A Conceptual Analysis of Differential Item Functioning in Terms of a Multidimensional Item Response Model

Gregory Camilli
Rutgers, The State University of New Jersey

Differential item functioning (DIF) has been informally conceptualized as multidimensionality. Recently, more formal descriptions of DIF as multidimensionality have become available in the item response theory literature. This approach assumes that DIF is not a difference in the item parameters of two groups; rather, it is a shift in the distribution of ability along a secondary trait that influences the probability of a correct item response. That is, one group is relatively more able on an ability such as test-wiseness. The parameters of the secondary distribution are confounded with item parameters by unidimensional DIF detection models, and this manifests as differences between estimated item parameters. However, DIF is confounded with impact in multidimensional tests, which may be a serious limitation of unidimensional detection methods in some situations. In the multidimensional approach, DIF is considered to be a function of the educational histories of the examinees. Thus, a better tool for understanding DIF may be provided through structural modeling with external variables that describe background and schooling experience. *Index terms: differential item functioning, factor analysis, IRT, item bias, LISREL, multidimensionality.*

A test item is considered to function differently for two groups if the probability of a correct response is associated with group membership for examinees of comparable ability (Holland & Thayer, 1988). In this case, it is natural to assume that there are one or more variables, in addition to the target ability, that account for or explain a group difference in item performance. From this perspective, differential item functioning (DIF) is evidence of additional abilities.

In this paper, a mathematical model is proposed to describe how group differences in distributions of abilities, which are distinct from the target ability, influence the probability of a correct item response. These abilities are referred to as secondary abilities and may or may not be correlated with the target or primary ability. The groups being compared are referred to as the reference and focal groups (Holland & Thayer, 1988). In practice, the reference group is often a majority group and the focal group is a minority group.

There are reasons other than multiple abilities (or multidimensionality) that give rise to DIF. Unique linguistic training can result in a noun behaving as a false cognate. For example, Italians use the phrase "incidente stradale" to signify a car accident. DIF could result from the use of the word "incident" in English test items with Italian bilinguals, but it is awkward to call this linguistic difference a "secondary ability." Estimation errors in item parameters may also lead to the appearance of DIF. However, this paper is concerned with the conceptual value that a formal mathematical structure imparts to DIF. The advantage to this approach is that complex observed phenomena can be analyzed in terms of how they correspond to relatively simple features of a model that contains equations and coefficients that can be used to make deductions. Thus, a simple change in one group's covariance

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 2, June 1992, pp. 129-147

0146-6216/92/020129-19$2.20

structure might "predict" numerous complex changes in unidimensional estimates of item parameters that are used to measure DIF.

Often used interchangeably with DIF, the term *item bias* usually implies qualitative aspects of test items, in addition to DIF. If DIF is observed and can be attributed to a feature of the item irrelevant to the test construct, then the item may have a biasing effect on estimated ability. An evaluation of the fairness of the item requires further information about the test and the context of the test's use. DIF, rather than bias or fairness, is the focus of this paper. Nonetheless, a better statistical description of DIF should facilitate its explanation.

Nonparametric approaches to DIF, such as the Mantel-Haenszel (MH) procedure, are now widely known and easily implemented. There are two important reasons, however, for studying DIF in terms of theoretical models. First, when concerned with the efficacy of a nonparametric technique, it is helpful to evaluate the technique against a variety of item response theory (IRT) models. Because IRT models have been shown to describe real test item responses accurately, they serve as an "important testing ground" (Holland & Thayer, 1988, p. 143). For example, Holland and Thayer showed that with the assumption of a Rasch model, the MH log-odds ratio is an unbiased estimate of the difference in *b* parameters for two groups. But Swaminathan and Rogers (1990) showed that the MH technique may be ineffective with two-parameter IRT models in which two groups have crossing item response functions (IRFs). Thus, arguing that any particular method is appropriate for detecting DIF is equivalent to rejecting certain classes of IRT models as plausible descriptions of a given set of item responses. Second, contingency table methods, like most statistical procedures, assume conditional independence of observations. Thus, if two or more conditioning variables are necessary, but only a single variable (say total score) is used, then it is likely that this assumption is violated. In IRT, this is the assumption of local independence. This latter problem is usefully analyzed within the context of a multidimensional IRT model.

Many statistical techniques for detecting DIF have been assessed exclusively within the framework of unidimensional IRT models (Shepard, Camilli, & Williams, 1985; Thissen, Steinberg, & Wainer, 1988). In contrast to the multiple ability perspective, DIF is operationalized with unidimensional models as a difference in the IRFs for two groups. This difference can be represented in terms of item parameters, such as difficulty or discrimination. However, two important premises of this approach are that (1) the target ability (Shealy & Stout, in press) is statistically extracted from the item responses, and (2) the groups in question are anchored over this target ability—not just placed on the same standardized scale—before comparing IRFs or item parameters. That is, in both groups the probability of a correct response must be a function of the identical target ability. This approach has a number of weaknesses:

1. If more than one ability contributes to item responses, then unidimensional models estimate a composite of the underlying abilities. Thus, groups cannot be anchored over a single target ability for a valid comparison of IRFs.
2. If a test is unidimensional within two groups, but some item parameters differ, this suggests that different abilities are being assessed. A single ability for anchoring IRFs does not exist. In this case, DIF is not well defined (Crocker & Algina, 1986, p.378). (By fiat, a kind of multidimensionality is created by combining two different unidimensional structures; however, this paper is concerned with more complex instances of multidimensionality.)
3. Unidimensional models do not provide a useful mechanism to aid in the interpretation of DIF. In contrast, a multidimensional model suggests that it is prudent to isolate and interpret secondary abilities. This search focuses on the abilities of examinees, not solely on the properties of test items.

The purpose of this paper is to review multidimensional item response models as they relate to DIF and to describe how unidimensional techniques for detecting DIF work with multidimensional data. One important area of investigation concerns the distinction between DIF and impact. If a group difference in item performance is computed for examinees of comparable ability, the net result is DIF; however, if differences are computed for unmatched examinees the result is impact. More formally, impact is the portion of a group difference in item performance that can be attributed to the group difference in the target ability. The significance of this distinction is discussed elsewhere (Angoff, 1982; Camilli, in press; Holland & Thayer, 1988). By using multidimensional IRT models as a theoretical testing ground, it is shown below that DIF and impact are not separated into tidy parcels by unidimensional techniques, including the MH procedures; rather, they are intricately confounded. Nonparametric approaches to DIF that make use of the total test score do not automatically circumvent this problem. With a multidimensional test, the total score is some weighted function of all underlying abilities. Consequently, the average total score for a group, as well as the difference in group averages, are both functions of primary and secondary abilities. It is shown below that impact is confounded with DIF whenever underlying abilities are correlated.

The total score is often used for establishing comparability (the conditioning variable) in addition to providing a point estimate of ability, but it may not be sufficient for establishing independence. Consequently, for examinees with the same total score, group differences in primary ability may still exist and may also contribute to the statistical estimate of DIF. (Examinees from different demographic groups can arrive at the same total score with different mixtures of underlying abilities.) Hunter (1975) showed how unidimensional techniques can lead to false impressions of DIF and argued more generally that "no conceivable methods that ignore the structure of the test can work" (p. 2).

This paper extends previous work on multidimensionality, in particular that of Wang (1985) and Shealy and Stout (in press). Wang showed how multidimensional item parameters are estimated (or expressed) by unidimensional IRT methods, with applications to both DIF and test equating. Shealy and Stout showed how DIF and test bias could be formulated in terms of a general multidimensional model for test items with monotone IRFs. In contrast to the latter work, this paper examines multidimensionality in terms of a parametric structure both on IRFs (for which a normal or logistic ogive is assumed) and on latent abilities (for which a multivariate normal distribution is assumed). The advantages of a parametric approach include the relative simplicity of mathematical results and the use of these results for developing applications in terms of existing multidimensional software.

A number of terms and concepts from factor analysis are used in this paper. *Unidimensional item structure* refers to a set of factor coefficients in which each item loads on a single common factor, and *multidimensional item structure* refers to a set that contains item loadings on two or more common factors. It will become apparent that these loadings, or pattern coefficients, are the discrimination parameters of the items. Other terms are used from regression analysis because the factor models employed below are fundamentally standard linear models.

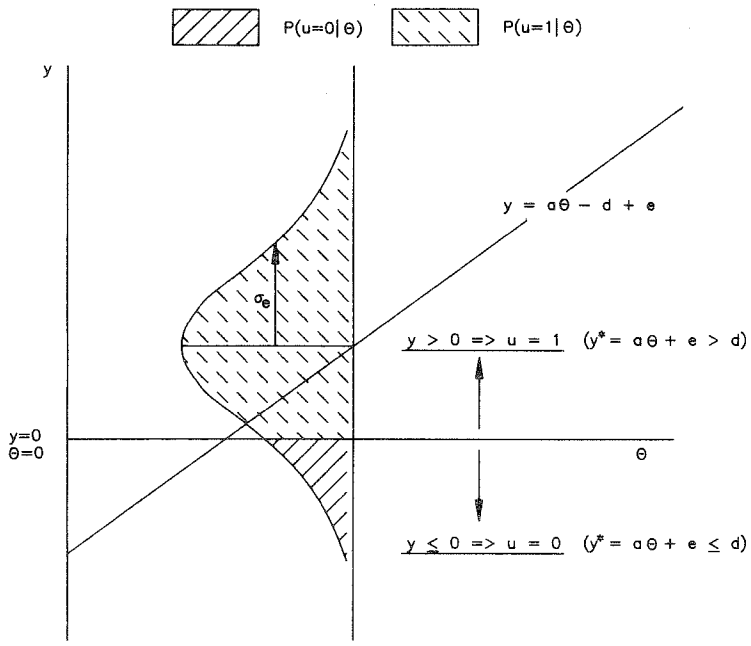
A Two-Dimensional Item Response Model

A number of models have been proposed for multidimensional item structures. The most widely known of these are the compensatory (Bock & Aitkin, 1981; Samejima, 1974; Torgerson, 1958), and the noncompensatory (Sympson, 1978) models. Because compensatory models are of special interest to DIF studies, and the noncompensatory model can be closely approximated by a compensatory model (Wang, 1985), only the latter is considered here. The normal ogive form without a guessing parameter for an item with M factors is

$$P(u = 1|\Theta) = \Phi\left(\sum_{i=1}^{M} a_i\theta_i - d\right) \quad , \tag{1}$$

where $\Phi(\cdot)$ is the normal distribution function. The model is an application of a simple linear equation relating a propensity (or probit) score y, to a set of underlying abilities. In this model, y represents a continuous measurement containing error [Thurstone's (1927) "discriminal process"], because it is a degree of ability that is assessed by a particular (fallible) item—it is an artifact of the measurement process that only a dichotomous response is observed. The probit continuum is dichotomized at 0 to model an observed correct-incorrect item response; that is, for example, $u = 1$ (for $y > 0$) and $u = 0$ (for $y \leq 0$) model a response on a multiple-choice test. Figure 1 provides an illustration

Figure 1
Diagram of the Slope/Intercept Item Response Model for One Latent Ability, θ (The Point $y = 0$ Dichotomizes the y Continuum to Model Observed Correct/Incorrect Responses)



of a model with a single latent ability. In the case of two latent abilities (say θ_1 and θ_2), the latent model for an item in slope/intercept form is

$$y = a_1\theta_1 + a_2\theta_2 - d + e \quad , \tag{2}$$

where σ_e^2 is the variance of measurement errors, d is the intercept or item difficulty, and a_1 and a_2 are slope coefficients, or discriminations, that are scaled so that $\sigma_e^2 = 1$ (see Bock & Aitkin, 1981, pp. 455–456; Lord & Novick, 1968, p. 375). (The probit distributions move up and down the y axis depending on the intercept, d. Therefore, distributions for different items are dichotomized at different points.) The logistic form of this item response model is more common, but with the introduction of a scaling constant, the two models give virtually identical results (Birnbaum, 1968). Note that with more than one latent ability, the more common expression of the model embodying $a(\theta - b)$ cannot be used.

This model has interesting implications for DIF studies because it allows an examinee with a higher level of one type of ability to compensate for a lower level of another type. For example, test-wise examinees may use this ability (θ_2) to compensate for lack of reading ability (θ_1) on a test item, and the level of test-wiseness may vary across demographic groups. As a result, DIF is observed. The converse of this property is that a very low level of one ability can deflate moderately high levels of another. For example, poor reading ability is likely to interfere with solving math word problems. Thus, such models provide a mechanism to study factors that are conceptually distinct from the target ability, but may have a significant compensatory effect. This model can easily be extended to include a number of secondary traits; but for the sake of simplicity, most examples in this paper emphasize the two-factor model with one target and one secondary ability.

DIF as Multidimensionality

If multidimensionality holds, the definition of DIF given by Kok (1988, p. 269) and Shealy and Stout (in press) follows. Specifically, the multidimensional definition is that

$$\varepsilon_{\theta_2}[P_1(u = 1|\theta_1,\theta_2)|\theta_1] \neq \varepsilon_{\theta_2}[P_2(u = 1|\theta_1,\theta_2)|\theta_1] \quad , \tag{3}$$

where P_i signifies the probability of a correct answer in group i, θ_1 is the target ability, and θ_2 is a factor other than θ_1 that affects the correctness of an item response. (Note that θ_2 could be a vector of several secondary abilities, for example, test-wiseness and the ability to work quickly.) Informally, this signifies that for any two groups of examinees, all of whom have the same value of θ_1, the influence of the second ability (θ_2) on a correct response is different. That is, θ_2 provides a relative advantage for one of the groups. Equality in Equation 3 is maintained only when

$$G_1(\theta_2|\theta_1) = G_2(\theta_2|\theta_1) \quad , \tag{4}$$

where G_i signifies the distribution of θ_2 for examinees with fixed values of θ_1; that is, the conditional distribution of θ_2. On the other hand, DIF is encountered when $G_1 \neq G_2$. In this case, θ_2 does not have equivalent effects for persons of Groups 1 and 2 with equal θ_1. The central feature of the multidimensional approach is that DIF does not manifest itself as differences between item parameters (a, b, or c, for example), but as differences between the parameters of G_1 and G_2.

More specifically, assume that for a given group the regression of θ_2 on θ_1 is given by

$$\theta_2 = \nu + \beta\theta_1 + \epsilon \quad , \tag{5}$$

where ν is the intercept and β is the slope. The mean and variance of θ_2 for a fixed value of θ_1 are

$$\mu' = \nu + \beta\theta_1 \tag{6}$$

and

$$\sigma^2_\epsilon = \sigma^2_2(1 - \rho^2) \quad , \tag{7}$$

where σ^2_2 is the variance of θ_2 and ρ is the correlation of θ_1 with θ_2. If $\theta_2|\theta_1$ is normally distributed within groups, then its distribution G is completely described by the parameters μ' and σ^2_ϵ. Therefore, DIF can be defined as failure of one or both of the following equalities to hold across two groups:

$$\mu'_1 = \mu'_2 \tag{8}$$

and

$$\sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2} \quad . \tag{9}$$

When either of the above conditions does not hold, differences in the estimated unidimensional item parameters may be created, but such differences are only indications that marginal distributions are not equivalent. Equivalence problems arise when the item parameters change. In this case, it can be argued that different abilities are being measured, and the comparison of item parameters can be likened to an "apples and oranges" problem. Indeed, only if the measuring mechanism remains more or less constant across groups is there a common denominator for the comparison. When item responses are unidimensional with different IRFs, the target abilities are different (Crocker & Algina, 1986).

It should be noted, however, that Kok's (1988) definition does not demonstrate how differences in the groups' marginal distributions give rise to empirical measures of DIF obtained with unidimensional detection methods. As will be demonstrated below, the differences in Equations 8 and 9 each make a unique contribution. With a multidimensional test, item parameters estimated with unidimensional IRT software may differ for two groups because they are confounded with the parameters of G_1 and G_2. In this case, unidimensional techniques are useful to the degree that they provide a reasonable substitute for methods that directly estimate the parameters of the latent ability distributions.

A more familiar description of the conditions that may lead to DIF is based on the analysis of covariance (ANCOVA) model. Let the dependent variable be defined as the secondary ability (θ_2) and the covariate as the primary ability (θ_1). In standard ANCOVA, group means on θ_2 are adjusted for initial systematic differences in θ_1, assuming homogenous within-group regressions. The average difference between groups is then given by the difference in intercepts, $\nu_1 - \nu_2$, and is uniform throughout the range of θ_1. As applied to DIF analysis, a nonzero value of $\nu_1 - \nu_2$ means that the primary ability cannot account for group differences on other factors: A potential advantage remains for one group at all levels of θ_1. However, a more complicated model results if regression slopes are not equal, in which case the relative advantage varies across the range of θ_1. Other complex situations arise when the conditional variances are not homogenous across groups, or when they are heteroscedastic across the range of θ_1. Note that because items tend to discriminate only in a relatively narrow band of ability, DIF is highly interactive with all models except the uniform $\nu_1 - \nu_2$ ANCOVA model. Although this is a complex analogy, ANCOVA theorists have studied such biases in statistical adjustments for some time (Cronbach, Ragosa, Floden, & Price, 1977; Ragosa, 1980). Thus, DIF can be conceived as the failure of the primary ability to control group differences on secondary abilities.

Matrix Formulation of the General Case

Suppose a test is composed of K items. The terms in the general (latent) item response model with M dimensions are defined as follows:

$\mathbf{Y}$ is a $K \times 1$ column vector of probit scores on each item,
$\mathbf{A}$ is a $K \times M$ matrix of slope or discrimination parameters,
$\boldsymbol{\Theta}$ is a $M \times 1$ column vector of latent abilities,
$\mathbf{D}$ is a $K \times 1$ column vector of difficulty parameters, and
$\mathbf{E}$ is a $K \times 1$ column vector of random measurement errors.

The general latent model for an examinee can then be written as

$$\mathbf{Y} = \mathbf{A}\boldsymbol{\Theta} - \mathbf{D} + \mathbf{E} \tag{10}$$

and

$$\hat{\mathbf{Y}} = \mathbf{A}\boldsymbol{\Theta} - \mathbf{D} \quad , \tag{11}$$

where a row of $\mathbf{A}$ contains the M discriminations for a particular item, $\boldsymbol{\Theta}$ contains the M latent abilities, and $\hat{\mathbf{Y}}$ is the vector of predicted or expected item responses.

The mean vector ($\boldsymbol{\mu}_Y$) and variance-covariance matrix ($\boldsymbol{\Omega}_Y$) of $\mathbf{Y}$ are then given as expectations over examinees

$$\boldsymbol{\mu}_Y = \varepsilon(\mathbf{Y}) = \mathbf{A}\boldsymbol{\mu}_\Theta - \mathbf{D} \tag{12}$$

and

$$\boldsymbol{\Omega}_Y = \varepsilon[(\mathbf{Y} - \boldsymbol{\mu}_Y)(\mathbf{Y} - \boldsymbol{\mu}_Y)'] = \mathbf{A}\boldsymbol{\Omega}_\Theta\mathbf{A}' + \boldsymbol{\Omega}_E \quad , \tag{13}$$

where $\varepsilon[\mathbf{EE}'] = \boldsymbol{\Omega}_E$. By assumption, $\varepsilon[\mathbf{E}] = 0$ and $\boldsymbol{\Omega}_E$ is a diagonal matrix, that is, the measurement errors are uncorrelated across item responses.

A One-Dimensional Approximation to Multifactor Item Response Models

Ability

The vector $\mathbf{Y}$ contains the probit scores y_j for individual items. Although y_j is a continuous variable, it is unobserved: only a dichotomy along y_j is observed. That is, the observed item response, u_j, is a 0-1 variable. Theoretically, it is preferable to compute and factor analyze $\boldsymbol{\Omega}_Y$, the covariance matrix of the y_j, rather than that of the u_j. This is the motivation for full-information factor analysis (Bock, Gibbons, & Muraki, 1988). The objective here would be to extract the structure

$$\boldsymbol{\Omega}_Y = \mathbf{A}\boldsymbol{\Omega}_\Theta\mathbf{A}' + \boldsymbol{\Omega}_E \quad . \tag{14}$$

A number of studies have concluded that unidimensional models extract an ability that is a mixture of the underlying dimensions (Ackerman, 1989; Dorans & Kingston, 1985; Reckase, 1979). Wang (1985) demonstrated that unidimensional models appear to extract the first factor of the matrix $\boldsymbol{\Omega}_Y$, and that the first factor could in turn be represented as a weighted combination of latent abilities, labeled the *reference composite.*

Wang defined the reference composite as the factor score corresponding to the first principal component of $\mathbf{A}\boldsymbol{\Omega}_\Theta\mathbf{A}'$. (Actually, this is the expected value of the reference composite because the error component $\boldsymbol{\Omega}_E$ is being ignored in this treatment for the sake of simplicity.) That is, the reference composite is the score corresponding to the first principal component of

$$\begin{aligned}\boldsymbol{\Omega}_{\hat{Y}} &= \varepsilon[(\hat{\mathbf{Y}} - \boldsymbol{\mu}_Y)(\hat{\mathbf{Y}} - \boldsymbol{\mu}_Y)'] \\ &= \boldsymbol{\Omega}_Y - \boldsymbol{\Omega}_E \\ &= \mathbf{A}\boldsymbol{\Omega}_\Theta\mathbf{A}' \quad .\end{aligned} \tag{15}$$

Define the following

$$\varepsilon(\boldsymbol{\Theta}) = \boldsymbol{\mu} \tag{16}$$

and

$$\varepsilon[(\boldsymbol{\Theta} - \boldsymbol{\mu})(\boldsymbol{\Theta} - \boldsymbol{\mu})'] = \boldsymbol{\Omega}_\Theta = \mathbf{TT}' \quad , \tag{17}$$

where the matrix $\mathbf{T}^{-1}$ transforms the (possibly) correlated variables in $\boldsymbol{\Theta} - \boldsymbol{\mu}$ to uncorrelated variables, that is, $\mathbf{T}^{-1}$ normalizes $\boldsymbol{\Theta}$ (see Winer, 1971, section 2.6). In particular, let the Cholesky decomposition of $\boldsymbol{\Omega}_\Theta$ be given by $\boldsymbol{\Omega}_\Theta = \mathbf{TT}'$. (Any factorization works; however, the Cholesky method

preserves the relative signs of the composite variables described below.) Assume that $\boldsymbol{\Omega}_{\hat{\mathbf{Y}}}$ is not full rank, that is, $M < K$. The diagonalization of $\boldsymbol{\Omega}_{\hat{\mathbf{Y}}}$ can then be represented as

$$\mathbf{U}'\boldsymbol{\Omega}_{\hat{\mathbf{Y}}}\mathbf{U} = \begin{pmatrix}\mathbf{D}^{1/2}\\ 0\end{pmatrix}(\mathbf{D}^{1/2}\ 0) \quad . \tag{18}$$

Now according to the Eckert-Young theorem (see Mulaik, 1972), if $\mathbf{V}$ diagonalizes $\mathbf{T}'\mathbf{A}'\mathbf{A}\mathbf{T}$ then

$$\mathbf{V}'[\mathbf{T}'\mathbf{A}'\mathbf{A}\mathbf{T}]\mathbf{V} = \mathbf{D} \tag{19}$$

and

$$\mathbf{U}' = \mathbf{D}^{-1/2}\mathbf{V}'\mathbf{T}'\mathbf{A}' \quad , \tag{20}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices. The vector of factor scores of $\boldsymbol{\Omega}_{\hat{\mathbf{Y}}}$ is then

$$\boldsymbol{\Upsilon} = \mathbf{D}^{-1/2}\mathbf{U}'(\hat{\mathbf{Y}} - \boldsymbol{\mu}_{Y}) \quad . \tag{21}$$

On substituting Equation 20 for $\mathbf{U}'$ in Equation 21 and simplifying,

$$\boldsymbol{\Upsilon} = \mathbf{W}'(\boldsymbol{\Theta} - \boldsymbol{\mu}_{\Theta}) \quad , \tag{22}$$

where

$$\mathbf{W}' = \mathbf{V}'\mathbf{T}^{-1} \quad , \tag{23}$$

and

$$\mathbf{W}'^{-1} = \mathbf{T}\mathbf{V} \quad . \tag{24}$$

The expected reference composite is then determined by the weights in the first column of $\mathbf{W}$. In the two-dimensional case, $\boldsymbol{\Upsilon}$ is given by

$$\boldsymbol{\Upsilon} = \begin{pmatrix}\upsilon_1\\ \upsilon_2\end{pmatrix} = \begin{bmatrix}w_{11}(\theta_1 - \mu_{\theta_1}) + w_{12}(\theta_2 - \mu_{\theta_2})\\ w_{21}(\theta_1 - \mu_{\theta_1}) + w_{22}(\theta_2 - \mu_{\theta_2})\end{bmatrix} \quad , \tag{25}$$

where w_{ij} is an element of $\mathbf{W}$. The factor score υ_1 is the reference composite, and the second factor score υ_2 is orthogonal to υ_1.

Item Parameters

The item parameters extracted by a unidimensional IRT model can be represented as the expected value of an IRF anchored over the reference composite υ_1 with respect to the remaining υ_i. In the two-dimensional case, this is

$$\varepsilon_{\upsilon_2}[P(u = 1|\upsilon_1,\upsilon_2)|\upsilon_1] = \int_{-\infty}^{+\infty} P(u = 1|\upsilon_1,\upsilon_2)G(\upsilon_2|\upsilon_1)d\upsilon_2 \quad . \tag{26}$$

The vector of discriminations for a single item can be defined as the row vector $\mathbf{A} = (a_1, a_2)$. The two-dimensional probit model then can be reparameterized as

$$\begin{aligned} a_1\theta_1 + a_2\theta_2 - d + e &= \mathbf{A}\boldsymbol{\Theta} - d + e\\ &= \mathbf{A}(\boldsymbol{\Theta} - \boldsymbol{\mu}_{\Theta}) + \mathbf{A}\boldsymbol{\mu}_{\Theta} - d + e\\ &= \mathbf{A}\mathbf{W}'^{-1}\mathbf{W}'(\boldsymbol{\Theta} - \boldsymbol{\mu}_{\Theta}) - (d - \mathbf{A}\boldsymbol{\mu}_{\Theta}) + e\\ &= \mathbf{A}\mathbf{W}^{-1'}\boldsymbol{\Upsilon} - (d - \mathbf{A}\boldsymbol{\mu}_{\Theta}) + e\\ &= \boldsymbol{\Lambda}\boldsymbol{\Upsilon} - (d - \mathbf{A}\boldsymbol{\mu}_{\Theta}) + e\\ &= \boldsymbol{\Lambda}\boldsymbol{\Upsilon} - \delta + e \quad . \end{aligned} \tag{27}$$

The vector $\mathbf{\Lambda}$ contains the reparameterized slope coefficients

$$\mathbf{\Lambda} = \begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix} = \begin{pmatrix}a_1w^{11} + a_2w^{12}\\ a_1w^{21} + a_2w^{22}\end{pmatrix} \quad , \tag{28}$$

where w^{ij} is an element of $\mathbf{W}^{-1}$. Given $\mathbf{\Lambda}$, Υ, δ, the solution to the integral in Equation 26 after simplification (see Lord & Novick, 1968, pp. 376–377) gives the following one-dimensional approximations:

$$\alpha = \frac{\lambda_1 + \beta\lambda_2}{[1 + \lambda_2^2\sigma_2^2(1 - \rho^2)]^{1/2}} \tag{29}$$

and

$$\delta = \frac{(d - \mathbf{A}\mu_\Theta) - \lambda_2(\mu_2 - \beta\mu_1)}{\lambda_1 + \beta\lambda_2} \quad , \tag{30}$$

where β is the regression slope of υ_2 on υ_1,
ρ is the correlation of μ_2 with μ_1,
σ_2^2 is the variance of υ_2, and
μ_i is the mean of υ_i.

By use of Equations 22, 23, and 24 it can be shown that $\varepsilon[\Upsilon] = 0$ and $\varepsilon[\Upsilon\Upsilon'] = I$. Thus, $\beta = \rho = \mu_i = 0$ and $\sigma_2^2 = 1$. Equations 29 and 30 then simplify to

$$\alpha = \frac{\lambda_1}{(1 + \lambda_2^2)^{1/2}} \tag{31}$$

and

$$\delta = \frac{d - \mathbf{A}\mu_\Theta}{\lambda_1} \quad . \tag{32}$$

Wang (1985) derived the equations for the general case of M dimensions. Ackerman (1988) used Wang's results to predict DIF in simulated tests, but did not attempt to clarify mathematically how latent distributions affect group differences in α and δ.

The one-dimensional approximation to the normal ogive can then be given in the standard form

$$\varepsilon_{\upsilon_2}[P(u = 1|\upsilon_1,\upsilon_2)|\upsilon_1] = P(u = 1|\upsilon_1) = \Phi[\alpha(\upsilon_1 - \delta)] \quad . \tag{33}$$

Now υ_1 can be written as

$$\begin{aligned}\upsilon_1 &= w_{11}(\theta_1 - \mu_{\theta_1}) + w_{12}(\theta_2 - \mu_{\theta_2})\\ &= (w_{11}\theta_1 + w_{12}\theta_2) - (w_{11}\mu_{\theta_1} + w_{12}\mu_{\theta_2})\\ &= \upsilon_1^* - \mu_{\upsilon_1} \quad ,\end{aligned} \tag{34}$$

and Equation 33 can be rewritten with this substitution to give

$$P(u = 1|\upsilon_1) = \Phi\{\alpha[\upsilon_1^* - (\delta + \mu_{\upsilon_1})]\} \quad . \tag{35}$$

The above substitution facilitates comparisons of the latent models across groups because it results in the IRFs being anchored over the untransformed reference composite υ_1^* rather than the group deviation composite υ_1.

The Comparison of Unidimensional Item Parameters

Equation 35 provides the basic mechanism for comparing the parameters of IRFs for two groups approximated by unidimensional models. Let $\boldsymbol{\Omega}_{\Theta_i}$ be the covariance matrix of $\boldsymbol{\Theta}$ for group i, and assume for the moment that $\boldsymbol{\Omega}_{\Theta_1} \neq \boldsymbol{\Omega}_{\Theta_2}$. Label $\mathbf{W}$, the matrix of composite weights, as $\mathbf{W}_i$ for group i. Because $\mathbf{W}_i$ depends on $\boldsymbol{\Omega}_{\Theta_i}$, the composite weights will also differ according to group. That is, each group's reference composite represents a different weighted composite of the underlying abilities. The differences in difficulty and discrimination parameters can then be given as

$$\Delta\alpha = \alpha_1 - \alpha_2 \tag{36}$$

and

$$\Delta\delta = [\delta_1 + \mu_{\upsilon_{1(1)}}] - [\delta_2 + \mu_{\upsilon_{1(2)}}] \quad . \tag{37}$$

It is important to note that if $\boldsymbol{\Omega}_{\Theta_1} \neq \boldsymbol{\Omega}_{\Theta_2}$, then $\Delta\alpha$ and $\Delta\delta$ are not well defined because group IRFs are anchored over different reference composites. The scales of υ_1 and υ_2 should be linked to provide comparable item parameter estimates. This can be done by scaling both with unit variances, but this scaling does not yield equal units on a common construct.

In practice, item parameters are linked either through anchor items (Lord, 1980; Thissen, Steinberg, & Wainer, in press) or with ad hoc linear transformations (Linn, Levine, Hastings, & Wardrop, 1980; Stocking & Lord, 1983). The anchor item method would extract an average or common reference composite based on either the combined or pooled groups estimate of $\boldsymbol{\Omega}_{\Theta}$, because item parameters for the reference and focal groups would be simultaneously calibrated. A combined groups estimate of $\boldsymbol{\Omega}_{\Theta}$ confounds within and between covariances and results in potentially large distortions in $\mathbf{W}$, leading to distortions in the reference composite and estimates of $\Delta\delta$. In this case, $\mathbf{W}$ is based in part on the covariance matrix

$$\boldsymbol{\Omega}_{\Theta_{1+2}} = p\boldsymbol{\Omega}_{\Theta_1} + q\boldsymbol{\Omega}_{\Theta_2} + pq(\boldsymbol{\mu}_{\Theta_1} - \boldsymbol{\mu}_{\Theta_2})(\boldsymbol{\mu}_{\Theta_1} - \boldsymbol{\mu}_{\Theta_2})' \quad , \tag{38}$$

where p is the proportion of examinees in group 1, and $q = 1 - p$. A calibration analysis that combines groups, but also estimates latent population parameters separately for both groups, is perhaps less prone to confounding $\mathbf{W}$ with differences in group centroids.

Both types of confounding described above will not necessarily distort $\Delta\alpha$, because the αs depend only on the common reference composite and invariant structural coefficients (a_i). The effects of ad hoc linking of separate-sample calibrations are more difficult to characterize; however, they are almost certain to produce DIF in the α parameters when group covariances on Θ are not equal (as in Equation 31). Indeed, a major source of ambiguity in applied studies of DIF is the use of linked separate-sample calibrations—even with anchor test methods, differences in latent population distributions are rarely taken into account. This problem was first noted by Wang (1985).

Now assume that group covariances are equal so that $\mathbf{W}_1 = \mathbf{W}_2 = \mathbf{W}$. Equation 37 can now be simplified in the two-dimensional case to give

$$\Delta\delta = (a_1/\lambda_1 - w_{11})[\mu_{1(2)} - \mu_{1(1)}] + (a_2/\lambda_1 - w_{12})[\mu_{2(2)} - \mu_{2(1)}] \quad , \tag{39}$$

where $\mu_{i(j)}$ is the mean on θ_i for group j. Also, w_{11} and w_{12} are the weights for the reference composite, and w_{21} and w_{22} are weights for the variate orthogonal to the reference composite. It is important to note that DIF in the difficulty parameter is a weighted function of group differences on θ_1 and θ_2. Thus, unidimensional estimates of ability confound DIF with group differences on the target ability (θ_1) when multidimensionality holds. This is because the term $\mu_{1(2)} - \mu_{1(1)}$, which appears in the

theoretical formula for DIF given in Equation 39, represents impact—the unmatched group difference in target ability. Items that show no DIF will satisfy the equation $\Delta\delta = 0$. Although the general circumstances in which this is true are complex, a sufficient condition derived from Equation 39 is that

$$\frac{a_1}{a_2} = \frac{w_{11}}{w_{12}} \text{ and } w_{12} = w_{21} \quad . \tag{40}$$

The first part of this compound condition indicates that the ratio of the actual discriminations must equal the ratio of reference composite weights. It can be shown that DIF is not confounded with group differences on the target ability when this first part is satisfied. Reckase, Ackerman, and Carlson (1988) showed that a simulated test comprised of such items is statistically indistinguishable from a unidimensional test which, from the multidimensional perspective, cannot show DIF. With an actual test, however, it seems very unlikely that any given item will satisfy this constraint. The second part of the above condition suggests that symmetry of the matrix W plays an interesting role in determining DIF, but this issue was not investigated here.

It is useful to show how DIF would be formulated mathematically if a set of anchor items could be obtained that was unidimensional in the target ability, θ_1, because this would solve the comparability problem. A number of researchers have suggested the use of a short subtest as a proxy for target ability. This has been referred to as the "valid subtest" (Shealy & Stout, in press) and the "designated anchor" (Thissen et al., in press). However, although Shealy and Stout required the subtest to be unidimensional (depending only on the target ability), Thissen et al. suggested that the subtest items could be selected partially on the basis of policy demands. It has also been suggested that the designated anchor must reflect test content and, therefore, test dimensionality (Wang, 1985). It is interesting to note that the designated anchor may have requirements similar to common items in common-item equating designs.

To demonstrate how DIF would be defined mathematically with a unidimensional anchor, the quantity $\Delta\delta$ can be calculated by obtaining the W matrix (the composite weights) so that it satisfies the conditions

$$\upsilon_1 = (\theta_1 - \mu_1)/\sigma_1 \quad , \tag{41}$$

$$\varepsilon[\Upsilon] = 0 \quad , \tag{42}$$

and

$$\varepsilon[\Upsilon\Upsilon'] = \mathbf{I} \quad . \tag{43}$$

In this case

$$\mathbf{W}' = \begin{pmatrix} \sigma_1^{-1} & 0 \\ -\beta\sigma_\epsilon^{-1} & \sigma_\epsilon^{-1} \end{pmatrix} \quad , \tag{44}$$

where $\sigma_\epsilon = \sigma_2(1 - \rho^2)^{1/2}$ and β is the regression of θ_2 on θ_1. Assuming that group covariances are equal, the following result is obtained when the appropriate substitutions are made into Equation 30:

$$\begin{aligned} \Delta\delta &= \sigma_1^{-1}\left(\frac{a_2}{a_1 + \beta a_2}\right)\{-\beta[\mu_{1(2)} - \mu_{1(1)}] + [\mu_{2(2)} - \mu_{2(1)}]\} \\ &= \sigma_1^{-1}\left(\frac{a_2}{a_1 + \beta a_2}\right)[\nu_2 - \nu_1] \quad . \end{aligned} \tag{45}$$

This result indicates that even if IRFs could be anchored over the target ability, DIF in the difficulty parameter is still confounded, in an absolute sense, with the target ability. This will be the case unless θ_1 and θ_2 are uncorrelated, or unless Equation 8 holds, in which case the DIF is uniformly 0 across items. However, only the terms involving *a* parameters in Equation 45 vary across items, though the potential for DIF remains uniform over θ_1. The *relative variation* is solely a function of the ratio a_2/a_1. This ratio acts as a lens through which $\upsilon_2 - \upsilon_1$ is magnified or reduced. All things considered, this is not an unreasonable index of DIF. However, if the within-group regressions of θ_2 on θ_1 are not parallel, DIF indices may vary systematically with group differences on θ_1 across different samples of reference and focal group examinees.

If it is not assumed that σ_ϵ is equivalent across groups, then even with an anchor test the analysis of DIF becomes much more complicated. The mathematical expression for DIF in this case is not given here, but a basic outline for its derivation is provided. First, obtain the composite transformation matrices and their inverses separately for both groups. Second, obtain λ_1 and λ_2 with $\sigma_{\epsilon_1} \neq \sigma_{\epsilon_2}$. Then substitute these terms into Equations 28 through 30. The resulting expression contains a number of ratios in which the σ_ϵ and *d* terms appear in both the numerators and denominators. Thus, the net effect is not easily characterized. A similar derivation could be performed assuming $\sigma_{\epsilon_1} = \sigma_{\epsilon_2}$, but that $\beta_1 \neq \beta_2$.

A Numerical Example Using Simulation

This section demonstrates the use of LISREL (Jöreskog & Sörbom, 1989) for estimating the parameters of the latent ability distributions for two or more groups. With observed variables, the required input consists of within-group correlations, variances, and centroids. This is problematic with dichotomous item responses because the continuous metric is not observed. Methods for approximating these quantities are discussed, as well as how to specify a model so that LISREL will estimate $\boldsymbol{\mu}_\Theta$ and $\boldsymbol{\Omega}_\Theta$.

The computer program LISCOMP (Muthén, 1987) has recently become available for multigroup factor analysis with latent means. The mathematical foundation for this program was given by Muthén and Christofferson (1981), and a numerical example for one latent factor was given by Muthén and Lehman (1985). The method presented below is highly similar to that employed by LISCOMP. However, the current method permits a more didactic presentation of DIF as a multigroup modeling problem. It is also more general in the sense that it can be applied in situations in which item formats allow guessing, though in this case both tetrachoric correlations and item *p* values (described below) must be corrected for chance. Nonetheless, LISCOMP should be considered the best available choice in situations without guessing and with smaller numbers of items.

Data Generation

Multidimensional procedures were applied to randomly generated item responses in a simulation study to demonstrate that certain procedures can recover the parameters of latent distributions with simplified test data. In the first step of this analysis, item responses were generated for a two-factor model with 10 items for a reference and a focal group. The reference group had means (0.0, 0.0) and standard deviations (SDs) (1.0, 1.0) on the dominant and secondary abilities, respectively, and the focal group had means (0.0, –1.0) and SDs (.5, 1.5). The abilities were uncorrelated. This model represented a situation with one dominant factor and a weaker secondary factor. Its two key features were that there was no guessing and that items were equivalent in difficulty. Obviously, these are not features of real test data; however, the purpose was to demonstrate a result in principle.

In the second step, tetrachoric correlations were computed for the reference group, and the dimensionality of the simulated data was assessed. In the third step, the dimensional structure from the previous step was used to estimate latent means and covariances for the reference ($N = 2{,}000$) and focal groups ($N = 2{,}000$).

Data Preparation

There are several limitations to the factor analysis of dichotomous item responses. First, factor analysis is usually performed on the tetrachoric correlations because it is well-known that the use of phi coefficients confounds factor coefficients with item difficulties (Bock, Gibbons, & Muraki, 1988; Carroll, 1945). With real data, tetrachorics may still be confounded with item difficulty. This problem was avoided in the present study by simulating item responses for items of moderate difficulty without guessing. Tetrachorics can either be computed by a specialized program, or by PRELIS. For this study, software was written to obtain conditional maximum likelihood estimates of the tetrachorics.

When using one group, means, variances, and correlations of the underlying item variables are not simultaneously identified. However, with two groups, it was possible to apply a simplifying assumption that led to the identification of the second group's means relative to the first. It was assumed that the underlying variables are normally distributed with identical SDs of 1.0 across items and groups. (See Torgerson, 1958 for a discussion of this assumption.)

Let the quantity d be defined as

$$d = \Phi^{-1}(q_R) \quad , \tag{46}$$

where q_R is the proportion of examinees in the reference group answering the item incorrectly. The point d dichotomizes the underlying y^* metric of the reference group (see Figure 1), where in the two-dimensional model given by Equation 2 $y^* = a_1\theta_1 + a_2\theta_2 + e$. Note that if $y > 0$, this implies that $y^* > d$. Because the y^* metric is unobserved, it is common to assume that it is normal, with a mean of 0 and a variance of 1. The reference group's means and SDs on y^* are then set at 0 and 1, respectively, and d is defined in terms of this metric. Because d is invariant and y^* assumed normal, it can then be shown that for the second group's mean on an item

$$(d - \mu)/\sigma = d - \mu = \Phi^{-1}(q_F) \quad . \tag{47}$$

Because $\sigma = 1$ by assumption, this implies that

$$\mu = \Phi^{-1}(q_R) - \Phi^{-1}(q_F) \quad . \tag{48}$$

In summary, group correlations and relative means may be obtained under the assumption that variances are equal to 1 across groups and items. (It is possible that relative group variances are obtainable in principle, and approximate solutions for these quantities would be highly desirable.)

Dimensionality Assessment

This analysis proceeded in two steps. First, an exploratory factor analysis was performed on all 10 items using the method of maximum likelihood. This was done because LISREL is primarily used for confirmatory analyses, and in real applications a preliminary model must be developed to define the dimensionality of the solution. Such a solution can then be employed to constrain some coefficients in the LISREL $\mathbf{A}$ matrix. The eigenvalues extracted initially are given in Table 1. There is no evidence for a second factor based on a simple scree test. It appears that the correct factor structure is unlikely to be found in such conditions. However, due to the large sample size, the second factor was found

Table 1
Maximum Likelihood Factor Loadings, Communalities, and Eigenvalues from the Exploratory Factor Analysis

Variable	Factor 1	Factor 2	Commu-nality	Initial Eigenvalue
F1	.14	.08	.02	4.04
F2	.36	.13	.15	.97
F3	.53	.25	.35	.86
F4	.54	.28	.38	.77
F5	.52	.25	.33	.67
F6	.66	.10	.44	.64
F7	.60	.11	.38	.61
F8	.59	.18	.39	.58
F9	.74	−.10	.56	.52
F10	.83	−.27	.77	.34
Eigenvalue	3.43	.38		
Percent of Variance	34.3	3.8		
Cumulative Percent	34.3	38.1		

to be statistically significant, and the rotated two-factor solution (Table 2) bore a strong resemblance to the actual structure. The limitations of factor analysis as an exploratory technique were discussed in more detail by Mulaik (1972, Chapter 15).

If it is assumed that the zero loadings on Factor 2 are known, and that the abilities are orthogonal, the situation is much different. In this case, the diagonally weighted least squares (DWLS) algorithm yielded the confirmatory solution given in Table 2. As can be seen, the factor coefficients were estimated with a high degree of accuracy. The latter estimated factor structure was then input to LISREL (to model the effect of estimation errors), and this common model was fit to both the reference and focal groups in a multigroup analysis with structured means. At this step, latent means and covariance structures were estimated.

The actual reference and orthogonal composites can be obtained through standard matrix operations using Equations 19 and 23. To do this, the actual pattern matrix in Table 2 is taken as A, and

Table 2
Rotated Pattern and Structure Loadings From the Exploratory Analysis, and Model Specified and Estimated (Est) Loadings For the Confirmatory Solution (Factor Correlation = .746)

Pattern Matrix		Structure Matrix		Confirmatory Solution			
				Factor 1		Factor 2	
Factor 1	Factor 2	Factor 1	Factor 2	Model	Est	Model	Est
.18	−.02	.16	.11	.2	.17	0	0.00
.34	.05	.38	.31	.4	.40	0	0.00
.59	−0.00	.59	.44	.6	.60	0	0.00
.65	−.04	.61	.43	.6	.61	0	0.00
.59	−.01	.58	.42	.6	.58	0	0.00
.44	.26	.64	.60	.6	.64	.2	.18
.43	.22	.59	.54	.6	.60	.2	.16
.53	.11	.62	.51	.6	.62	.2	.10
.17	.61	.63	.74	.6	.63	.4	.42
−.02	.89	.64	.87	.6	.64	.6	.60

the factor correlation matrix is taken as $\mathbf{I}$. (Both of these quantities are standard output in factor analysis programs.) In the current analysis, the reference composite υ_1 and the orthogonal component υ_2 are given by $\upsilon_1 = .94\theta_1 + .33\theta_2$ and $\upsilon_2 = -.33\theta_1 + .94\theta_2$. Although υ_1 has positive weights for the two latent abilities, υ_2 has a positive weight on θ_2 and a negative weight on θ_1. The composite variable weights for the DWLS estimates of $\mathbf{A}$ were highly similar to those reported above.

A Structural Model

The structural model for the simulated item responses was primarily a measurement model for the item probit scores that related the 10 items to the latent abilities. LISREL terminology is used here in order to facilitate applications. Accordingly, the 10×1 vector of probit scores was relabeled $\mathbf{X}$, and the usual symbol for the vector of latent variables $\boldsymbol{\Theta}$ was changed to $\boldsymbol{\Xi}$, and elements of $\boldsymbol{\Xi}$ were labeled ξ. The measurement model for probit scores is then given by

$$\mathbf{X} = \tau_X + \Lambda_X \boldsymbol{\Xi} + \delta \quad . \tag{49}$$

This represents the regression of the xs on the factors ξ. Accordingly, the 10×1 vector τ_X contains the $-d$s where d is defined above in Equation 46, and the 10×2 matrix Λ_X contains the item discriminations, (i.e., the slope coefficients). The 2×1 $\boldsymbol{\Xi}$ vector contains the two latent abilities θ_1 and θ_2 relabeled ξ_1 and ξ_2.

In general, the parameters of the measurement model are considered to be invariant across groups. The multidimensional perspective of DIF is that differences in item responses are due to group differences in the means (which are labeled as the 2×1 vector $\boldsymbol{\kappa}$) and 2×2 covariance matrices (labeled $\boldsymbol{\Phi}$) of the ξ variables. (Φ is not being used here to denote the cumulative normal distribution function.) Consequently, the objective is to estimate $\boldsymbol{\kappa}$ and $\boldsymbol{\Phi}$. Fortunately, once Equation 49 is specified, these quantities are automatically computed and printed.

The LISREL commands for this analysis were written with the following specifications. First, $\boldsymbol{\tau}$ (TX), $\boldsymbol{\Lambda}$ (LX), and group correlation matrices (KM) were read and fixed. The $\boldsymbol{\Lambda}$ matrix of estimated discrimination parameters was obtained from the reference group's LISREL dimensionality assessment (see Table 2). Group means (ME), standard deviations (SDs), and item difficulties $\boldsymbol{\tau}$ were also read. The standardization constraint (ST) on $\boldsymbol{\Phi}$ in the reference group and fixed constraint (FI) on $\boldsymbol{\tau}$ and $\boldsymbol{\Lambda}$ effectively determined the scale of the ξ. All parameters were specified as invariant across groups except for $\boldsymbol{\kappa}$ and $\boldsymbol{\Phi}$. The former was free (FR) in both groups; its scale was determined by group mean vectors $\boldsymbol{\tau}$ and $\boldsymbol{\Lambda}$. The latter was specified as a correlation matrix in the reference group (ST) and a free symmetric matrix (FR,SY) in the focal group.

Results

The main results consist of three quantities: $\boldsymbol{\Phi}_F$, $\boldsymbol{\kappa}_R$, and $\boldsymbol{\kappa}_F$ (because $\boldsymbol{\Phi}_R = \mathbf{I}$). The estimates obtained were:

$$\boldsymbol{\Phi}_F = \begin{pmatrix} .37 & -.01 \\ -.01 & 2.87 \end{pmatrix} \quad , \tag{50}$$

$$\boldsymbol{\kappa}_R = (0.00 \quad 0.00) \quad , \tag{51}$$

and

$$\boldsymbol{\kappa}_F = (.08 \quad -.85) \quad . \tag{52}$$

These results indicate that the reference and focal group means on the dominant and secondary abilities

were fairly close to the actual parameters. Likewise, the focal group SDs for ξ_1 and ξ_2 (.61 and 1.69) were close to the actual values (.5 and 1.5). The implication for DIF is that if ξ_1 were used as a matching variable, then focal group examinees would score lower than expected on items requiring more of ξ_2 (and higher than expected on items requiring less of ξ_2).

This preliminary result, which describes differential test functioning, may have interesting applications to cultural studies as well as psychometric studies. The LISREL analysis can also, in principle, be extended to more than two demographic groups, and other explanatory variables can be included in Equation 49 (Muthén, 1988).

Discussion

There are several limitations to this analysis. First, as mentioned above, the data were generated with simplifying assumptions. However, the assumption of equal variances for the *x*s was violated by the data generation model, which is necessarily the case if the factor coefficients are invariant but the ability covariance matrix is not. The procedure seems to be somewhat robust to this assumption. Second, this approach was ad hoc in the sense that different methods of numerical estimation were employed at different stages. For example, the initial analysis was done with DWLS, but the multigroup analysis was done with unweighted least squares. Given these limitations, the above results should be interpreted with caution.

The most serious drawback to this approach is probably the estimation of tetrachoric correlations in the presence of differential item difficulty and guessing. A partial solution to guessing was given (Carroll, 1945) by use of a correction procedure. As an alternative to full-information factor models, improved methods for estimating tetrachoric correlations would lend the power of structural equations analyses to the field of DIF.

With multidimensional item responses, it seems both useful and necessary to analyze the data with an appropriate multidimensional item analysis. As applied to DIF analysis, the item parameters (as well as other measurement coefficients) for two groups would be constrained to be equal, and the parameters of the latent distributions could be estimated. Then a more satisfying description of the secondary abilities that are associated with the DIF would be gained. Although the utility of latent ability models is limited by the rotational indeterminacy of the factor structure, some a priori knowledge of the true factor structure may lead to a satisfactory result.

The current approach makes DIF a property of the examinee, not the item. The secondary traits are likely to result from the educational histories of individuals. This has important implications for studying DIF—a tool is provided for incorporating background variables directly into the model. This seems more useful than the methods of (1) using a multivariate matching strategy that includes both background variables and total score, or (2) observing whether DIF indices are reduced relative to indices computed by using total score as a single matching variable—especially because empirical research employing the multivariate matching design has yielded inconsistent results.

One way to conceptualize why some DIF is difficult to explain is facilitated by the standard factor analytic approach. Suppose a test is composed of a primary factor (say θ_1) and other common factors (say θ_2), and unique factors (say ς). Differences in item parameters per se are practically indistinguishable from differences in $G(\varsigma|\theta_1)$. Furthermore, such differences may be extremely difficult to explain for the same reasons that weak secondary factors are often difficult to interpret in factor analysis. It is more likely that DIF, when it can be explained, is due to differences in $G(\theta_2|\theta_1)$. Nevertheless, DIF arising from unique factors undoubtedly exists (and may be the most common variety). For example, DIF caused by a word with a different frequency of usage in two groups may be difficult to interpret if only one item on a test contains this word. False cognates are more

interesting examples of unique linguistic differences between English and Spanish-English speaking examinees (Schmitt, 1988). The notion and role of unique factors on long tests has been given an interesting treatment by Shealy and Stout (in press) in their discussion of "essential dimensionality," and by Holland (1990) in his discussion of the "Dutch identity."

A major finding of this study is that if multidimensionality holds, then unidimensional DIF methods are likely to confound DIF with impact. Because the estimated ability—that is, the reference composite—is partly a function of ability on secondary (and possibly irrelevant) abilities, true DIF in terms of the target ability (as indexed by a_2/a_1) may be obscured. Equation 39 describes the conditions under which true DIF may be present but undetected. This may serve to explain why Oshima (1989) concluded that mean secondary trait differences had only small effects on the invariance of IRFs with tests comprised of at least 10% to 20% multidimensional items, but had larger effects on tests comprised of only 5% of such items. Note that the number of multidimensional items as well as the size of the loadings on secondary dimensions influences the reference composite. More specifically, the reference composite varies directly with $\mathbf{A}$ if $\mathbf{\Omega}_\Theta$ is held fixed. As the reference weights on secondary factors increase, more of the group difference in secondary ability means may be mapped into υ_1, and not into the DIF statistic. This effect is stronger for highly correlated traits, although the effects of many unique factors may tend to cancel (Shealy & Stout, in press). It is important to note that although θ_2 can be called a secondary ability, it may be intentionally measured. Consequently, items showing DIF due to such a factor would not be labeled as biased, although bias could result later by use of unidimensional scoring or equating methods.

Some researchers have argued that impact—the difference between the average test scores of two groups—is not an appropriate index for evaluating item bias (Anrig, 1988; Holland & Thayer, 1988; Linn & Drasgow, 1987). The current technology, however, which includes the Mantel-Haenszel method, does not automatically distinguish DIF from impact. There are dangers to conditioning DIF statistics on ability estimated with an incorrect model. The use of such statistics to aid the identification of test items that are biased must be appropriately qualified.

References

Ackerman, T. A. (1988, April). *An explanation of differential item functioning from a multidimensional perspective.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13,* 113–127.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 97–116). Baltimore MD: Johns Hopkins University Press.

Anrig, G. R. (1988). ETS replies to Golden Rule on "Golden Rule." *Educational Measurement: Issues and Practices, 7,* 1, 20–21.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading MA: Addison-Wesley.

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Camilli, G. (In press). The case against item bias detection techniques based on internal criteria. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice.* Hillsdale NJ: Erlbaum.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika, 10,* 1–19.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

Cronbach, L. J., Ragosa, D. R., Floden, R. E., & Price, G. G. (1977). *Analysis of covariance in nonrandomized*

experiments: Parameters affecting bias. Stanford CA: Stanford University, Stanford Evaluation Consortium, School of Education.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 22,* 249–262.

Holland, P. W. (1990). The Dutch identity: A new tool for the study of item response models. *Psychometrika, 55,* 5–18.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–146). Englewood Cliffs NJ: Erlbaum.

Hunter, J. E. (1975, December). *A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement items.* Paper presented at the National Institute of Education Conference on Test Bias, Annapolis MD.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7 user's reference guide.* Mooresville IN: Scientific Software.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 263–275). New York: Plenum Press.

Linn, R. L., & Drasgow, F. (1987). Implication of the Golden Rule for test construction. *Educational Measurement: Issues and Practices, 6,* 13–17.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1980). An investigation of 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.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Mulaik, S. A. (1972). *The foundations of factor analysis.* New York: McGraw-Hill.

Muthén, B. (1987). *LISCOMP: Analysis of linear structural relations with a comprehensive measurement model.* Mooresville IN: Scientific Software.

Muthén, B. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 213–238). Englewood Cliffs NJ: Erlbaum.

Muthén, B., & Christofferson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46,* 407–419.

Muthén, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10,* 133–142.

Oshima, T. (1989). *The effect of multidimensionality on item bias detection based on item response theory.* Unpublished doctoral dissertation, University of Florida, Gainesville.

Ragosa, D. R. (1980). Comparing nonparallel regression lines. *Psychological Bulletin, 88,* 307–321.

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

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25,* 193–302.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39,* 111–121.

Schmitt, A. P. (1988). Language and cultural characteristics that explain item functioning for Hispanic examinees on the Scholastic Aptitude Test. *Journal of Educational Measurement, 25,* 1–13.

Shealy, R., & Stout, W. (In press). An item response theory model for test bias. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning: Theory and practice.* Hillsdale NJ: Erlbaum.

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

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

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361–370.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82–98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

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

Thissen, D., Steinberg, L., & Wainer, H. (In press). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice.* Hillsdale NJ: Erlbaum.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273–286.

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: McGraw-Hill.

Wang, M. (1985). *Fitting a unidimensional model to*

multidimensional item response data: The effects of latent space misspecification on the application of IRT. Unpublished manuscript, University of Iowa.

Winer, B. J. (1971). *Statistical principles in experimental design.* New York: McGraw-Hill.

Author's Address

Send requests for reprints or further information to Gregory Camilli, Rutgers University, Graduate School of Education, 10 Seminary Place, New Brunswick NJ 08903, U.S.A.