = -A_{jk}, \tag{37}$$

$$C_{jk} = \frac{2}{\hat{a}_{jR}^2}\left\{\frac{Y_{jk}\exp(Y_{jk})}{1 + \exp(Y_{jk})} - \ln[1 + \exp(Y_{jk})]\right\}, \tag{38}$$

and

$$D_{jk} = -\frac{\hat{a}_{jR}^2}{\hat{a}_{jF}^2} C_{jk}. \tag{39}$$

Detection of DIF using Lord's χ^2. Lord's χ^2 (1980) has been shown to be an effective means of detection of DIF in the dichotomous model (Candell & Hulin, 1987; McCauley & Mendoza, 1985). This same statistic also can be used to test the hypothesis that the parameters estimated for a graded response item are the same between reference and focal groups. For example, the vector of item parameter estimates for the reference group for the two-parameter IRF is

$$\hat{\xi}_{jR} = [\hat{a}_{jR}, \hat{b}_{jR}]', \tag{40}$$

and for the focal group it is

$$\hat{\xi}_{jF} = [\hat{a}_{jF}, \hat{b}_{jF}]'. \tag{41}$$

The χ^2 statistic is computed as

$$\chi_j^2 = \hat{\xi}_j'\hat{\Sigma}_j^{-1}\hat{\xi}_j, \tag{42}$$

where $\hat{\xi}_j$ is the vector of differences between parameter estimates $[\hat{a}_{jF} - \hat{a}_{jR}, \hat{b}_{jF} - \hat{b}_{jR}]'$, (i.e.,

$\hat{\xi}_j = \hat{\xi}_{jF} - \hat{\xi}_{jR}$), and $\hat{\Sigma}_j^{-1}$ is the inverse of the variance-covariance matrix for $(\hat{a}_{jF} - \hat{a}_{jR})$ and $(\hat{b}_{jF} - \hat{b}_{jR})$.

Because the parameter estimates from the reference group are independent of those from the focal group,

$$\hat{\Sigma}_j = \hat{\Sigma}_{jR} + \hat{\Sigma}_{jF} , \tag{43}$$

where $\hat{\Sigma}_{jR}$ is the matrix of variance-covariance estimates from the reference group, and $\hat{\Sigma}_{jF}$ is the matrix of variance-covariance estimates from the focal group. Under the null hypothesis, Lord's χ^2 for the 2PM has two degrees of freedom.

For the graded response model with m_j categories, Lord's χ^2 can be extended to include the vectors of item parameter estimates and their respective estimated variance-covariance matrices. For example, the vector of item parameter estimates for the reference group would be

$$\hat{\xi}_{jR} = [\hat{a}_{jR}, \hat{b}_{j1R}, \ldots, \hat{b}_{j(m_j-1)R}]' , \tag{44}$$

and the variance-covariance matrix would be

$$\hat{\Sigma}_{jR} = \begin{bmatrix} \mathrm{Var}(\hat{a}_{jR}) & \mathrm{Cov}(\hat{a}_{jR}, \hat{b}_{j1R}) & \ldots & \mathrm{Cov}(\hat{a}_{jR}, \hat{b}_{j(m_j-1)R}) \\ & \mathrm{Var}(\hat{b}_{j1R}) & \ldots & \mathrm{Cov}(\hat{b}_{j1R}, \hat{b}_{j(m_j-1)R}) \\ & & \ddots & \vdots \\ & & & \mathrm{Var}(\hat{b}_{j(m_j-1)R}) \end{bmatrix} . \tag{45}$$

The vector of item parameter estimates and the estimated variance-covariance matrix for the focal group can be defined similarly. There are m_j degrees of freedom for this extension of Lord's χ^2 for the graded response model with m_j categories.

Example: Simulated DIF

Method

Data simulation. The DIF detection procedures described above were investigated using a simulated DIF procedure. Data were generated for a reference group and a focal group for the graded response model using the computer program GENIRV (Baker, 1986). A 40-item test with five categories per item was simulated with 1,000 examinees. Generating parameters for the reference and focal groups for θ were normal (0,1). Generating parameters for the reference group for each b_{jk} were normal (0,1), and for the a_j were uniform over the interval [1.0, 2.0]. For each item, four location parameters were selected randomly from the standard normal distribution, arranged in increasing order, and then paired with a randomly selected a_j. The generating parameters for the reference group and the focal group are given in Table 1.

DIF was simulated in the focal group by modifying either the a_j or $\mathbf{b}_j$ in the first six items of the test generated for the reference group (see Table 1). For Items 1 and 2, only the location parameters ($\mathbf{b}_j$) were modified. For Items 3 and 4, only the discrimination parameters (a_j) were modified. Finally, for Items 5 and 6, both $\mathbf{b}_j$ and a_j parameters were modified. Item parameter estimates were obtained using the computer program MULTILOG (Thissen, 1991). Program default settings were used.

In the calculation of the DIF statistics, the off-diagonal terms in the variance-covariance matrix were not provided by the version of MULTILOG used; however, item parameter estimates are known to be correlated even though item discrimination and item location parameters are assumed to be independent (Baker, 1985). Although the estimated covariance terms are typically much smaller than the variance terms, to the extent that this is not the case, ignoring the covariance terms (as was

Table 1
Item Parameters Used to Generate the Datasets

	Reference Group					Focal Group				
Item	a_{jR}	b_{j1R}	b_{j2R}	b_{j3R}	b_{j4R}	a_{jF}	b_{j1F}	b_{j2F}	b_{j3F}	b_{j4F}
1	1.01	−.80	−.52	.44	.92	1.01	−.30	−.02	.94	1.42
2	1.22	−.33	.51	.89	1.34	1.22	.67	1.51	1.89	2.34
3	1.94	−.57	−.16	.22	1.14	1.44	−.57	−.16	.22	1.14
4	1.53	−1.34	−.33	.24	1.47	.53	−1.34	−.33	.24	1.47
5	1.27	−1.12	−.23	.19	.54	.77	−.62	.27	.69	1.04
6	1.32	−1.40	−.59	.41	.76	.32	−.40	.43	1.41	1.76
7	1.75	−.91	−.44	.05	.44	1.75	−.91	−.44	.05	.44
8	1.49	−1.44	.18	.44	.89	1.49	−1.44	.18	.44	.89
9	1.81	−1.10	−.43	.06	.55	1.81	−1.10	−.43	.06	.55
10	1.25	−1.97	−1.33	−.57	−.02	1.25	−1.97	−1.33	−.57	−.02
11	1.34	−1.54	−.83	.18	.88	1.34	−1.54	−.83	.18	.88
12	1.00	−1.10	−.82	−.42	2.06	1.00	−1.10	−.82	−.42	2.06
13	1.80	−1.48	−.96	.13	.65	1.80	−1.48	−.96	.13	.65
14	1.62	−.62	−.38	−.06	1.89	1.62	−.62	−.38	−.06	1.89
15	1.80	−.83	−.33	.22	1.24	1.80	−.83	−.33	.22	1.24
16	1.36	−1.69	−.50	.08	.63	1.36	−1.69	−.50	.08	.63
17	1.04	−1.61	−.90	.34	1.96	1.04	−1.61	−.90	.34	1.96
18	1.30	−.56	−.07	.41	.83	1.30	−.56	−.07	.41	.83
19	1.87	−.31	.19	.45	1.79	1.87	−.31	.19	.45	1.79
20	1.37	−.93	.04	.99	1.27	1.37	−.93	.04	.99	1.27
21	1.88	.26	.46	.98	1.39	1.88	.26	.46	.98	1.39
22	1.23	−1.36	−1.05	−.86	−.01	1.23	−1.36	−1.05	−.86	−.01
23	1.11	−1.90	−.25	.24	.89	1.11	−1.90	−.25	.24	.89
24	1.14	−.74	−.24	.19	1.65	1.14	−.74	−.24	.19	1.65
25	1.99	−.21	.29	.70	.86	1.99	−.21	.29	.70	.86
26	1.88	−.47	.42	.93	1.81	1.88	−.47	.42	.93	1.81
27	1.46	−.50	−.18	.52	.81	1.46	−.50	−.18	.52	.81
28	1.47	−1.25	.25	.77	1.02	1.47	−1.25	.25	.77	1.02
29	1.15	−.62	−.15	.47	1.03	1.15	−.62	−.15	.47	1.03
30	1.24	−.34	.44	.89	1.45	1.24	−.34	.44	.89	1.45
31	1.98	−.30	−.10	.15	.58	1.98	−.30	−.10	.15	.58
32	1.82	−1.08	−.64	−.08	.90	1.82	−1.08	−.64	−.08	.90
33	1.81	−.22	.31	.54	.86	1.81	−.22	.31	.54	.86
34	1.37	−1.76	−1.03	−.54	.12	1.37	−1.76	−1.03	−.54	.12
35	1.19	−.32	.10	.41	1.98	1.19	−.32	.10	.41	1.98
36	1.27	−1.64	−1.14	1.22	1.69	1.27	−1.64	−1.14	1.22	1.69
37	1.58	−1.72	−.63	.30	.66	1.58	−1.72	−.63	.30	.66
38	1.00	−1.77	−1.27	−.72	−.27	1.00	−1.77	−1.27	−.72	−.27
39	1.07	−2.31	−.69	−.36	1.25	1.07	−2.31	−.69	−.36	1.25
40	1.39	−1.35	−.17	.36	.71	1.39	−1.35	−.17	.36	.71

necessary here), may possibly degrade the utility of the measures presented in this paper. With respect to the present example, GENIRV, which was used to generate the simulated test data, provided independent parameters for each item. Consequently, the absence of the off-diagonal terms from the computations should have had little effect on the present results. Availability of these off-diagonal terms in future IRT computer programs, however, clearly is important, if only to be able to determine their impact on DIF detection.

Parameter recovery. Item parameter estimates for both the reference and focal groups were equated

to the underlying metric using the TCC method of the computer program EQUATE 2.0 (Baker, 1993). Linear equating coefficients obtained for the reference group were $A = .97788$ and $K = .0078$, and for the focal group were $A = .98615$ and $K = -.00858$. Root mean square differences (RMSDs) and correlations then were calculated between the generating parameters and the parameter estimates. Table 2 shows that the RMSDs between the generating parameters and item parameter estimates were small for both the reference and focal groups, and that the correlations were high. These results indicate that recovery was very good in both the reference group and focal group datasets.

Table 2
RMSDs and Correlations (*r*s) Between Estimates and True Values

	a		b_1		b_2		b_3		b_4	
Group	RMSD	r	RMSD	r	RMSD	r	RMSD	r	RMSD	r
Reference	.110	.950	.071	.993	.063	.993	.055	.995	.089	.989
Focal	.071	.982	.077	.993	.045	.996	.063	.994	.077	.992

Comparison of item parameters. In order to compare either ITSFs or item parameter estimates, it is first necessary to place the estimates onto the same metric. The following procedure was used for placing focal group item parameters onto the reference group metric and then calculating each DIF statistic:

1. Calibrate the items in the focal and reference groups separately.
2. Using the TCC method (Baker, 1992), calculate the linear equating coefficients to place the focal group item parameter estimates onto the scale of the reference group using the anchor items in both the reference and focal groups. [The anchor items were the 34 items (Items 7–40, see Table 1) common to both groups in which DIF was not simulated.] The coefficients were $A = 1.00235$ and $K = .01883$.
3. Transform all focal group item parameter estimates and their standard errors onto the reference group metric. (See Baker, 1992 for details on equating with the graded response model.)
4. Calculate DIF measures.

Results

Detection of DIF. DIF statistics for the extension of Lord's χ^2 and the areas between ITSFs are given in Table 3. The extension of Lord's χ^2 detected DIF in five of the six items in which DIF was simulated at $\alpha = .01$, and in all six items at the .05 level of significance. The DIF simulated in Item 3, which was missed at $\alpha = .01$, had been modified by shifting the discrimination parameter. This type of DIF may be difficult to detect in graded response items. In addition, DIF was detected in Item 18 at $\alpha = .05$, but it had not originally been simulated as a DIF item.

Z(SA) detected four of the six simulated DIF items at both $\alpha = .01$ and $\alpha = .05$. Further, three items not originally simulated as DIF items (Items 15, 18, and 33) also were identified as DIF items at $\alpha = .05$. Items 15 and 18 also were identified at $\alpha = .01$. Z(SA) failed to detect Items 3 and 4 as DIF items; DIF in these two items was simulated by modifying only the discrimination parameters (see Table 1). Although no significance test is available for the UA_j between ITSFs, note that five of the six UAs for the simulated DIF items were all larger than the values of the UAs for the 34 non-DIF items.

Area differences between individual BRFs. The SAs and UAs between ITSFs were described as composites of the SAs and UAs between individual BRFs. For this reason, the two-tailed tests of areas between individual BRFs were conducted at the reduced $\alpha/(m_j - 1)$ level of significance. For this example, the individual BRFs were tested at $\alpha = .0125$ and $\alpha = .0025$ for the .05 and .01 overall α levels,

Table 3
Lord's χ^2 and Area DIF Measures

Item	χ^2	SA_j	$Var(SA_j)$	$Z(SA_j)$	UA_j
1	46.08**	2.10	.10	6.70**	2.10
2	129.96**	3.50	.12	9.93**	3.50
3	11.29*	−.03	.04	−.15	.82
4	54.36**	−.60	.19	−1.39	4.81
5	50.03**	1.84	.09	6.12**	2.65
6	84.91**	4.14	.78	4.70**	12.25
7	8.72	−.32	.03	−1.75	.34
8	.47	.13	.05	.56	.13
9	1.16	.01	.03	.05	.16
10	.38	.04	.12	.11	.32
11	3.57	−.16	.09	−.54	.59
12	3.84	−.62	.17	−1.49	.99
13	3.95	.27	.04	1.40	.27
14	8.91	.27	.05	1.24	.44
15	10.80	−.63	.05	−2.85**	.71
16	.37	−.04	.06	−.17	.05
17	.65	.09	.16	.22	.39
18	12.94*	−.69	.06	−2.81**	.96
19	6.32	−.23	.05	−1.05	.69
20	2.32	−.15	.07	−.56	.31
21	1.89	−.16	.04	−.84	.16
22	.39	.01	.07	.04	.17
23	3.29	.53	.12	1.56	.53
24	3.63	.11	.09	.36	.78
25	1.69	−.00	.03	−.02	.26
26	2.85	.18	.05	.76	.42
27	2.14	.15	.06	.61	.15
28	.86	−.20	.05	−.89	.21
29	.61	−.13	.07	−.48	.15
30	2.16	−.21	.08	−.74	.21
31	.85	.05	.03	.31	.16
32	1.42	.14	.04	.72	.17
33	6.64	−.44	.03	−2.49*	.44
34	5.35	.33	.09	1.10	.70
35	2.65	−.28	.11	−.85	.83
36	.96	−.01	.11	−.03	.19
37	2.79	−.25	.05	−1.08	.29
38	1.10	.17	.17	.41	.78
39	1.09	.16	.17	.39	.17
40	3.82	−.35	.07	−1.36	.42

*$p < .05$; **$p < .01$.

respectively. Results for the significant SAs and UAs between individual BRFs are given in Tables 4 and 5.

For $Z(SA)$ (Table 4), differences between BRFs were detected for four of the six simulated DIF items at $\alpha = .0125$; at $\alpha = .0025$, four simulated DIF items (Items 1, 2, 5, and 6) had one or more significant differences between individual BRFs. None of the individual BRFs were significant for Items 3 or 4 at either $\alpha = .0125$ or $\alpha = .0025$. Items 15, 18, and 33 were identified by SA_j as DIF items; however, only Item 15 had a significant SA_{jk}, SA_{jl}, at $\alpha = .0125$.

Table 4
SA Measures for Individual BRFs, for Items with Significant SAs

	1st BRF			2nd BRF			3rd BRF			4th BRF		
Item	SA_{j1}	Var	Z	SA_{j2}	Var	Z	SA_{j3}	Var	Z	SA_{j4}	Var	Z
1	.42	.02	2.82*	.47	.02	3.52**	.54	.02	3.63**	.67	.04	3.53**
2	.97	.01	8.18**	.82	.02	5.30**	.83	.03	4.46**	.88	.05	3.87**
3	−.01	.01	−.08	.01	.01	.14	.06	.01	.69	−.10	.02	−.77
4	−.06	.07	−.23	−.14	.03	−.79	−.14	.02	−1.13	−.25	.07	−.97
5	.40	.03	2.49	.47	.02	3.52**	.46	.02	3.39**	.50	.03	3.02**
6	1.04	.11	3.10**	1.04	.10	3.28**	1.09	.25	2.21	.97	.32	1.71
15	−.10	.01	−.90	−.04	.01	−.49	−.11	.01	−1.16	−.38	.02	−2.69*
18	0.00	.01	−.01	−.19	.02	−1.46	−.25	.01	−2.20	−.25	.02	−1.95
33	−.09	.01	−1.03	−.11	.01	−1.29	−.14	.01	−1.70	−.10	.01	−1.01

*$p < .0125$; **$p < .0025$.

For Z(UA) (Table 5), all four of the BRFs for each of the six items with simulated DIF were significant at $\alpha = .0125$. In addition, one BRF in Items 15 and 18 in which DIF was not simulated, were significant at $\alpha = .0125$.

Table 5
UA Measures for Individual BRFs, for Items with Significant UAs

	1st BRF			2nd BRF			3rd BRF			4th BRF		
Item	H_{j1}	Var	Z	H_{j2}	Var	Z	H_{j3}	Var	Z	H_{j4}	Var	Z
1	−.42	.02	−2.82*	−.47	.02	−3.52**	−.54	.02	−3.63**	−.67	.04	−3.53**
2	.97	.01	8.18**	.82	.02	5.30**	.83	.03	4.46**	.88	.05	3.87**
3	−.26	.01	−3.12**	−.26	.01	−3.13**	−.27	.01	−3.20**	−.27	.01	−3.18**
4	−1.43	.11	−4.35**	−1.43	.11	−4.37**	−1.43	.11	−4.38**	−1.44	.11	−4.42**
5	−.66	.03	−3.74**	−.69	.03	−4.18**	−.69	.03	−4.13**	−.71	.03	−4.12**
6	−3.54	1.38	−3.02**	−3.54	1.37	−3.02**	−3.56	1.37	−3.04**	−3.53	1.39	−2.99*
15	.13	.01	1.37	.10	.01	1.19	.13	.01	1.59	.38	.02	2.70*
18	.17	.01	1.46	.24	.01	2.07	.28	.01	2.63*	.28	.01	2.43
33	−.09	.01	−1.03	−.11	.01	−1.29	−.14	.01	−1.70	−.10	.01	−1.01

*$p < .0125$; **$p < .0025$.

Discussion

The definition of DIF based on ITSFs presents a somewhat different view of DIF, but one that is completely consistent with previous definitions (cf. Marascuilo & Slaughter, 1981). Further, the measures of DIF described in this paper for the graded response model appeared to be useful for detection of the kinds of DIF simulated in the example. Although the extension of Lord's χ^2 seemed to be the most effective, differences among the measures were minor. Additional careful study on both simulated as well as real datasets clearly is needed for these statistics.

The SAs and UAs between individual BRFs also were useful in locating DIF among the categories within the item. Note that the translation of a significant SA between BRFs in the reference and focal groups is not directly attributable to a single response category. This is because the BRF represents the cumulative probability of a response above category k. Therefore, when this SA measure is significant, the correct interpretation should be that the DIF is in the two adjacent categories separated by the BRF. If the SAs or UAs for the second BRF are significant, for example, the DIF identified by these measures is located in the second and third categories in the item.

One problem with using the UA between ITSFs, UA_j, is that the distribution of this statistic is not known. Given the type of function involved, monte carlo or resampling studies should be helpful for determining the distributional characteristics of this measure.

Analysis of DIF under IRT requires that all items to be compared are on the same metric. Linking, however, requires that DIF items be removed from calculations of the linking constants. In the example presented here, it was possible to identify the non-DIF items and use only those items for calculation of the linking constants. In real data, however, it is not possible to accurately identify DIF items a priori. In a practical testing situation, therefore, more complicated linking methods such as purification (Lord, 1980) or iterative linking (Candell & Drasgow, 1988) should be considered.

One problem with a measure such as the SA between ITSFs, SA_j, is that the resulting value may be 0 or at least nonsignificant even when the item still is functioning differentially. This was the case in Items 3 and 4 in the example. These two items were generated in part to simulate a type of DIF that is sometimes missed by the SA measure but may be detected by other measures. Therefore, it is important, when using a measure such as the SA, that other measures such as the extension of Lord's χ^2, $Z(UA_{jk})$, or the likelihood ratio test described by Wainer, Sireci, & Thissen (1991) be considered as well. These procedures should be considered complementary and should probably not be relied on as sole indicators of DIF.

The difference between two IRFs (i.e., DIF) may be more relevant in an area of the trait distribution in which there are more examinees. This raises the issue of the impact of DIF on specific populations. Wainer (1993) described several measures of the impact of DIF. This notion of impact can be applied to the graded response model in the context of item true scores. Recall that under IRT, the difference between IRFs is assumed to be independent of the trait distributions of either the focal or reference groups. However, estimates of the impact of DIF require integration of a measure of DIF, such as SA_j, over a trait distribution. In this regard, the choice of a trait distribution is generally that of the focal group (Wainer, 1993), because the focal group is normally the focus of the DIF study. The choice of the trait distribution is important because measures of the impact of DIF will differ depending on the specific group selected as the focal group.

Finally, in the calculation of the DIF statistics in the example presented above, the off-diagonal terms in the variance-covariance matrix were not used. This was because computer programs used for the graded response model, such as MULTILOG or PARSCALE (Muraki & Bock, 1991), do not provide these estimates. Availability of the correct estimates of these terms in future programs is desirable if only to be able to determine their impact on DIF detection.

Appendix: The Expectation and Variance of x

If X is distributed normal (μ, σ^2), Hogg & Craig (1978) showed that

$$E(|X - \mu|) = \left(\frac{2\sigma^2}{\pi}\right)^{1/2} \tag{46}$$

and

$$\begin{aligned} E(|X - \mu|) &= \int_{-\infty}^{\infty} |x - \mu| \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right] dx \\ &= \int_{-\infty}^{\mu} (-x + \mu) \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right] dx \\ &+ \int_{\mu}^{\infty} (x - \mu) \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right] dx . \end{aligned} \tag{47}$$

Let $(x - \mu)/\sigma = y$; then $x = \sigma y + \mu$ and

$$\begin{aligned} E(|X - \mu|) &= \int_{-\infty}^{0} \frac{-y\sigma}{(2\pi)^{1/2}\sigma} \exp\left(-\frac{y^2}{2}\right) |\sigma| dy + \int_{0}^{\infty} \frac{y\sigma}{(2\pi)^{1/2}\sigma} \exp\left(-\frac{y^2}{2}\right) |\sigma| dy \\ &= \frac{\sigma}{(2\pi)^{1/2}} \left[\exp\left(-\frac{y^2}{2}\right) \right]_{-\infty}^{0} + \frac{\sigma}{(2\pi)^{1/2}} \left[-\exp\left(-\frac{y^2}{2}\right) \right]_{0}^{\infty} \\ &= \frac{\sigma}{(2\pi)^{1/2}} + \frac{\sigma}{(2\pi)^{1/2}} = \frac{2\sigma}{(2\pi)^{1/2}} = \left(\frac{2\sigma^2}{\pi}\right)^{1/2}. \end{aligned} \tag{48}$$

When X is distributed normal $(0, \sigma^2)$, then

$$E(|X|) = \left(\frac{2\sigma^2}{\pi}\right)^{1/2} \tag{49}$$

and

$$\mathrm{Var}(|X|) = E(|X|^2) - [E(|X|)]^2 = \sigma^2 - \left(\frac{2\sigma^2}{\pi}\right) = \sigma^2\left(1 - \frac{2}{\pi}\right). \tag{50}$$

References

Baker, F. B. (1985). *The basics of item response theory.* Portsmouth NH: Heinemann.

Baker, F. B. (1986). *GENIRV: A program to generate item response vectors* [Computer program]. Madison: University of Wisconsin, School of Education, Department of Educational Psychology, Laboratory of Experimental Design.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16,* 87–96.

Baker, F. B. (1993). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement, 17,* 20.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28,* 147–162.

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

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12,* 253–260.

Candell, G. L., & Hulin, C. L. (1987). Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. *Journal of Cross-Cultural Psychology, 17,* 417–440.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement, 9,* 413–415.

Hogg, R. V., & Craig, A. T. (1978). *Introduction to mathematical statistics* (4th ed). New York: Macmillan.

Johnson, N. L., & Kotz, S. (1970). *Continuous univariate distributions: 1.* Boston: Houghton Mifflin.

Kim, S.-H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29,* 51–66.

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.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179–193.

Marascuilo, L. A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on χ^2 statistics. *Journal of Educational Measurement, 18,* 229–248.

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.

Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parameter scaling of rating data* [Computer program]. Chicago IL: Scientific Software.

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

Raju, N. S. (1990). Determining the significance of

estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14,* 197–207.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, No. 17.

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

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

Thissen, D. (1991). *MULTILOG user's guide* (Version 6.0) [Computer program]. Chicago IL: Scientific Software.

Thissen, D. (1992). *PLOTLOG for the MacIntosh* [Computer program]. Chapel Hill: University of North Carolina, L. L. Thurstone Psychometric Laboratory.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26,* 247–260.

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

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28,* 197–219.

Acknowledgments

Graphics for the figures presented in this paper were produced by the computer program PLOTLOG (Thissen, 1992).

Authors' Address

Send requests for reprints or further information to Allan S. Cohen, University of Wisconsin, 1025 W. Johnson, Madison WI 53706, U.S.A. Internet: cohen@tne.edsci.wisc.edu.