Methodology Review: Statistical Approaches for Assessing Measurement Bias

Roger E. Millsap, Baruch College, City University of New York

Howard T. Everson, The College Board

Statistical methods developed over the last decade for detecting measurement bias in psychological and educational tests are reviewed. Earlier methods for assessing measurement bias generally have been replaced by more sophisticated statistical techniques, such as the Mantel-Haenszel procedure, the standardization approach, logistic regression models, and item response theory approaches. The review employs a conceptual framework that distinguishes methods of detecting measurement bias based on either *observed* or *unobserved* conditional invariance models. Although progress has been made in the development of statistical methods for detecting measurement bias, issues related to the choice of matching variable, the nonuniform nature of measurement bias, the suitability of current approaches for new and emerging performance assessment methods, and insights into the causes of measurement bias remain elusive. Clearly, psychometric solutions to the problems of measurement bias will further understanding of the more central issue of construct validity. The continuing development of statistical methods for detecting and understanding the causes of measurement bias will continue to be an important scientific challenge. *Index terms: bias detection, differential item functioning, item bias, measurement bias, test bias.*

Given the widespread use of standardized tests in education, the issue of test or measurement bias is central to measurement theoreticians, practitioners, and educational policy makers. But what kind of evidence leads to the conclusion that a psychological or educational test is biased, either for or against a particular examinee group? In this review, "bias" refers to a systematic inaccuracy of measurement, a concept defined more explicitly below. Obviously, a complete answer to the measurement bias question requires thorough analyses of empirical data. Statistical methods to evaluate data in the investigation of measurement bias are reviewed. An extensive literature has evolved on this topic over the last 30 years, and more particularly, a number of important developments have come about in the last decade. This paper provides an up-to-date review of state-of-the-art methodological developments appearing in the literature since the publication of the *Handbook of Methods for Detecting Test Bias* (Berk, 1982).

The primary interest in this review is in group-level measurement bias. Methods for bias detection in both continuous and ordered-categorical (including dichotomous) measures are reviewed. Also, methods for use in testlets and multiple measures, as well as item-level methods are discussed. Methods relying on expert judgment are not discussed. Methods for detecting predictive bias typically found in personnel or educational selection contexts also are not reviewed. Studies of predictive bias may have implications for measurement bias, but the two forms of bias need not be related (Drasgow, 1982; Millsap & Meredith, 1992). Also, no discussion of the developments in the area of appropriateness or person-fit indexes is provided.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 17, No. 4, December 1993, pp. 297-334

0146-6216/93/040297-38$3.15

The phrase "measurement bias" has been adopted in this review primarily for grammatical convenience, while recognizing the additional connotations of the term "bias." Many researchers prefer the neutral phrase "differential item functioning" (DIF) when the measurement is an item score. This phrase is less appropriate when the measure is not an item, but rather a testlet or an entire set of items (e.g., factor analytic studies). Because this review includes such applications, *measurement bias* has been selected as a general label. A formal definition of bias is given below.

This review employs an analytic framework, presented in more detail in the next section of the paper, that distinguishes methods based on either an observed conditional invariance (OCI) or an unobserved conditional invariance (UCI) model of measurement bias. The various methods are classified and reviewed using this classification. A discussion of the outstanding methodological problems in detecting measurement bias, and an overview of promising directions for future research also are provided.

A Conceptual Framework

A distinction can be made between impact and measurement bias or DIF (Dorans & Holland, 1993). *Impact* refers to group differences in measured performance on tests or items. Because individuals commonly differ on the attributes measured by tests, impact is ubiquitous. For example, males typically score higher on average than females on standardized tests of mathematics such as the Scholastic Aptitude Test mathematics section (Wilder & Powell, 1989). Measurement bias or DIF, on the other hand, refers to differences in the functioning of a test or item among groups that are matched on the attribute measured by the test or item (Dorans & Holland, 1993; Scheuneman, 1979). As Dorans & Holland (1993) assert, it is critical when assessing measurement bias that performance differences among matched groups be examined to avoid Simpson's (1951) paradox. In this paradox, the direction of item impact is inconsistent with the direction of group differences among matched individuals. For example, a math item may be more difficult for females overall, yet may be less difficult for females within a group of examinees who have been matched on ability.

A Formal Definition of Measurement Bias

The formal definition of measurement bias provided here is intentionally stated at a general level to include many forms of measurement bias. Denote the observed scores provided by a measuring instrument as a random variable Y, which may be univariate (e.g., scores on a single item) or multivariate (e.g., scores on a set of items). Ordinarily, Y is discretely measured, but the number of possible values may be large. Examinees are divided into two or more populations on the basis of variables denoted as V, which can be multivariate. The variables in V are usually demographic information, such as ethnicity, gender, or age. It is assumed that these are known and measured without error. Finally, define W as a latent or unobserved variable for which Y is the intended observed indicator: Y is viewed as a measure of W. The latent variable W may be univariate or multivariate, depending on the nature of Y. Although latent variables in the measurement context often are viewed as continuous, W will be considered to be discrete for notational convenience in the following definition. UCI holds for Y in relation to W and V if

$$P(Y \mid W = w, V = v) = P(Y \mid W = w) \tag{1}$$

for all values of W and V. Here, $P(Y \mid W = w)$ is the conditional probability function for Y given that W assumes the value w. If Equation 1 holds for Y, the relationship between Y and the latent variable W is independent of group membership. Among individuals with common values on W, the distribution of Y is the same across populations defined by V. Y is unbiased as a measure of W with respect

to V if UCI in Equation 1 holds. Conversely, measurement bias in Y in relation to W occurs if UCI is violated.

This definition of bias flows from the distinction between impact and DIF and, as such, adequately represents many conceptions of bias found in the literature. If Y is taken to be a dichotomous item score variable, and W is a unidimensional latent trait, UCI in Equation 1 corresponds to Lord's (1980) definition of lack of item bias. In this situation, UCI implies identical item response functions (IRFs) among populations. Lord's (1980) definition forms the basis for methods of DIF detection in applications of item response theory (IRT; Thissen, Steinberg, & Wainer, 1988). Mellenbergh (1989) presented a definition of an "unbiased" item that is nearly identical to Equation 1, as did Kok (1988). If $\mathbf{Y}$ is a vector of observed continuous measures whose regression on $\mathbf{W}$ corresponds to the linear factor analytic model, with $\mathbf{W}$ a vector of factor score variables, then UCI in Equation 1 implies an invariant factor structure for $\mathbf{Y}$ (Meredith, 1990). Hence, UCI in Equation 1 can encompass the idea of factorial invariance as well.

The idea of conditioning on W in defining bias is important for distinguishing measurement bias from ordinary group differences, or impact. For example, it may be true that groups differ in score distributions on Y, or that

$$P(Y \mid V = v) \neq P(Y) \ . \tag{2}$$

There is a general consensus in the literature that Equation 2 is not sufficient to establish bias as defined above (Ackerman, 1992; Drasgow, 1987; Holland & Thayer, 1988; Lord, 1980). To the extent that examinee performance depends on W, and that groups differ on W, Equation 2 may indicate bias even if no bias exists. Empirical bias investigations usually proceed on the assumption that group differences on W are possible or likely. The validity of this assumption will not be debated here, and it will be assumed that the distribution of W differs across groups.

Bias and Dimensionality

An important issue in any bias investigation is the dimensionality of W, or the number of latent dimensions believed to underlie Y. If the bias investigation is to be meaningful, W in Equation 1 must be limited to include only those dimensions for which Y is intended to be an indicator. There may exist additional latent variables that influence Y in unanticipated ways. For example, suppose that Y is a reading comprehension item score that is intended as a measure of a unidimensional latent variable W_1 (e.g., reading ability). Suppose, however, that the content of the reading selection for Y is unusual and favors examinees with prior familiarity with this content. Hence, there is a second latent variable W_2 (i.e., prior knowledge). Given the intended purpose for Y, W_2 may be considered a nuisance variable (Ackerman, 1992; Kok, 1988; Shealy & Stout, 1993). The important issue here is that UCI in Equation 1 may hold for $W = (W_1, W_2)$, but not for $W = (W_1)$.

This example illustrates that some restrictions must be placed on W in Equation 1 if the bias investigation is to proceed. Thus, issues of construct validity are inextricably bound to issues of measurement bias (Ackerman, 1992).

Bias and Measurement Models

Some bias detection methods attempt to test UCI directly by first proposing a measurement model relating Y and W. Bias then is investigated by evaluating whether features of the model remain invariant over populations defined by V. In investigations of item bias, bias detection methods based on IRT are examples of this. IRT models have been proposed for ordered-categorical response formats as well as dichotomously-scored items. For continuous measurements, factor-analytic models are

commonly used. Methods for investigating factorial invariance are also examples of this. Collectively, these types of bias detection methods are denoted UCI methods.

Other detection methods proceed without formal specification of a measurement model relating Y and W. Instead, an observable random variable Z (possibly multivariate) is found that may serve as a stratifying variable for use in examining bias. Here, Z is intended as a proxy for W. For example, in studies of item bias, Z may be taken as the total test score. More generally, Z could include information external to the test under consideration. These methods investigate a form of invariance that parallels UCI in Equation 2, where

$$P(Y|Z = z, V = v) = P(Y|Z = z) . \quad (3)$$

If Equation 3 holds, the distributions of scores on Y among examinees with common values on Z are independent of group membership. If $Z = W$, Equation 3 is identical to UCI in Equation 1, and hence Y is unbiased. These methods use empirical data on Y, Z, and V to assess Equation 3 in hopes of inferring something about UCI in Equation 1. In the context of item bias detection, examples of these methods include the traditional χ^2 methods (Ironson, 1982; Marascuilo & Slaughter, 1981; Scheuneman, 1979; Shepard, Camilli, & Averill, 1981), the Mantel-Haenszel (MH) χ^2 method (Holland & Thayer, 1988; Mantel & Haenszel, 1959), standardization approaches (Dorans & Kulik, 1986), and logistic regression methods (Swaminathan & Rogers, 1990a). The form of invariance in Equation 3 will be referred to as OCI, and these bias detection models will be referred to as OCI methods.

As noted earlier, the OCI/UCI distinction provides a useful basis for organizing this review of the broad array of methods for detecting measurement bias that have been developed over the last decade. The next section includes a review of OCI methods, including loglinear models, the MH statistic, and methods based on logistic regression. This is followed by a review of UCI methods. Both sections discuss bias detection in dichotomous and polytomous measurements.

OCI Methods

Statistical methods for detecting unexpected item bias or DIF have been under development for nearly three decades (Angoff & Ford, 1973; Berk, 1982; Cardall & Coffman, 1964). Methods such as the traditional χ^2 approaches (e.g., Ironson, 1982; Marascuilo & Slaughter, 1981; Scheuneman, 1979; Shepard et al., 1981), loglinear models (Mellenbergh, 1982), the MH χ^2 (Holland & Thayer, 1988; Mantel & Haenszel, 1959), the standardization approach (Dorans, 1989; Dorans & Kulick, 1986), logistic regression techniques (Swaminathan & Rogers, 1990a), and, more recently, logistic discriminant function analysis (Miller, Spray, & Wilson, 1992) are classified here as OCI approaches. In this section, however, only those OCI methods that have emerged in the last decade are reviewed. A brief description of each method is presented, including its statistical assumptions and/or rationale. This is followed, where appropriate, by a discussion of the research highlighting the relative strengths and weaknesses of each method. Detailed reviews of the more traditional χ^2 methods are widely available elsewhere (e.g., Ironson, 1982; Ironson & Subkoviak, 1979; Osterlind, 1983; Rudner, Getson, & Knight, 1980a; Scheuneman & Bleistein, 1989; Shepard et al., 1981; Shepard, Camilli, & Williams, 1985).

Loglinear Models

Building on traditional χ^2 methods for studying item bias (Camilli, 1979; Scheuneman, 1979), a number of researchers have extended them to fit the more general theory of loglinear and logit models for contingency tables (e.g., Alderman & Holland, 1981; Kok, Mellenbergh, & Van der Flier, 1985;

Loyd, 1984; Mellenbergh, 1982; Van der Flier, Mellenbergh, Ader, & Wijn, 1984). Under this approach, item responses (both correct and incorrect) are classified in a three-way contingency table, which includes score level × group × item response (test scores are divided into score levels). Because dichotomous item response variables are used, the loglinear models are transformed into logit models; the logit is defined as the natural logarithm of the ratio of the number of correct to the number of incorrect item responses. Thus, model fitting and parameter estimation are implemented using corresponding loglinear models (Bishop, Fienberg, & Holland, 1975).

Following Bishop et al. (1975), the saturated loglinear model for the natural logarithm of the ith item response category ($i = 1, \ldots, m - 1$) in the kth score level ($k = 1, 2, \ldots, s$) and the jth group ($j = 1, 2, \ldots, g$) is:

$$\ln F_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)} . \quad (4)$$

The corresponding saturated logit model, where correct ($i = 1$) and incorrect ($i = 2$), is

$$\ln(F_{1jk}/F_{0jk}) = \alpha + \beta_k + \delta_j + (\delta\beta)_{jk} , \quad (5)$$

where

ln is the natural logarithm,
F_{1jk} is the expected frequency of correct item responses in the jth group,
F_{0jk} is the expected frequency of incorrect item responses in the jth group,
α is the overall item difficulty effect parameter,
β_k is the main score level effect,
δ_j is the main group effect, and
$(\delta\beta)_{jk}$ is the score level × group interaction effect.

A model for an unbiased item is created as a special case of the saturated model in Equation 5 by eliminating the terms δ_j and $(\delta\beta)_{jk}$. If this model fits the data for an item, the item is declared to be unbiased. Ordinarily, this "unbiased" model will be tested first.

A less restrictive model is created by eliminating only the $(\delta\beta)_{jk}$ term from the saturated model in Equation 5. An item that fits this model (and does not fit the unbiased model) shows "uniform" bias (Mellenbergh, 1982). Under uniform bias, the differences between groups in item performance are consistent across score levels. For example, one group may consistently correctly answer the item more frequently than the other group across all score levels. The interaction term $(\delta\beta)_{jk}$ represents "nonuniform" bias in which the group differences in performance vary over score levels.

Model testing is done using sample frequencies to estimate the expected frequencies of the contingency table and calculating the likelihood ratio G^2 statistic (Fienberg, 1980), which is distributed asymptotically as χ^2 (Kok et al., 1985; Mellenbergh, 1982; Van der Flier et al., 1984). The G^2 statistic is

$$G^2 = 2 \sum_{i=1}^{m-1} \sum_{j=1}^{g} \sum_{k=1}^{s} \ln f_{ijk}(f_{ijk}/\hat{F}_{ijk}) , \quad (6)$$

where $\hat{F}_{ijk}$ is the expected frequency estimate under the model tested, and f_{ijk} is the sample frequency. The degrees of freedom (df) will depend on the model being tested. The unbiased model is tested with $s(g - 1)$ df. Hierarchical tests between models that are logically nested can be conducted using differences in the respective G^2 statistics. These tests are useful for distinguishing uniform and nonuniform bias.

Loglinear models offer several advantages for bias detection. The models are flexible and can be extended to polytomous items, to multiple examinee groups, or to simultaneous bias detection in

several items (Agresti, 1990; Kelderman, 1989). The distinction in these models between uniform and nonuniform bias is important for understanding the nature of any bias that is present.

One difficulty in the use of these models may arise when a substantial portion of the items are biased, resulting in bias in the total score used to match examinees. Van der Flier et al. (1984) proposed that biased items be removed from the total score by iteratively applying the loglinear detection procedure. They demonstrated that this iterative approach leads to improved bias detection in simulated data.

A more fundamental problem with the loglinear approach is the assumption that the loglinear model adequately represents the data for purposes of bias detection. When the item responses fit the Rasch model, the response probabilities can be represented by a loglinear model (Cressie & Holland, 1983); however, data generated by more complex IRT models [e.g., the two-parameter logistic model (2PLM)] cannot. One aspect of this problem is the adequacy of the total score as a substitute for the unobserved latent trait. Under the Rasch model, the total score is a sufficient statistic for the latent trait, and is an adequate substitute (Lord & Novick, 1968). Sufficiency breaks down when the underlying response model is more complex. As a result, problems would be expected to be encountered in using loglinear models for bias detection in data generated by multiparameter IRT models. The robustness of the loglinear approach in such cases has not been studied thoroughly. Kelderman (1989) presented encouraging results in a small simulation study. The studies by Van der Flier et al. (1984) and Kok et al. (1985) did not properly address the issue because the examinee groups used did not differ in distributions of the latent trait (θ). More thorough studies of the robustness of the loglinear approach are needed.

Mantel-Haenszel Statistic

The MH statistic (Mantel & Haenszel, 1959), a contingency table method derived initially for use in biomedical research, was extended by Holland (1985) and subsequently by Holland & Thayer (1988) for use in detecting DIF. It has rapidly become one of the more widely used methods for detecting item-level measurement bias. Like the loglinear models, the MH statistic is a natural extension of the traditional χ^2 approaches described by Scheuneman (1979) and Marascuilo & Slaughter (1981). Others (e.g., Holland & Thayer, 1988; Thissen et al., 1988) have noted the relationship between the MH procedure and IRT-based methods for detecting measurement bias at the item level.

The MH procedure compares the performance of two groups of examinees—the reference and focal groups—on all the items in a given test, one item at a time. The group designated as the *focal group* is the group that is believed to be disadvantaged by the presence of DIF in the test. The group designated as the *reference group* serves as a comparison group for the purpose of DIF detection. Like the loglinear methods discussed above, the performance of comparable members of both groups are contrasted. Typically the total test score is the matching variable for establishing comparability between the groups.

The data for the MH procedure are contained in a $s \times 2 \times 2$ contingency table (where s designates the number of test score levels). Thus, at each score level k, individual item data from two groups of examinees can be arranged as a 2×2 table (see Table 1). From the $s \times 2 \times 2$ table for any given item, the following statistics are computed (see Holland & Thayer, 1988):

$$\text{MH } \chi^2 = \frac{\left[\left|\sum_{k=1}^{s} f_{1rk} - \sum_{k=1}^{s} \mathrm{E}(f_{1rk})\right| - .5\right]^2}{\sum_{k=1}^{s} \mathrm{Var}(f_{1rk})}, \tag{7}$$

where

$$E(f_{1rk}) = \frac{n_{1k} n_{rk}}{n_k}, \tag{8}$$

$$\text{Var}(f_{1rk}) = \frac{n_{1k} n_{0k} n_{rk} n_{fk}}{(n_k)^2 (n_k - 1)}, \tag{9}$$

$$\alpha_{MH} = \frac{\sum_{k=1}^{s} f_{1rk} f_{0fk} / n_k}{\sum_{k=1}^{s} f_{0rk} f_{1fk} / n_k}, \tag{10}$$

and

$$\Delta_{MH} = -\frac{4}{1.7} \ln(\alpha_{MH}) = -2.35 \ln(\alpha_{MH}), \tag{11}$$

where α_{MH} is the common odds ratio across the $s \times 2 \times 2$ tables for a particular test item. α_{MH} ranges from 0 to ∞, with values of 1.0 signifying no DIF. Values less than 1.0 indicate that the item is less difficult for examinees in the focal group, controlling for total test score. α_{MH} values greater than 1.0, on the other hand, indicate that the item is less difficult for examinees in the reference group.

Table 1
Mantel Haenszel $s \times 2 \times 2$ Contingency Table

	Score on Studied Item		
Group	1	0	Total
Reference	f_{1rk}	f_{0rk}	n_{rk}
Focal	f_{1fk}	f_{0fk}	n_{fk}
Total	n_{1k}	n_{0k}	n_k

The null hypothesis tested in the MH procedure is that α_{MH} across the s tables is equal to 1.0. This null hypothesis corresponds to Equation 3 with Z defined as the total test score and Y defined as the studied item score. The alternative hypothesis tested is that α_{MH} is not equal to 1.0. Note that this alternative hypothesis includes only a subset of the possible situations that depart from the null hypothesis. Only situations involving homogeneous α_{MH} across the s tables are considered, as discussed by Holland & Thayer (1988). The MH χ^2 test is the uniformly most powerful test against alternatives within this class.

To perform the test, the MH χ^2 value is referred to the χ^2 distribution with 1 *df*. As an index of DIF, the odds ratio estimator in Equation 10 is consistent and is efficient over a wide range of true odds ratio values (Holland & Thayer, 1988). The transformed estimator in Equation 11 expresses DIF in the metric of the Educational Testing Service's delta scale; Dorans & Holland (1993) labeled this estimator MH D-DIF.

Largely because the MH statistic is conceptually simple, relatively easy to use, and provides a χ^2 test of significance, it has become a widely used method for detecting measurement bias at the item level. Moreover, the MH procedure addresses the general problem with χ^2 methods of only providing tests of the null hypothesis and lacking a parameter estimate of the amount of DIF present in the item (Holland & Thayer, 1988).

As a result of this widespread use, the MH procedure also has been the focus of much research recently. The MH procedure, for example, has been compared with the standardization method (Dorans & Kulick, 1983, 1986; Wright, 1987), described below, and has yielded similar results. Others have

examined many factors thought to affect the stability of the statistic, including grouping of scores on the matching variable (Donoghue & Allen, 1991), the amount and type of DIF (Uttaro, 1992), inclusion of the studied item in the matching variable score (Donoghue, Holland, & Thayer, 1993), sample ability differences (Harvey, 1990), and sample size (Mazor, Clauser, & Hambleton, 1991; Ryan, 1991).

The MH procedure does have several disadvantages. First, the procedure is not designed to detect nonuniform bias. The MH procedure sacrifices some sensitivity to achieve greater power for detecting uniform bias (Holland & Thayer, 1988). Several studies have shown that the MH procedure has relatively low power for detecting nonuniform bias (Swaminathan & Rogers, 1990a, 1990b; Uttaro, 1992). This problem is of concern if the item responses are generated by non-Rasch IRT models.

The second problem concerns the adequacy of the total score as a substitute for the latent trait. Both theoretical studies (Meredith & Millsap, 1992; Millsap & Meredith, 1992; Zwick, 1990) and simulation studies (Uttaro, 1992) have shown that when the item responses are generated by complex IRT models, the MH procedure can falsely indicate DIF when no bias is present. This problem is more serious in short tests (less than 20 items). As test length increases, the total score becomes a better proxy for the univariate latent trait.

Several extensions of the MH procedure are available for polytomous item scores (Zwick, Donoghue, & Grima, 1993). One extension is based on the procedure given by Mantel (1963) that considers ordered categories. In this extension, index numbers are assigned to the ordered response categories, and examinee groups are compared on their mean responses, conditional on the total score group. A 1 df χ^2 statistic is used to test the null hypothesis of no DIF. The second extension is based on the generalized MH statistic (Mantel & Haenszel, 1959; Somes, 1986). This statistic considers examinee group differences in the entire response distribution across the m response categories within a given score group. A χ^2 statistic with $m - 1$ df is used to test the null hypothesis of no DIF. Zwick et al. (1993) used both of these procedures in simulated and real data and found that both procedures adhered to the nominal Type I error rate under no-DIF conditions in the simulations. They also found some differences between the two procedures in the form of DIF that was detected most readily. Both procedures require further study.

Standardization Method

Another DIF assessment technique which is highly related and, indeed, complementary to the MH procedure is the standardization approach developed by Dorans & Kulick (1983, 1986; Dorans & Holland, 1993). While the MH procedure is a statistically powerful technique for detecting measurement bias at the item level, the standardization method is a more easily understood procedure for describing and explaining the nature of the measurement bias.

The initial step in this method is to define the empirical item-test regressions for both the focal and reference groups. At score level Z, these regressions take the form $E_f(Y|Z)$ and $E_r(Y|Z)$, where Y is the item score. Thus, the definition of measurement bias in the standardization method implies

$$E_f(Y|Z) \neq E_r(Y|Z) \ . \tag{12}$$

The primary DIF statistic computed using the standardization method is commonly referred to as the standardized p difference (STD P-DIF), and is computed as follows:

$$\text{STD P-DIF} = \sum_{k=1}^{s} [W_k(P_{fk} - P_{rk})] \Big/ \sum_{k=1}^{s} (W_k) \ , \tag{13}$$

where $[W_k/\Sigma(W_k)]$ is the weighting factor at the kth score level that weights the differences in the proportions correct between the focal group ($P_{fk} = f_{1fk}/n_{fk}$) and the reference group ($P_{rk} = f_{1rk}/n_{rk}$).

These weighted differences are summed across the score levels to yield STD P-DIF. The STD P-DIF index ranges from −1.0 to +1.0, with positive values indicating that the item favors the focal group and negative values indicating the item favors the reference group. STD P-DIF values between −.05 and +.05 are considered negligible; values between −.10 and −.05 and .05 and .10 will cause the item to be flagged for further inspection. The next step is to try to understand the causes for the DIF. According to Dorans, Schmitt, & Bleistein (1992), these cutoff points work well in practice.

The results of the standardization method are usually in close agreement with those of the MH procedure (Dorans & Holland, 1993). The close relationship between the two procedures suggests that the problems found in the MH procedure also will affect the standardization procedure. For example, the standardization procedure also uses the total score as a substitute for the unobserved latent trait, and hence should encounter problems in data generated by multiparameter IRT models.

Recently, the standardization method has been extended to all response options (Dorans et al., 1992). This new approach, referred to as *comprehensive DIF* or CDIF (Dorans et al., 1992), may prove useful for understanding the differential functioning of distractors, as well as for identifying differential speededness on omitted items near the end of a test. This ability to identify differential speededness may help clarify the cause of DIF in some items and, in turn, reduce the noise in the matching variable used in the MH procedures. The standardization approach holds promise for furthering understanding of why a test item functions differentially for some groups and not others.

Logistic Regression Models

Swaminathan & Rogers (1990a) introduced a DIF procedure based on the logistic regression model, which is sensitive to both uniform and nonuniform DIF. For example, let Z be the observed proxy variable (usually the total test score) that is used to match individuals for purposes of bias detection. Let V be the indicator variable that identifies demographic group membership. In the full logistic regression model, the conditional probability that an examinee will correctly answer the test item, given Z and V, is

$$P(Y = 1 \mid Z, V) = \frac{\exp(\beta_0 + \beta_1 Z + \beta_2 V + \beta_3 ZV)}{[1 + \exp(\beta_0 + \beta_1 Z + \beta_2 V + \beta_3 ZV)]}, \tag{14}$$

where Y is the item score variable. Equation 14 can be rewritten as a linear model in the logit metric as

$$\log\left[\frac{P}{1 - P}\right] = \beta_0 + \beta_1 Z + \beta_2 V + \beta_3 ZV . \tag{15}$$

The regression parameters in Equation 15 can be estimated using maximum likelihood (Bock, 1975) and can be tested for significance. If the item is unbiased, only β_0 and β_1 should be nonzero. A model that includes β_0, β_1, and β_2 corresponds to an item that shows uniform bias. If the interaction parameter β_3 is nonzero, nonuniform bias is present. Swaminathan & Rogers (1990a) argue that "... the Mantel-Haenszel procedure can be thought of as being based on a logistic regression model where the ability variable is discrete and no interaction term between the group variable and ability is permitted" (p. 365).

The logistic regression procedure can be viewed as a reformulation of the loglinear approach for the case of dichotomous item scores with stratified sampling. Bishop et al. (1975) described the relationship between logit models and the general loglinear model. The tests of uniform and nonuniform bias described under Equation 15 above are equivalent to those given by the loglinear approach.

The logistic regression procedure is flexible because it can be extended to multiple examinee groups and to polytomous item scores (Agresti, 1990; Miller, Spray, & Wilson, 1992). Recent research

comparing the logistic regression procedure to the MH procedure (Swaminathan & Rogers, 1990a, 1990b) suggests that the logistic regression model is as powerful for detecting uniform DIF as the MH procedure and more powerful for detecting nonuniform DIF than the MH procedure. In their simulation studies, Swaminathan & Rogers (1990b) manipulated sample size (N = 250 vs. 500), test length (40, 60, or 80 items) and the nature of DIF (uniform or nonuniform). Under uniform DIF, both the logistic regression and the MH procedures performed equally well; however, the MH performed slightly better in the smaller sample. Only the logistic regression model, however, was able to consistently detect the nonuniform DIF. Finally, logistic regression procedures produce estimates of the regression coefficients which, when plotted, may be useful for locating DIF along the score scale (Miller et al., 1992).

Logistic regression procedures share the difficulties faced by loglinear methods. The use of the total score as a proxy for the latent trait will encounter problems if the item responses follow multiparameter IRT models. The simulations conducted by Swaminathan & Rogers (1990b) used relatively long tests, reducing the opportunity for problems to arise. In shorter tests, theory suggests that false indications of bias will be encountered in the no-DIF case.

The logistic regression procedure allows for the inclusion of curvilinear terms and other factors—such as examinee characteristics like test anxiety or instructional opportunity—that may be relevant factors for exploring possible causes of DIF. Logistic regression models also produce estimates of the regression coefficients which, when plotted, may be useful for determining where along the score scale the DIF is problematic (Miller et al., 1992). Thus, thoughtful use of this approach may further understanding of the theoretical nature of DIF.

Logistic Discriminant Function Analysis

The call for performance assessments relying on item formats other than the traditional correct/incorrect or dichotomously scored responses has generated interest in methods for detecting item bias when more than one response category is appropriate, commonly referred to as polytomous item responses. Miller et al. (1992) developed an extension of logistic regression procedures (Swaminathan & Rogers, 1990a) that holds promise for identifying DIF in polytomously scored items, an approach they termed logistic discriminant function analysis.

Thus, recasting the logistic regression model yields

$$P(Y|Z,V) = \frac{\exp(1-Y)(-\beta_0 - \beta_1 Z - \beta_2 V - \beta_3 ZV)}{[1 + \exp(-\beta_0 - \beta_1 Z - \beta_2 V - \beta_3 ZV)]}, \quad (16)$$

where the probability of observing each dichotomous response Y is modeled as a function of two predictor or explanatory variables, the observed test score Z and a group term V. Miller et al. (1992) argued that it is reasonable, employing a logistic form of the posterior probability used in discriminant analyses, to estimate $P(V|Z,Y)$ even though V is fixed and Y is random. Thus, the logistic discriminant function analysis for DIF detection with polytomously scored responses is written as

$$P(V|Z,Y) = \frac{\exp(1-V)(-\alpha_0 - \alpha_1 Z - \alpha_2 Y - \alpha_3 ZY)}{[1 + \exp(-\alpha_0 - \alpha_1 Z - \alpha_2 Y - \alpha_3 ZY)]}, \quad (17)$$

where the regression coefficients are denoted by α_i, i = 0, 1, 2, and 3 and, as in Equation 16, group membership is denoted by V = 0 for the focal group and V = 1 for the reference group. In Equation 17, however, the response variable Y need not be restricted to dichotomously scored categories, but can assume polytomously scored values.

Using simulated data for 25 items with four response categories, Miller et al. (1992) compared their approach to both the logistic regression model using a continuation ratio logit analysis (Agresti, 1990)

and the MH procedure with ordered responses to detect DIF in polytomously scored items. Two types of nonuniform DIF were simulated: (1) the item difficulty (b) parameters were equal for the focal and reference groups but the item discrimination (a) parameters differed, and (2) the a parameters were equal and the b parameters differed between groups. Their results suggested, not surprisingly, that the MH procedure did not detect the nonuniform DIF, and that both the logistic regression and linear discriminant function analysis methods were equally powerful in detecting nonuniform DIF due to differences in the a parameters. The logistic discriminant function analysis approach, however, was superior for detecting nonuniform DIF due to differences in the b parameters.

The procedures detailed by Miller et al. (1992) appear promising. Logistic regression models, although useful for detecting uniform and nonuniform DIF, become computationally cumbersome when applied to polytomously scored items. The logistic discriminant function analysis approach, on the other hand, may provide an elegant solution by treating the item response as an independent variable and requiring only one regression model per item. The efficacy of this method under other bias conditions requires further study.

UCI Methods

This section reviews developments in bias detection methods that operate within an assumed measurement model relating the observed measure Y to the latent variable W. The section is divided in three subsections. The first concerns methods for dichotomously scored Y measures, while the second addresses methods appropriate for polytomously scored measures. Both of these sections assume W to be unidimensional. The final section reviews methods that consider W as a multivariate latent variable.

Dichotomously Scored Measures

These methods assume that a unidimensional IRT model underlies performance on the studied measure. Throughout, it is assumed that parameter linking has been completed prior to the construction of any bias indexes. Before detailing the various methods, a brief discussion of the issues related to parameter linking and the related IRT problem of scale indeterminacy is presented.

Parameter linking. In the unidimensional IRT models for dichotomous items that are commonly used, the item and θ parameters are not identified without further constraints. For example, consider the three-parameter logistic model (3PLM)

$$P(Y = 1 \mid a, b, c, \theta) = c + (1 - c)\{1 + \exp[-Da(\theta - b)]\}^{-1}, \tag{18}$$

where c is the pseudo-guessing parameter, and D is the logistic scaling constant.

Let A and B be any constants. Then if $a^* = a/A$, $b^* = Ab + B$, and $\theta^* = A\theta + B$

$$P(Y = 1 \mid a, b, c, \theta) = P(Y \mid a^*, b^*, c, \theta^*). \tag{19}$$

Hence, the item and θ parameters are not unique and may be linearly transformed without altering the response probabilities. The same condition holds in the one-parameter logistic model (1PLM), the 2PLM, and in the normal ogive IRT model.

The problem of scale indeterminacy can be solved in several ways. One method is to scale θ so that its mean is 0 and its standard deviation (SD) is 1 for the group of examinees under study. Alternatively, the item difficulties could be scaled so that they have a mean of 0 and SD of 1 across the items under study. When multiple random samples are drawn from a single population (i.e., all examinees fit the same IRT model), separate standardizations can be done within each sample. The item parameters estimated within each sample will vary across samples because of sampling error. However, the separate estimates should be linearly related, as illustrated in Lord (1980). It should be possible to find values

for A and B that will linearly transform the item parameters in one sample to closely approximate those in another sample. In contrast to classical test theory, this invariance property is a well-known advantage of IRT (Lord & Novick, 1968).

The invariance property follows from the assumption that multiple samples are drawn from a common population, one in which all examinees can be fit by a common IRT model. The situation is quite different when studies of measurement bias are conducted. In these studies, it cannot be assumed that the groups under study belong to a common population. It is, in fact, this question that is being investigated. As a result, there may exist no values of A and B that will render the item parameters equivalent across the groups under study. It is still necessary to impose some constraints within each group for parameter estimation to proceed. An effort must be made to link the parameters across groups so that spurious group differences due to sampling, combined with the separate standardizations, are reduced. If ordinary linking procedures are used (e.g., linear transformations based on within-group item parameter estimates), the presence of some biased items may distort the development of the linking transformation. The problem is to find the proper linking transformation while avoiding distortions in the transformation due to biased items. This linking problem has been known for some time (Lord, 1980), but recent work (Candell & Drasgow, 1988; Kim & Cohen, 1992; Lautenschlager & Park, 1988; Park & Lautenschlager, 1990) has amplified the importance of the problem for bias investigations.

All linking procedures designed for bias investigations attempt to base the derivation of the linking transformation on only the unbiased items. These methods require a preliminary screening of biased items, followed by development of the linking transformation using the remaining items. The issue is whether a single screening is adequate, or whether several screening iterations followed by reevaluation of the linking transformation are needed. The question is important because the initial screening, which uses item parameter estimates whose metrics are linked using all items, may not properly identify all biased items. An iterative approach could improve the quality of the linking.

Lord (1980) suggested a single iteration procedure (attributed to Marco, 1977) that begins with an initial screen for biased items. The procedure for the 3PLM consists of the following steps:

1. Estimate item parameters for the total sample combining data from all groups. Standardize on the b parameters (i.e., force the mean b to be 0, with SD = 1).
2. Fix the guessing (c) parameters to the values in Step 1, and estimate the b and a parameters separately in each group again standardizing on the b parameters. Evaluate each item for bias.
3. Remove all items that are biased.
4. Combine groups and estimate θ using the reduced item set.
5. Using the θ estimates from Step 4, estimate the item parameters for all items (biased and unbiased) within each group separately.
6. Evaluate each item for bias using the item parameter estimates from Step 5.

In this procedure, the presence of biased items will distort the parameter estimates in Step 2, and may induce misclassification of items as biased or unbiased at that stage. The second screen in Step 6 attempts to correct any misclassifications.

Park & Lautenschlager (1990; Park, 1988) presented a modified version of Lord's procedure that iteratively improves the estimation. This modified procedure consists of the following steps:

1. Combine the groups and estimate θ for all examinees.
2. Separate the groups, and within each group estimate all item parameters using the θ estimates from Step 1. Evaluate all items for bias.
3. Remove the biased items.
4. Using the remaining items, combine the groups and estimate θ for all examinees.

5. Separate the groups, and within each group estimate all item parameters using the θ estimates from Step 4.
6. Evaluate all items for bias.
7. Repeat Steps 3–6 until the same items are identified as biased in successive iterations.

One difficulty with both Lord's procedure and this modified procedure is that they are computationally complex (Candell & Drasgow, 1988; Kim & Cohen, 1992). Repeated estimation of the item and θ parameters is required.

An alternative procedure suggested by Segall (1983) does not require repeated parameter estimation. Instead, a linking function is applied, and this function is iteratively improved. The procedure involves the following steps:

1. Item parameters are estimated within each group separately.
2. A linking function is developed to link all items.
3. After linking, all items are evaluated for bias. Biased items are removed.
4. A new linking function is developed using the remaining items, and applied to all items.
5. All items are evaluated for bias and the biased items are removed.

The last two steps are repeated until the same items are identified as biased on successive iterations. This procedure has been used with real data (Drasgow, 1987), and in simulations (Candell & Drasgow, 1988; Park & Lautenschlager, 1990).

The accumulated evidence suggests that iterative linking can improve the accuracy of bias detection in comparison to single iteration linking (Candell & Drasgow, 1988; Kim & Cohen, 1992; Lautenschlager & Park, 1988; Park & Lautenschlager, 1990). Lautenschlager & Park (1988) used simulated data in which biased items were created by introducing a second latent trait dimension. They found that when the linking transformation was based on all items with no attempt to remove biased items, misclassifications of items as biased or unbiased often resulted. The false negative rate, or the proportion of biased items that were not detected as biased, was fairly high in many of the conditions studied. The false positive rate was also high in some conditions.

McCauley & Mendoza (1985) used simulated data to evaluate several bias indexes, basing parameter linking on all items without removal of biased items. They reported some evidence of elevated false positive rates that were possibly due to the linking procedure. Several studies have directly compared single iteration linking to multiple iterations linking (Candell & Drasgow, 1988; Kim & Cohen, 1992; Park & Lautenschlager, 1990). In general, all studies have found improved bias classification accuracy with multiple iterations. Park & Lautenschlager (1990), however, found that very high proportions of biased items could lead to misclassification even under iterative linking. They found, for example, that the most serious classification problems arose with items that were weakly biased and were classified as unbiased. Similarly, Kim & Cohen (1992) varied group sample sizes and found that iterative linking is most needed in small samples (i.e., $N = 300$ per group, using a 2PLM).

In addition to the choice between single or multiple iterations, the linking function itself must be selected. Selecting the linking function means selecting the constants A and B. There are two broad classes of methods for deriving values for the linking constants. One class derives the linking constants using only first and second moment information from the distributions of the item parameter estimates in the two groups (Linn, Levine, Hastings, & Wardrup, 1981; Loyd & Hoover, 1980; Marco, 1977; Vale, 1986; Warm, 1978). These will be denoted *moment methods*. The second class uses additional information deriving constants that minimize group differences in item or test response functions (Divgi, 1985; Stocking & Lord, 1983). These linking methods will be denoted *characteristic curve methods*. The characteristic curve methods use more information and might be expected to be more accurate, but they also require more computation. In practice, poorly estimated curves could degrade the quality

of the linking.

The available research does not clearly support the superiority of either type of method in all situations. Stocking & Lord (1983) found that their characteristic curve method was more accurate than the moment method developed by Linn et al. (1981), but Candell & Drasgow (1988) reached the opposite conclusion. Baker & Al-Karni (1991) found no differences between the characteristic curve method and a moment method developed by Loyd & Hoover (1980). Using simulated data, Kim & Cohen (1992) compared three linking methods, including the characteristic curve method, a weighted moment method, and a minimum χ^2 method (Divgi, 1985). The characteristic curve and minimum χ^2 methods were slightly better than the moment method. In large samples ($N = 600$), the methods performed equally well. Overall, the evidence tends to favor the characteristic curve methods, but the question remains open. Apparently, the answer depends on conditions such as sample size, the number of biased items, the nature of the true model, and the length of the test.

Area measures. Area measures of bias express the difference between the reference and focal group IRFs as some function of the area between the IRFs, calculated over a selected interval on the θ scale. Thus, by defining $P_R(\theta)$ and $P_F(\theta)$ as the IRFs on the studied item for the reference and focal groups, respectively, a general definition of an area measure is

$$A = f_s[P_R(\theta) - P_F(\theta)] , \tag{20}$$

defined for θ in the interval $S = (\theta_L, \theta_U)$ where L and U indicate lower and upper values, respectively.

Many choices for the function f and the interval boundaries are possible. Area measures differ by whether (1) absolute, unsigned, or signed differences are used, (2) the interval S is bounded or unbounded, (3) continuous integration or discrete approximation is used in f, and (4) the differences in f are equally weighted or differentially weighted.

Early area measures (Ironson & Subkoviak, 1979; Rudner, 1977; Rudner, Getson, & Knight, 1980a, 1980b) used bounded intervals with discrete approximation. Rudner (1977) proposed the unsigned index

$$R = \sum_{j=-3}^{3} |D_j| \Delta, \text{ for } D_j = P_R(\theta_j) - P_F(\theta_j) \tag{21}$$

with $S = (-3, +3)$ and Δ a small interval (e.g., $\Delta = .005$). The interval S is divided into 600 intervals, each of width $\Delta = .005$, and summation is performed across these 600 intervals. The measure R is easily converted to a signed measure by removing the absolute value operator, allowing both positive and negative differences to be summed.

Linn et al. (1981) also used the measure R. Linn et al. (1981) and Ironson & Subkoviak (1979) defined the "base high" index as

$$R_H = \sum_{S_H} (D_j) \Delta , \tag{22}$$

and Linn et al. (1981) defined the "base low" index as

$$R_L = R - R_H . \tag{23}$$

For R_H, the summation only includes intervals in which the IRF for the designated "base" group is above the IRF for the other group. R_L is then the total area in R minus R_H. The R_H and R_L indexes allow the investigator to understand the direction of bias in absolute terms, while avoiding the cancellation that may occur in the signed index. Linn et al. (1981) proposed another unsigned index,

$$\text{LSS} = \sum_{j=-3}^{3} [(D_j \Delta)^2]^{1/2} , \tag{24}$$

which essentially provides the root mean squared difference between the IRFs.

Shepard, Camilli, & Williams (1984, 1985) proposed variations on these signed and unsigned indexes that restricted the summation to only the θ values found in the samples under study:

$$SOS_1 = \frac{1}{N}\sum_{j=1}^{N} D_j^2 , \tag{25}$$

$$SOS_2 = \frac{1}{N}\sum_{j=1}^{N} D_j^2/\sigma_{D_j}^2 , \tag{26}$$

$$SOS_3 = \frac{1}{N}\sum_{j=1}^{N} |D_j|(D_j) , \tag{27}$$

and

$$SOS_4 = \frac{1}{N}\sum_{j=1}^{N} |D_j|(D_j)/\sigma_{D_j}^2 . \tag{28}$$

In these indexes, the interval S is bounded, but the boundaries are determined by the available data. In these indexes, $N = N_R + N_F$ is the total sample size, and θ_j is the jth person's θ value in D_j, regardless of group membership. SOS_3 is identical to SOS_1 except that the sign of the difference is retained and replaced after squaring. Use of either of these indexes requires that estimated θ values be available.

Indexes that provide for differential weighting of the IRF differences also have been proposed. Linn et al. (1981) created

$$R_W = \sum_{j=-3}^{3} |D_j|\Delta/\sigma_{D_j} \tag{29}$$

from R by weighting each difference in inverse proportion to the estimated standard error (SE) of the difference, which gives small weights to differences with large SEs, denoting uncertainty in estimation. Linn et al. (1981) found that weighted and unweighted indexes gave essentially identical results.

Shepard et al. (1984) created weighted SOS indexes (SOS_2 and SOS_4). They used the reciprocals of the estimated variances of the differences as weights. They found that although the weighted and unweighted indexes did not markedly differ, the weighted indexes provided more interpretable results. This advantage for the weighted indexes was not found in Shepard et al. (1985), in which smaller sample sizes were used.

All of the above measures use discrete approximation. More recently, continuous integration over either bounded or unbounded intervals has been used (Kim & Cohen, 1991; Raju, 1988, 1990). These indexes have the general form

$$A_c = \int_s f[P_R(\theta) - P_F(\theta)]d\theta , \tag{30}$$

with $S = (-\infty, +\infty)$ in the unbounded case or $S = (\theta_L, \theta_U)$ in the bounded case. In nearly all cases, either the unsigned function

$$f = |P_R(\theta) - P_F(\theta)| \tag{31}$$

or the signed function

$$f = [P_R(\theta) - P_F(\theta)] \tag{32}$$

is used. Raju (1988) derived closed form expressions for both cases when $S = (-\infty, +\infty)$ and both

IRFs are either one-, two-, or three-parameter logistic functions. In the 3PLM case, A_c in Equation 30 is infinite if the cs differ between groups. When the IRFs are both based on the 1PLM, the signed and unsigned indexes have the same absolute values. Interestingly, only the unsigned indexes are influenced by any group differences in the a parameters in the 2PLM and the 3PLM. The signed indexes in these models are a function only of the b parameters and, if present, the common value of the c parameter.

Kim & Cohen (1991) developed closed-form expressions for signed and unsigned indexes using Equations 31 and 32 in the bounded case, in which both the signed and unsigned indexes are influenced by group differences in the a and c parameters under the 2PLM or the 3PLM. These indexes are finite when groups differ with respect to the c parameter in the 3PLM.

A long recognized difficulty with all of the area measures is that the SEs of the measures are unknown, making it difficult to evaluate the statistical significance of any differences found (Linn et al., 1981). The use of weighted indexes, such as those proposed by Linn et al. (1981), still does not accomplish this goal. Raju (1990) presented asymptotic SE formulas for Raju's (1988) unbounded signed and unsigned measures. These SEs can be used to generate z tests of significance under normality assumptions. In an analysis of items from a vocabulary test, Raju (1990) found that the asymptotic test statistics gave sensible results and were fairly consistent with MH statistics calculated on the same data. More empirical study of these test statistics is needed. At present, no SEs for bounded area indexes are available.

An alternative approach to the significance problem is to construct confidence bands around the IRFs, or around their differences. Linn et al. (1981) suggested the use of separate bands for each IRF, and illustrated their use. These bands are built using the estimated SE of the difference between the IRFs, evaluated at each θ value. Similarly, Thissen & Wainer (1990) and Lord & Pashley (1988) developed "confidence envelopes" for logistic IRFs. These envelopes are built by first deriving confidence bounds on the parameter vector and then translating these bounds in terms of the IRF, in a manner similar to that used to construct confidence bands in linear regression. More recently, Pashley (1992) extended this methodology to produce confidence bands for the difference between IRFs from different examinee groups and illustrated the approach with the 3PLM. An advantage of the methods developed by Thissen & Wainer (1990), Lord & Pashley (1988), and Pashley (1992) is that the confidence bands developed are simultaneous, rather than pointwise as in the Linn et al. (1981) approach. Joint probability or confidence statements have a firmer basis in the simultaneous approach.

In general, there appears to be little advantage in using discrete approximations if simple difference functions (such as Equations 31 or 32) are used, and if the 1PLM, 2PLM, or 3PLM are fit. Computation of the area is simplified by the methods of Raju (1988) and Kim & Cohen (1991). This advantage of the continuous integration measures may be lost if more complicated functions are used, such as those that use differential weighting. Weighted continuous indexes comparable to those used with the discrete approximation measures are not widely used, although Raju (1988) discussed weighting by the prior distribution of θ.

The choice between bounded and unbounded area measures remains unclear. One disadvantage of the unbounded measures is that they are infinite when there are group differences in the c parameter in the 3PLM. If these measures are used with a 3PLM, common c parameters must be used. Kim & Cohen (1991) compared bounded and unbounded area measures under the 3PLM using real data in which bias was experimentally manipulated (Subkoviak, Mack, Ironson, & Craig, 1984). Few differences were found between the two types of area measures in the detection of the biased items or in false positive rates. Subkoviak et al. (1984) also used the bounded 3PLM continuous area measure with group differences in c parameter values, and found this measure to be slightly superior to the

other measures, both bounded and unbounded.

A disadvantage of the bounded measures is that their value depends on the endpoints of the selected interval—this choice is arbitrary to some extent. On the other hand, IRF differences in the extreme regions of the θ scale carry few implications for practical measurement, and should be excluded by using a bounded measure. Raju (1988) noted that the bounded area measures may be substantially smaller than the unbounded measures. More comparative studies will be needed to resolve the debate.

The choice between signed and unsigned area indexes is also not clear. In the 1PLM case, or when neither the a nor c parameters differ between groups, the signed and unsigned continuous indexes have the same absolute values. Comparative studies have been performed using real data (Cohen, Kim, & Subkoviak, 1991; Ironson & Subkoviak, 1979; Kim & Cohen, 1991; Raju, 1990; Shepard et al., 1981; Shepard et al., 1984, 1985; Subkoviak et al., 1984) and simulated data (McCauley & Mendoza, 1985; Shepard et al., 1985). No clear superiority for either signed or unsigned indexes has emerged from these studies. The two indexes may not correlate highly across items because of the bipolar nature of the signed index.

Many of the above studies used the 3PLM with common c parameter values. In this case, items showing bias must have group differences in a parameters before performance differences will appear between signed and unsigned indexes. The Shepard et al. (1985) simulation studies used the 3PLM, but created bias only as a function of the b parameters, which reduced the chances of finding differences between signed and unsigned indexes. No substantial differences were found. Raju (1990) used his asymptotic SEs to conduct significance tests for both signed and unsigned indexes. In real data on a 40-item vocabulary test, the unsigned index identified more items as biased than did the signed index. It is not clear whether differences in the test statistics in the two cases account for this phenomenon. Although a substantial number of studies have compared signed and unsigned indexes in real data, simulation studies are rare and more are needed.

Wald statistics. Lord (1980) proposed that the null hypothesis of identical IRFs across groups be tested using a test for equality of item parameters under an assumed parametric model. To illustrate, suppose that the 3PLM is found to fit in both groups. Lord (1980) recommended that the test for parameter equality be confined to the b and a parameters, with c parameters constrained to be equal across groups.

Under maximum likelihood estimation (joint or marginal), an estimate of the 2×2 covariance matrix for the parameter estimates is available, calculated separately for each group. Let the covariance matrix estimates for the focal and reference groups be $\mathbf{S}_F$ and $\mathbf{S}_R$, respectively. These are found as the inverses of the information matrices within each group. The null hypothesis to be tested is

$$H_0: a_F = a_R, \quad b_F = b_R , \tag{33}$$

stating that item parameters are equal across groups. To conduct the test, a 2×1 vector of parameter estimate differences is formed as

$$\mathbf{V}' = [\hat{a}_F - \hat{a}_R, \quad \hat{b}_F - \hat{b}_R] , \tag{34}$$

with the estimated covariance matrix $\mathbf{S} = \mathbf{S}_F + \mathbf{S}_R$. Then standard theory implies that under H_0,

$$\chi^2 = \mathbf{V}'\mathbf{S}^{-1}\mathbf{V} , \tag{35}$$

which has a central χ^2 distribution with 2 *df* in large samples. The *df* are equal to the number of constraints placed on the parameters under H_0. This is a large sample test.

The test just described for the 3PLM can be constructed in an analogous way for the 1PLM or the 2PLM. In the 1PLM case, the χ^2 test statistic reduces to

$$\chi^2 = \frac{(\hat{b}_F - \hat{b}_R)^2}{S_R^2 + S_F^2}, \tag{36}$$

where S_R^2 and S_F^2 are the estimated variances of $\hat{b}_R$ and $\hat{b}_F$, respectively. This statistic is simply the square of the z statistic proposed by Wright, Mead, & Draba (1976; "the Draba statistic") for the Rasch model. The χ^2 in Equation 36 is evaluated with $df = 1$.

Thissen et al. (1988) pointed out that the test statistic in Equation 35 is a member of the class of Wald (1943) statistics that are frequently used with maximum likelihood estimation. Amemiya (1985) and Rao (1973) presented the relevant asymptotic theory that underlies these statistics. Although much is known about the large sample behavior of these statistics, the small sample behavior has not been thoroughly investigated. The minimum sample size required for convergence of Equation 35 to the χ^2 distribution in IRT applications is unknown at present. An important question in understanding the behavior of the statistic in finite samples concerns the adequacy of the estimator **S** (the covariance matrix for the parameter estimates). Under joint maximum likelihood estimation, **S** may be poor even in large samples in the 2PLM and the 3PLM. McLaughlin & Drasgow (1987) demonstrated using simulations that SE estimates were negatively biased under these models with joint maximum likelihood estimation. Type 1 error rates for Lord's test were inflated as a result. This problem may be lessened under marginal maximum likelihood estimation (MMLE) because the consistency of the item parameter estimates is not an issue with this estimation method. The sample behavior of **S** under MMLE needs to be investigated.

Lord's χ^2 test has been compared to other IRT test procedures in a number of studies (McCauley & Mendoza, 1985; Shepard et al., 1981; Shepard et al., 1984, 1985; Thissen et al., 1988). Nearly all of these studies have used the test with the 2PLM or the 3PLM. A common finding is that the performance of the statistic correlates fairly closely with that of unsigned area indexes (Shepard et al., 1981; Shepard et al., 1984; McCauley & Mendoza, 1985). Studies of the statistic using the 3PLM have used common c parameter values, as recommended by Lord (1980). MMLE methods with prior distributions on the c parameters can reduce the numerical problems encountered in estimating these parameters. Improved estimation may permit wider use of the χ^2 statistic in testing all three parameters (a, b, and c) for equality. More studies are needed to evaluate the accuracy of the statistic in this case, especially with moderate to small samples.

One criticism of Lord's χ^2 statistic is that the null hypothesis may be rejected even when the unsigned area between IRFs is fairly small throughout the θ range in which most data appear. Artificial examples can be constructed to illustrate this (Linn et al., 1981), and the phenomenon also appears in real data analyses (Shepard et al., 1984). The argument here parallels the argument about the merits of bounded versus unbounded area indexes. Both Lord's χ^2 statistic and the unbounded area indexes have the advantage of mathematical tractability, permitting the use of SEs and formal hypothesis tests. The best approach may be to supplement the χ^2 with some calculation of a bounded area index when the χ^2 is significant.

Likelihood ratio methods. An alternative procedure for testing the equality of item parameters between groups is based on the likelihood ratio for two models: In Model 1 (M_1) the item parameter values on the studied item may vary between groups, and in Model 0 (M_0) these parameter values are constrained to equality between groups. Thissen et al. (1988) and Thissen, Steinberg, & Gerrard (1986) described the use of this test procedure. In the likelihood ratio methods, a subset of items must be selected to serve as "anchor" items. These anchor items are assumed to be unbiased and link the metric for parameter estimation. Assuming that this requirement is met, the likelihood ratio test statistic (LR) is calculated as

$$LR = -2\ln[L_0/L_1] \; . \tag{37}$$

In Equation 37, L_1 is the value of the likelihood function for M_1, and the numerator L_0 is the value of the likelihood for M_0. In general, $L_0 < L_1$, and $LR > 0$. *LR* has a χ^2 distribution in large samples under the null hypothesis that the model yielding L_0 fits. The *df* for *LR* are the number of constraints required to derive M_0 from M_1. For example, suppose that the studied item is assumed to follow a 3PLM and that M_1 permits all three parameters to be free. If M_0 constrains all three parameters to equality, L_R will have $df = 3$. Amemiya (1985) and Rao (1973) provided a more detailed discussion of the theory underlying *LR*.

As noted by Thissen et al. (1986), *LR* is easily extended to permit simultaneous tests of bias for multiple items. In this extension, more than one studied item is constrained under M_0, which allows conclusions about the presence of bias in a set of items. However, rejection of M_0 would leave open the question of which items are biased. Presumably, a post hoc item-by-item search could be initiated with appropriate controls on the Type I error rate. Another extension of *LR* in Equation 37 would involve tests for bias across more than two examinee groups.

LR should, in theory, give results close to those given by Lord's χ^2 test, provided that samples are large enough. Asymptotically, Wald and *LR* statistics converge to the same distribution under M_0 (see Buse, 1982, for a description of the relationship between these statistics). In practice, the two statistics may differ to some degree, in part because of the difficulty in estimating the covariance matrix required in the Wald procedure. Thissen et al. (1988) provided a good discussion of the relative merits of the two statistics in bias applications.

At least two difficulties arise in any practical use of the *LR* procedure. First, a set of unbiased anchor items must be available. This may not be a problem in large-scale testing, in which item banks are available containing prescreened items. Prescreened items can be administered along with the target items. In other situations, no items that are known to be unbiased may be available. There is some risk in this case that the anchor items will include some biased items. The influence that such items may have on the distribution of the test statistic is unclear, and requires further study. This problem is analogous to the parameter linking problem discussed earlier.

A second problem is the lack of software for performing the required calculations. Calculation of the likelihood values requires multiple-group simultaneous estimation. The parameter constraints in M_0 are equality constraints that operate across groups, and the constraints should be imposed during simultaneous estimation. At present, the only widely available program that can easily perform the estimation is MULTILOG (Thissen, 1990). Neither LOGIST (Wingersky, Barton, & Lord, 1982) nor BILOG (Mislevy & Bock, 1984) will permit the required estimation under parameter equality constraints that operate across groups.

Approximate procedures. Two other procedures have been proposed that use θ estimates, but do not perform direct comparisons of item parameters. Both procedures assume that the measures in each group can be fit adequately by a unidimensional IRT model.

Linn & Harnisch (1981) developed a procedure that assumes that each group can be fit by a 3PLM. The detection procedure begins by combining the groups and estimating all item and θ parameters under a 3PLM. Following this estimation, the *target group* (i.e., the group against which bias is suspected) is selected. The θ scale is divided into intervals in the target group using the θ parameter estimates. Linn & Harnisch (1981) created five intervals in their example. For example, let P_{ij} be the estimated probability of passing the ith item for the jth person in the target group. This probability is estimated using the parameter estimates obtained earlier. The estimated proportion of examinees who should pass the ith item in the gth subgroup is

$$P_{ig} = \frac{1}{n_g} \sum_{j \in g} P_{ij} , \tag{38}$$

where n_g is the number of examinees in the gth subgroup. Using the response data, the actual number of examinees who passed the ith item in the gth subgroup can be found. Let this number be O_{ig}. Both the expected and actual numbers of passing examinees for the entire target group are formed as

$$P_{i.} = \sum_{g=1}^{m} n_g P_{ig} \Big/ \sum_{g=1}^{m} n_g \tag{39}$$

and

$$O_{i.} = \sum_{g=1}^{m} n_g O_{ig} \Big/ \sum_{g=1}^{m} n_g , \tag{40}$$

respectively. Bias indexes then may be calculated, both for subgroups and for the entire target group, as

$$D_{ig} = O_{ig} - P_{ig} , \ (g = 1, \ldots, m) \tag{41}$$

and

$$D_{i.} = O_{i.} - P_{i.} . \tag{42}$$

In addition, *standardized difference scores* may be calculated as

$$Z_{ig} = \frac{1}{n_g} \sum_{j \in g} \left\{ \frac{Y_{ij} - P_{ij}}{[P_{ij}(1 - P_{ij})]^{1/2}} \right\} \tag{43}$$

and

$$Z_{i.} = \sum_{g=1}^{m} n_g Z_{ig} \Big/ \sum_{g=1}^{m} n_g , \tag{44}$$

where Y_{ij} is the score for the jth person on the ith item.

The bias indexes in Equations 41 through 44 measure discrepancies between the actual responses in the target group and those predicted under a common 3PLM. Large discrepancies are expected when the 3PLM with parameter estimates obtained from the pooled sample does not fit the target group. The logic of the method is that if it is assumed that both groups are fit by a (possibly different) 3PLM, large discrepancies may indicate that parameter values differ in the two groups. Parameter differences then would imply bias. Linn & Harnisch (1981) proposed this method for use when sample sizes are too small to conduct 3PLM estimation within the target group alone. Instead, estimation is done in the larger pooled sample.

Linn & Harnisch (1981) illustrated the use of their procedure in real data on a 46-item math test, but did not present extensive studies of the method. Two comparative studies using the procedure in combination with other procedures have been conducted (Ironson, Homan, Willis, & Signer, 1984; Shepard et al., 1985). Both studies compared the Linn & Harnisch (1981) procedure to the delta plot method (Angoff & Ford, 1973) and Lord's χ^2 procedure. Ironson et al. used real data in which bias was manipulated by altering the reading level on some math story problem items. The target group was defined by scores on an independent reading test. The Linn & Harnisch (1981) bias indexes were

found to correlate more highly with manipulated bias than did the indexes provided by the other two methods. Shepard et al. (1984) used both real math test data and simulated data. Biased items in the math test had been identified previously using large samples (Shepard et al., 1984). The target group sample size was 300 in both datasets. The Linn & Harnisch (1981) method performed much better than the delta plot method in detecting biased items, and performed slightly better than Lord's χ^2 procedure.

These studies support the value of the method in small samples, but the properties of the method have not been thoroughly studied. For example, Linn & Harnisch (1981) noted that biased items may disrupt the initial parameter estimation, leading to later distortions in the bias indexes. The effects of varying proportions of biased items on detection accuracy are unknown. Other questions that could be addressed with simulated data include: (1) How sensitive is the method in detecting bias? and (2) How does the sensitivity vary with test length?

A second bias detection procedure was presented by Hulin, Drasgow, & Komocar (1982). This procedure assumes that the measure under study can be fit by a 2PLM in all groups. The procedure begins by estimating 2PLM item and θ parameters separately in each group, and then "standardizing" the b parameters within each group (i.e., fixing the mean to 0 and the SD to 1). Next, the θ scale is divided into intervals within each group. The empirical IRFs, or the proportions of examinees passing the studied item, are plotted for each group using the midpoint of each interval on the θ scale. Let $O_k(\theta_g)$ be the observed proportion passing the studied item in the gth interval within the kth group. These observed proportions are transformed to logits

$$L_{gk} = \log\left[\frac{O_k(\theta_g)}{1 - O_k(\theta_g)}\right] . \tag{45}$$

Note that if the observed proportions fit the 2PLM perfectly, then

$$L_{gk} = Da(\theta_g - b) . \tag{46}$$

This logit is a linear function of θ_g. If the groups under study have identical IRFs on the studied item, these linear functions should be identical. Hulin et al. (1982) proposed that these linear functions be regarded as regression functions. The functions may be plotted and a standard F test for equality of regression functions may be applied. The item is considered to be biased if these functions are not equal.

Although this method has been used with real data involving language translations of attitude scales (Hulin et al., 1982; Hulin, Drasgow, & Parsons, 1983), no systematic studies of the method's properties have been published. As in Linn & Harnisch's (1981) method, biased items may distort the initial θ estimates, leading to detection inaccuracies. Violations of the 2PLM assumptions also can be expected to produce inaccuracies. The robustness of the method to these violations could be studied with simulated data. Finally, advances in IRT parameter estimation methods may have rendered this method obsolete, because estimation within the 2PLM is no longer difficult.

Polytomous Measures

Polytomously scored measures are becoming more important due to increased emphasis on performance-based measurement and constructed-response items. Ability or achievement test items that are awarded partial credit are polytomously scored, as are many attitude and personality scales (e.g., Likert scale attitude items). Testlets (Wainer & Kiely, 1987) are another source of polytomous measures. A testlet consists of a cluster of related test items that are scored together. A typical example is a reading comprehension testlet in which a paragraph is followed by a series of questions

about the paragraph. The examinee's score is usually some function of the number of questions answered correctly. The bias detection procedures discussed below assume that the observed measures fit a unidimensional IRT model.

IRT models for polytomous measures have a long history (Rasch, 1960), but there has been relatively little work on bias detection for these measures. This is due in part to the wider use of dichotomously scored measures in educational and employment testing and to the complexity of IRT models for polytomous measures. More parameters are needed per measure than in the dichotomous case. Furthermore, the increase in response options means that the number of possible response patterns across measures is greatly increased, complicating the assessment of fit. In some cases, goodness-of-fit tests will have little power except in very large samples.

In spite of these problems, bias detection in polytomous IRT models is important. Wainer, Sireci, & Thissen (1991) and Shealy & Stout (1991) described some advantages in considering testlets as the unit of analysis in bias detection. Small amounts of bias at the individual item level may accumulate across items, producing larger biases at the testlet level. An empirical example of this was given by Wainer et al. (1991). Bias "cancellation" also may occur if small biases in different directions cancel at the testlet level. Another reason for using testlets is that some tests may not meet the usual local independence assumptions at the item level. Violations of the assumption may come about as a result of the item structure—a group of items may refer to a common stimulus (e.g., a reading selection). Dependence could also arise for subtle, unanticipated reasons, as illustrated in Wainer & Kiely (1987). Rosenbaum (1988) presented some results extending the local independence assumptions to the testlet or "item bundle" level. In this extension, the idea of conditional association developed earlier by Holland & Rosenbaum (1986) was applied to testlets, resulting in useful theorems that characterize unidimensional models for testlets. This work is valuable in developing a theoretical basis for bias investigations at the testlet level.

Models for polytomous data. Thissen & Steinberg (1986) presented a useful taxonomy of IRT models for polytomous measures. If the possibility of guessing is ignored, Thissen & Steinberg demonstrated that nearly all existing models can be viewed as special cases of either Samejima's (1969) graded response model or Bock's (1972) nominal response model. Thissen & Steinberg denoted the class of models derived from Samejima's (1969) model as "difference models" and the class defined by Bock's (1972) model as "divide-by-total" models. These labels loosely describe the functional form of the response functions in each case.

There are several important special cases within the "divide-by-total" class. Thissen & Steinberg (1986) showed that all of the cases can be derived from Bock's model through parameter restrictions. Two special cases are the partial credit model (PCM; Masters, 1982) and the rating scale model (Andrich, 1978). Both of these models are members of the Rasch family of IRT models (Masters & Wright, 1984). An important feature of these models is that given items fitting the model, the sum of the item scores is a sufficient statistic for the θ parameter. The availability of a sufficient statistic is important in bias detection applications (Meredith & Millsap, 1992; Zwick, 1990).

The PCM includes $m - 1$ b parameters for each m-category item. Masters (1982) described these as "step difficulties," viewing completion of the item as involving m steps. An examinee's score on the item denotes the number of steps completed. The steps may differ in difficulty, and the step difficulties need not be ordered. Also, step difficulties may differ among items. No a parameter is included to allow different levels of discrimination among items. The model can be generalized to include such parameters (Muraki, 1992), but their inclusion destroys the sufficiency property.

The rating scale model can be derived as a special case of the PCM by restricting each step difficulty to be the sum of two parameters, one that depends only on the item, and one that depends only

on the step. The latter can be viewed as "threshold" parameters, and have common values across items. This model is intended for use with items having common response scales, such as Likert scaled items in attitude measurement. The model is generally not appropriate for items that differ in the number of response categories.

Bock's (1972) nominal model is the most general model in the divide-by-total class. The model includes difficulty parameters for each item and category, and discrimination or slope parameters for each item and category. Many special cases of the Bock model that are intermediate between the Rasch models and the general Bock model can be derived through appropriate parameter restrictions.

Parameter linking. As is true for dichotomously scored measures, the parameters in the models for polytomous measures are generally not identified without further constraints. In Bock's nominal model for example, constraints are needed to identify the difficulty and discrimination parameters. In the PCM, the sum of all of the step parameters across items is constrained to be 0. The general problem of implementing identification constraints while linking metrics across groups has not received much attention in the literature. As noted earlier, the presence of biased items can greatly complicate the linking process. These complications are expected to occur in the polytomous case as well. This problem should receive more attention if polytomous models are to be used fruitfully in bias investigations.

Evaluating bias. Once an appropriate model has been found for the groups under study, the invariance of the IRFs over groups can be tested. Potentially, any of the three major approaches reviewed for the dichotomous case could be used here: area statistics, likelihood ratio procedures, or Wald statistics.

Area statistics will be cumbersome to use with polytomous measures. Each measure has multiple response categories. If there are m categories, there will be $m - 1$ different response functions under the model. Each of these could be compared across groups, resulting in $m - 1$ area statistics for a single item. The separate area measures could be combined into a composite index. Given these multiple response functions, there are more opportunities for sign reversals in the area measures than in the dichotomous case. Biases may reverse directions across score categories. Although area measures can theoretically be calculated, there are no published examples of their use in polytomous data.

Likelihood ratio procedures have been used in bias investigations of polytomous measures, and in goodness-of-fit testing outside of the bias context (Thissen & Steinberg, 1986; Thissen, Steinberg, & Mooney, 1989; Wainer et al., 1991). The procedure is similar to that used in the dichotomous case, but there are potentially more hypotheses to investigate. Group differences in response functions may appear in many forms. For example, suppose that Bock's (1972) nominal model is used with items having m response categories. In the full model there will be $2(m - 1)$ item parameters for each item. Any one of these may differ between groups. If the UCI null hypothesis is rejected, the investigator must decide which of the potential $2(m - 1)$ parameters to test for invariance. A careful sequence of nested hypothesis tests could be used, perhaps beginning with tests for invariance of all difficulty or all discrimination parameters.

Thissen et al. (1989) demonstrated the use of Bock's nominal model in testlet data on four reading comprehension testlets, but did not use the procedure for bias detection. Thissen & Steinberg (1986) illustrated the use of Samejima's graded response model and Bock's model in fitting several types of measures, again outside the context of bias investigation.

Wainer et al. (1991) used Bock's model in a bias investigation of another reading comprehension test consisting of four testlets. MULTILOG (Thissen, 1990) was used for estimation and hypothesis testing in this study and in the two previous studies. An interesting feature of the Wainer et al. (1991)

study is that both internal and external matching criteria were used. In the internal case, only the data on the four testlets were used in the analysis. Hence, θ estimates were based on the equivalent of a four-item test. In the external case, examinee data on six multiple-choice items that were external to the four testlets were included in the analysis. These external items were fit using the 3PLM. The four testlets then were analyzed individually, each in turn being included with the six-item external anchor. Parameters for the external anchor were constrained to be equal across groups. Two models were fit: the "no bias" model, M_0, in which the testlet's parameters were constrained to be equal across groups, and an unconstrained model, M_1, in which the testlet's parameters were permitted to vary across groups. The log likelihoods for these models then were used in a likelihood ratio test, in a manner identical to that described earlier for the dichotomous case.

Wainer et al. (1991) noted that using the external anchors provided an advantage in efficiency relative to the internal anchor. In their example, the external anchor consisted of six dichotomous items, creating two possible score patterns. In contrast, the internal anchor consisted of four testlets, each with 10 response categories, resulting in 10 possible patterns. The reduction in the number of score patterns with the external anchor provided computational savings. To be useful, however, the items used in the external anchor should themselves be free of bias.

Wald statistics also could be used in tests of UCI under any of the polytomous models reviewed here. These models require more parameters than in the dichotomous case, resulting in larger *df* if simultaneous invariance restrictions are applied. SE estimates or error covariance matrices are required. Covariance matrices could be obtained as a byproduct of maximum likelihood estimation. These are estimated separately for each group, and then are substituted for $\mathbf{S}_F$ and $\mathbf{S}_R$ in the $\mathbf{S}$ matrix in Equation 35. Group differences in the parameter estimates are substituted for $\mathbf{V}'$ in Equation 34. The test statistic then is given in Equation 35.

Ferrara & Walker-Bartnick (1990) presented an example of the use of the PCM in a bias investigation of direct writing samples in an essay test. Essays were scored by raters using modified holistic procedures, and the ratings were averaged to yield a score that included seven categories. The statistical phase of the investigation used the Draba (1978) statistic to assess group differences in individual item parameters. As noted earlier, the square of the Draba statistic can be viewed as a single *df* Wald statistic. There are at least two difficulties in using the Draba statistic in the polytomous case. First, multiple tests are required for each item due to the many parameters. Second, the individual parameter tests do not reflect the covariances among the different parameter estimates. The individual tests are evaluated independently, but the estimates are not independent. A solution to this problem is to conduct simultaneous tests using the parameter estimate covariance matrix, assuming that this matrix is available.

Multidimensional IRT Models

All of the bias detection procedures discussed thus far have been based on the assumption of a unidimensional latent trait underlying test performance. Most practical applications of IRT make this assumption. However, the definition of UCI used here does not require that the latent trait space be unidimensional. Some of the earliest attempts to study measurement bias employed factor analysis in a multidimensional context (e.g., Thurstone, 1947). More recently, measurement bias itself has been viewed as evidence of multidimensionality (Ackerman, 1992; Kok, 1988; Shealy & Stout, 1993). Hence there is reason to examine multidimensional IRT models and their role in bias detection.

Recently there have been some important contributions to the methods available for assessing dimensionality. These new procedures may be useful in bias investigations as preliminary analyses. Holland & Rosenbaum (1986) presented theorems that give conditions that must be met under any uni-

dimensional, monotone latent variable model. These conditions involve forms of association among the observed measures fit by the model. The conditions can be evaluated empirically. This provides a very general test for unidimensionality in that no specific parametric form need be assumed for the underlying measurement model.

Stout (1987) presented a different method for assessing dimensionality that is also nonparametric. This procedure provides a large-sample statistical test for dimensionality that has performed well in simulations. A third development is the full-information factor analysis procedure developed by Bock and his colleagues (Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988). This factor analytic method can be applied to dichotomous item data, but avoids the known pitfalls encountered when factoring phi coefficients (see Hattie, 1985, for a discussion). Also, modifications are included to adjust for guessing and for omitted items. This method has been implemented in the TESTFACT program (Wilson, Wood, & Gibbons, 1987). The procedure can efficiently handle up to 100 items, with up to five factors.

The bias detection procedures reviewed here all assume that there are multiple latent variables that influence Y. One or more of these variables are the "traits" that Y is intended to measure, but there may be additional influences as well. For now, both types of latent variables will be included in W. In addition, these bias detection procedures specify a measurement model relating observed measures Y to the latent variables W. Bias is investigated as violations of UCI, possibly limiting UCI to a subset of the latent variables in W (as explained below).

Two cases can be distinguished. First, measures in Y may all be continuous. In this case, the most commonly used measurement model is the common factor model. The second case arises when Y contains dichotomous or ordered-categorical data. In this case, both factor analytic and multidimensional item response models have been used.

Continuous measures. The study of measurement bias in continuous measurements has largely relied on the common factor model. Factorial invariance across populations has been of interest since the early days of factor analytic theory (Thurstone, 1947). The development of efficient estimation methods for restricted factor analysis has enabled researchers to test hypotheses for factorial invariance under normal theory (Jöreskog, 1971). There are several good reviews of restricted factor methods for tests of invariance (Bollen, 1989; Byrne, 1989; Byrne, Shavelson, & Muthén, 1989; Rock, Werts, & Flaugher, 1978). Because this area is already well-documented, only two recent developments will be discussed: testing invariance in latent mean structures and the use of asymptotically distribution-free (ADF) estimation.

Although theory for investigating invariance in latent mean structures in addition to covariance structures has been available for some time (Sörbom, 1974), investigators have been slow to use these methods. Byrne et al. (1989) reviewed the literature and found only two published empirical studies that investigated group differences in latent mean structures. Yet as illustrated by these authors, an analysis that includes latent means may uncover group differences that cannot be studied if only covariance structures are analyzed. Also, the analysis of mean structures permits the investigator to study invariance in both intercepts and factor loadings. Invariance in both sets of parameters is required under UCI. Millsap & Everson (1991) described how invariance hypotheses can be tested within a latent means model. These procedures were illustrated in Everson, Millsap, & Rodriguez (1991) in an examination of gender differences in the Test Anxiety Inventory (Spielberger, Gonzalez, Taylor, Anton, Algaze, Ross, & Westberry, 1980). A thorough investigation of measurement bias within the factor analytic model should include the analysis of latent mean structures in addition to covariance structures.

Another development that will affect the use of factor analysis in bias investigations is the greater

use of robust or ADF estimators (Bentler, 1983; Browne, 1984). Traditionally, the maximum likelihood estimation methods used in programs such as LISREL have required assumptions of multivariate normality for the latent variables or for the conditional distributions of observed latent variables (Jöreskog, 1967; Lawley & Maxwell, 1971). These assumptions are required if the SEs and test statistics produced under maximum likelihood procedures are to have useful interpretations. LISREL VII (Jöreskog & Sörbom, 1989) now includes options for using general weight matrices in the fit functions that implement the ADF estimation. A practical limitation in using these new estimation methods is that they are computationally unwieldy if the number of observed variables is moderate (25–33 variables). LISREL VII (Jöreskog & Sörbom, 1989) offers a second option that simplifies computation, but gives only large-sample approximations to the parameter sampling variances. These developments permit tests of UCI under very general distributional assumptions, giving greater flexibility in practical applications.

Discrete measures. Multidimensional models for discrete measures that are scored with a small or moderate number of categories also must be considered. Examples include Likert scaled items used in attitude or personality measurement, ability or achievement test items that are scored for partial credit, or testlets. Dichotomously scored items are included here as well. There are two broad types of multidimensional models for these items: factor analytic models and multidimensional IRT models.

Factor analytic models are now available for use in both dichotomous and ordered-categorical data. These models generally assume that a latent scalar random variable w^* underlies a given measured variable. Responses on the measured variable are determined by w^* in combination with a set of threshold parameters $t = (t_1, \ldots, t_{m-1})$, where m is the number of response categories. The latent scalar w^* is, in turn, given factor analytic representation in the model.

The full-information factor analysis model of Bock et al. (1988) is designed for dichotomously scored measures. A single threshold parameter is required for each measure. The threshold parameter becomes a difficulty parameter within a normal ogive model. The latent scalar w^* is given a common factor representation, with multiple common factors and a single unique factor. The measured response probability is a normal ogive function of the difference between the threshold and w^*. The TESTFACT program (Wilson et al., 1987) that implements this model is not designed for bias investigations, however. Multiple examinee groups cannot be analyzed simultaneously, and factor pattern elements cannot be fixed to nonzero values.

The model developed by Muthén (1984) applies generally to any ordered-categorical or dichotomous measure. In this model, the measured response is determined directly by the value of w^* in relation to the threshold values. The common factor representation for w^* includes parameters for both mean structures and covariance structures. This model is implemented within the LISCOMP program (Muthén, 1987). LISCOMP permits simultaneous analyses for multiple examinee groups. Parameter constraints that operate within or across groups may be imposed. The program provides χ^2 goodness-of-fit indexes and significance tests for individual parameters.

LISCOMP can be used to conduct likelihood-ratio tests of hypotheses involving parameter invariance over groups. A variety of parameter restrictions could be considered here. Invariance in item thresholds, factor loadings, and factor intercepts would be of immediate interest. Complete UCI in Equation 1 also would require invariance in the unique variances (Muthén & Lehman, 1985). Under multivariate normality assumptions for the latent scalars and factor scores, invariance in the common factor covariance structure is not required for UCI.

LISCOMP provides a highly flexible and useful tool for bias investigations in the multidimensional context. One present limitation of the program is that it is computationally efficient only for a moderate

number of observed measures. Muthén (1984) stated that the estimation method used in LISCOMP is practically useful for up to 15 to 20 variables. Another practical limitation is that large samples are needed to obtain useful estimates of SEs for testing fit. More research is needed to determine the required sample sizes and to study the behavior of the estimates in smaller samples.

Another method of analysis for ordered-categorical data would use the PRELIS program (Jöreskog & Sörbom, 1989) to estimate tetrachoric or polychoric correlations, followed by simultaneous multiple-group factor analysis in LISREL. In addition to estimating the needed correlation matrix, PRELIS provides estimates of a weight matrix used for weighted least squares estimation in LISREL. The maximum likelihood estimation option in LISREL does not give the correct SEs and test statistics for the discrete variable case. Weighted least squares, combined with the PRELIS weight matrix, will yield the correct large-sample SEs and test statistics (see Bollen, 1989, for a discussion).

LISREL permits a wide range of invariance hypotheses to be tested. The hypotheses are essentially identical to those tested in the continuous case. In the discrete case, these hypotheses refer to the factor structure for the latent response variables w^*, rather than the observed variables. Threshold parameters are not part of the LISREL model because the program works directly with the correlations among w^* as provided by PRELIS. Also, it is unclear whether factor means and intercepts may be modeled in LISREL using PRELIS input. There are no examples of such analyses in the LISREL manual, nor any published examples. As is true for LISCOMP, the test statistics provided by LISREL require large samples for accurate significance levels, but more research is needed for specific recommendations.

Multidimensional models have been proposed within IRT for use with discrete measures (Fraser, 1987; Hirsch & Miller, 1991; McKinley & Reckase, 1983; Reckase, 1985; Reckase & McKinley, 1991; Samejima, 1974; Sympson, 1978). These models have largely been developed for use with dichotomous items. The general ordered-categorical case appears not to have been studied intensively. Multidimensional models are not used widely in practical measurement work, and have found even less use in direct assessments of bias in IRT. No published examples of studies that investigated UCI or bias within a parametric multidimensional model were found.

A number of researchers have proposed conceptions of bias that are built on an assumption of a multidimensional latent space (Ackerman, 1992; Kok, 1988; Shealy & Stout, 1991). In these conceptions, test item performance is determined by a latent "target" trait that the test is intended to measure, and one or more additional latent "nuisance" traits. These nuisance traits may influence performance, but are irrelevant to the purpose of the test. For example, Kok (1988) posited a "test-wiseness" nuisance trait that denotes familiarity with "contextual clues" that may aid in solving the item. Another example might be "reading ability" as it influences math test performance in a story-problem item.

The existence of these nuisance traits is not sufficient to violate UCI however. Another requirement is that the groups being compared must differ in their distributions on these nuisance traits. To illustrate, let W_t be the target trait, and W_n be a possibly vector-valued nuisance trait. Then if bias is present, it must be true that

$$P(W_n \mid W_t, V) \neq P(W_n \mid W_t) \ . \tag{47}$$

Equation 47 says that there are group differences in the conditional distribution of W_n given W_t (Kok, 1988). Finally, the multidimensional formulation assumes that in the "complete" latent space $\mathbf{W} = \{W_t, W_n\}$, UCI in Equation 1 holds. Given these assumptions, Kok (1988) showed that UCI will not hold with respect to W_t alone, or that

$$P(Y \mid W_t, V) \neq P(Y \mid W_t) \ . \tag{48}$$

The implication is that a unidimensional IRT model should demonstrate that Y is biased, although in this situation the bias can be attributed to group differences in nuisance trait distributions.

Ackerman (1992) presented a detailed discussion of how bias might appear in a unidimensional latent trait analysis of an item set, even though UCI may hold in the complete latent space. Reckase (1985; Reckase & McKinley, 1991) proposed a multidimensional 2PLM for a single item Y in the complete latent space as

$$P(Y = 1 \mid \mathbf{a}', b, \mathbf{w}) = \frac{\exp(\mathbf{a}'\mathbf{w} + b)}{1 + \exp(\mathbf{a}'\mathbf{w} + b)}, \tag{49}$$

where $\mathbf{a}'$ is a vector of discrimination parameters, one for each latent dimension, and b is the item difficulty. Ackerman (1992) considered the special case in which there are two latent dimensions—the target dimension and the nuisance dimension. He then considered what might happen when a unidimensional 2PLM is fit to items with true multidimensional IRFs. Wang (1986) gave expressions for the unidimensional 2PLM item parameters and trait variable in terms of parameters from the multidimensional 2PLM. Ackerman (1992) demonstrated that given the inequality in Equation 47, UCI will be violated in the unidimensional model even if UCI holds in the multidimensional model. This means that group differences in the distribution of $\mathbf{W} = \{W_t, W_n\}$ may lead to bias within a unidimensional context. For example, group differences in means, variances, or covariances between W_t and W_n may result in bias as defined in relation to W_t. Oshima & Miller (1990) demonstrated that even if the only difference between the groups is in the correlation between the two latent dimensions, bias may appear in the unidimensional model and can be detected by area statistics.

This multidimensional perspective on bias is valuable in providing a coherent theoretical account of why bias may appear when unidimensional latent variable models are used. At least one important bias detection procedure has evolved from this perspective (Shealy & Stout, 1991), and is discussed below. It is important to recognize that although UCI may hold for a measure in the complete latent space, this latent space may be quite different from the latent dimension for which the measure is the intended indicator. This distinction between the "intended" latent dimension and the "complete latent space" lies at the heart of the problem of construct validity. It is, of course, possible that there are multiple "intended dimensions" for a given measure Y (e.g., Y as an achievement test score). If UCI does not hold for the measure under study in relation to its intended latent dimension (or multiple intended dimensions), then that measure is biased, for all practical purposes. It is incorrect to say that this bias is illusory simply because there exists an even larger latent space within which the measure may not be biased. Oshima & Miller (1990) appeared to suggest this when urging caution in the use of unidimensional bias detection procedures for measures that are "unbiased" in the multidimensional space.

Recent Developments

Some recent work has appeared that will influence future developments in bias detection that does not fit easily within the OCI/UCI categories. For example, the SIBTEST detection procedure proposed by Shealy & Stout (1991) crosses boundaries between OCI and UCI methods. Other work has expanded the scope of traditional IRT. Much of this work follows a trend toward nonparametric or semiparametric modeling of the IRF.

The SIBTEST Procedure

Shealy & Stout (1991) proposed a bias detection procedure that builds on the multidimensional

conception of bias described above. Their procedure is intended for use with dichotomous measures, and may be used to detect bias that is present simultaneously in a set of test items. The procedure begins by identifying a subset of the items that constitute the "valid subtest"; that is, a group of items that are believed to measure only the target trait. Hence UCI holds for these items in relation to the target trait. A total score Z is calculated for the valid subtest items, and the sample is stratified on Z. Within each stratum, "adjusted" means are calculated for the studied score Y in the reference and focal groups. If more than one item is being studied, Y is defined as the total score across items. Finally, a summary test statistic is computed as a weighted average of the differences between the reference and focal group adjusted means on Y, averaging across strata defined by Z. This statistic resembles the standardization index of Dorans & Kulick (1986). Under the null hypothesis of UCI for Y in relation to the target trait, the distribution of the test statistic is available. The null hypothesis can be tested using a z test. The test is sensitive to unidirectional bias in which, at all values of the target trait, the reference group is expected to score as well or better than the focal group.

Formally, this test procedure resembles OCI detection procedures in conditioning on an observed score Z, the valid subtest. The procedure departs from the usual OCI pattern in applying an adjustment to the mean of Y prior to comparing the groups on these means. This adjustment attempts to remove that portion of the group mean difference that is attributable to group differences in target trait distributions. Particularly in short tests, conditioning on Z may not fully control for these prior trait level differences because Z is imperfect as a measure of the target trait. In this way, the procedure resembles UCI detection methods without imposing a formal measurement model.

The Shealy/Stout detection procedure has been implemented in the computer program SIBTEST (Shealy, Stout, & Rossi, 1991). The current version of the program will accept up to 80 items and 3,000 examinees. Preliminary simulation evidence has shown that the SIBTEST procedure has Type I error frequencies that adhere closely to the nominal error rate set by the user. The evidence also indicates that the procedure has acceptable power for detecting unidirectional bias (Shealy & Stout, 1991b). The behavior of the procedure under other bias conditions has not yet been studied extensively.

Developments in IRT

Rosenbaum & Holland described ways of testing assumptions such as unidimensionality or local independence in IRT (Holland, 1981; Holland & Rosenbaum, 1986; Rosenbaum, 1984). Holland & Rosenbaum (1986) noted that traditional IRT models make three fundamental assumptions: unidimensionality of the latent trait, monotonicity of the IRF, and local independence. These assumptions have implications for the covariance structure among items that meet the assumptions. These implications are testable, even without specification of a parametric form for the IRF. Holland (1981) and Rosenbaum (1984) described some test procedures. Taken together, this work provides a nonparametric basis for deciding whether a set of items could be fit by any monotone, unidimensional, locally independent model.

Rosenbaum (1985, 1987a) discussed some nonparametric tests that can be used when two examinee populations are being studied. Rosenbaum (1985) considered the case in which UCI holds under a unidimensional model but the trait distributions are stochastically ordered. This condition can be shown to have testable consequences for the item response patterns. Certain order relations must hold between the populations for the expected values of functions of the item responses. Although Rosenbaum (1985) viewed violations of these order relations as evidence against ordering of the latent trait distributions, the violations also could be due to violations of UCI. In this sense, the order relations offer a nonparametric check on UCI under stochastic ordering. Rosenbaum (1987b) described ways of comparing IRFs of items that all follow a monotone, locally independent IRT model. He

defined the latent odds ratio and considered items with odds ratios that are proportional over values of the latent trait. Items following the Rasch model have proportional latent odds, but more general examples can be found as well. Given two examinee populations and two items that have proportional odds ratios in both populations, Rosenbaum (1987b) described a nonparametric test for equality of the relative difficulties of the two items. He noted that this provides a test of UCI for the items.

There also has been much work in the last decade that extends the boundaries of IRT by weakening assumptions of unidimensionality or local independence. Stout (1987, 1990) developed theory for test items that are "essentially unidimensional" in the context of an infinite item pool. These items are dominated by a single latent trait, with additional traits possibly influencing small numbers of items. This concept is probably more useful than strict unidimensionality as a model for real test data. Stout (1990) defined the related notion of "essential independence" as an alternative to strict local independence.

A set of items are essentially independent if the average covariance between pairs of items in the set, conditional on the latent trait, is close to 0. Essential independence and unidimensionality have testable consequences apart from any parametric specification of the IRF. Junker (1991) extended the essential independence concept to include polytomous items and showed that maximum likelihood estimates of the latent trait are consistent under essential independence. Jannerone (1987) presented a family of "conjunctive" item response models that exhibit local dependence, yet may have useful statistical properties (e.g., consistent estimators or sufficient statistics) similar to those of traditional models.

A different direction has been toward the elimination of fully parametric IRFs, with substitution of semiparametric functions. A recent example of this is provided by Ramsey & Winsberg (1991), who presented a modeling approach that uses monotone regression spline functions in place of the usual parametric functions. MMLE is used to estimate the spline coefficients. Preliminary simulation evidence indicated that the method can efficiently recover the shapes of known IRFs. This semiparametric approach, and others like it (Drasgow, Levine, Williams, McLaughlin, & Candell, 1989), permit wide flexibility in modeling test item responses. It is too early to tell whether the added flexibility also will give greater scope for bias detection.

Discussion

The past decade has produced important developments in statistical methods for bias detection. Both UCI and OCI methods have been actively pursued. Continued interest in methods for bias detection is likely, given the wide social concern for fairness and equal treatment. Although progress has been made in developing detection methods, problems still remain. This section focuses on some of these problems and considers the prospects for their solution.

OCI Methods

New developments in OCI methods for dichotomous measures have now replaced the earlier χ^2 and delta plot methods. The MH procedure is computationally efficient, works well in samples of moderate size, and can accurately detect bias under some realistic conditions. When all items follow the Rasch model, the MH procedure detects bias with adequate power. The procedure often fails to detect bias when the focal and reference response functions intersect—the "nonuniform" bias case. Nonuniform bias may occur when the 2PLM or the 3PLM hold for the studied item. In some cases, logistic regression or loglinear methods will have more power.

The MH procedure falsely indicates bias under some conditions. One such condition occurs when the studied item score Y is excluded from the total score Z, and Z is itself unbiased. Here Z and Y

will be conditionally independent given the latent variable W under standard local independence assumptions. The problems created by conditional independence affect all OCI methods and are not unique to the MH procedure (Meredith & Millsap, 1992). As noted by Holland & Thayer (1988), the studied item should be included in the sum that defines Z.

The standardization procedure is closely related to the MH procedure, and the two procedures give results that are very similar (Dorans & Holland, 1993). The problems that affect the MH procedure also should affect the standardization procedure.

All OCI methods have two additional problems that must be considered in practice. First, bias in the matching variable Z will reduce the accuracy of Z as a proxy for W. This problem has been known for some time (e.g., Van der Flier et al., 1984). It is not clear whether the presence of biased items in Z will produce spurious indications of bias in studied items or whether it will mask the bias in some items. The usual approach to this problem has been to iteratively purify Z by successively removing biased items. The adequacy of this strategy, however, has not been thoroughly investigated and warrants more attention.

The second problem concerns the adequacy of Z as a proxy for W in general, apart from the presence of bias in Z. OCI methods rely on Z to control for group differences in W. Recent theoretical work has shown that, in general, Z will be adequate as a proxy for W when Z is a sufficient statistic for W (Meredith & Millsap, 1992; Millsap & Meredith, 1992; Zwick, 1990). In the item bias case, if all items follow a Rasch model and Z is an unweighted total score including the studied item, Z is a sufficient statistic for W. The total score Z is not sufficient for W under multiparameter logistic models such as the 2PLM or the 3PLM. Under these models, OCI methods may falsely indicate bias when the focal and reference groups are stochastically ordered on W. This error is especially likely in short tests (e.g., 20 items or less). In long tests, "near" sufficiency may be achieved even under these models (Meredith & Millsap, 1992). The required test length is difficult to specify generally, but could be determined using simulations.

One approach to achieving sufficiency for Z in relation to W is to expand Z to include information on examinees that is external to the test, but is relevant to W. In this case, Z could be a vector-valued variable containing the total score on the test under study and other scores on tests that are alternative measures of W. As more information is added to Z, the risk of including measures that are themselves biased may increase. This expansion strategy requires that unbiased external measures be available, and this requirement may be difficult to fulfill in practice. Also, larger samples will generally be needed to accommodate matching with a multivariate Z.

Although OCI methods for dichotomous measures have received considerable attention, the ordered-categorical case is less well understood. Two practical difficulties contribute to this situation. First, the larger number of categories results in many potential values for Z in moderate or long tests. The contingency table that is produced after conditioning on Z will be sparse unless the sample is large. Score categories for Z may be combined, but at the possible cost of reducing the quality of Z as a proxy for W. The second problem is that bias can assume a variety of forms in the ordered-categorical case. Complete uniform bias in every item score category is likely to be the exception, rather than the rule. OCI methods that are sensitive to nonuniform bias are needed in this case.

Finally, the performance of OCI methods under multidimensionality in W has not been studied. Although OCI methods make no explicit assumptions concerning W, the use of a univariate Z carries a tacit assumption of unidimensionality in W. If W is multivariate, the univariate Z will ordinarily not be a sufficient statistic for W, and may be a poor proxy in general. It is surprising that this problem has received no attention.

UCI Methods

Three general UCI methods for detecting bias under unidimensional models for dichotomous measures are now available: area methods, Wald statistics, and *LR* tests. Prior to the use of any of these methods, the problem of parameter linkage must be solved. This problem is solved implicitly in the simultaneous estimation procedures that lead to *LR* tests, provided that appropriate anchor items are available. Iterative linking methods are preferred for area indexes and Wald statistics.

More research is needed to investigate the advantages of different linking methods. Once linkage is achieved, both Wald statistics and *LR* tests have firm statistical foundations for large samples. Not enough is known about their behavior in small or moderate samples. Simulation studies of this question would be useful. One disadvantage of these methods is that they provide no easily interpretable index of the size of the bias. Judgments of statistical significance are vulnerable to the effect of sample size.

Area measures can supply the needed indexes. Unweighted continuous area measures are computed easily. Although weighted measures should be useful in theory, the evidence to date has not demonstrated their advantage over unweighted indexes. The choice between bounded or unbounded measures remains unclear. At present, it appears that a useful UCI strategy would generate either *LR* or Wald tests, with area measures used as indexes of the size of any bias found.

UCI methods for bias detection in ordered-categorical data are still in the early stages of development. Factor-analytic methods have progressed rapidly in the last decade. Both LISCOMP (Muthén, 1987) and PRELIS/LISREL (Jöreskog & Sörbom, 1989) offer flexible systems for bias investigations using *LR* procedures. IRT models exist for ordered-categorical data, but these models have not yet been used extensively for bias investigations. In such applications, Wald statistics or *LR* tests are likely to be used. More research on the development of area measures or other effect size indicators is needed for the ordered-categorical case. The accuracy of bias detection using these models in practice has not been investigated thoroughly. This is a fertile area for new research.

The sample size requirement for UCI methods continues to be a practical problem. Developments in the past decade in MMLE and Bayesian estimation methods have improved the situation to some degree for IRT models (Mislevy & Stocking, 1987). Sample sizes that are adequate for estimation within a single sample may be inadequate for between-sample statistical comparisons, however. Parameter estimation within models for ordered-categorical data will usually require larger samples than the dichotomous case. With few exceptions, bias investigation with UCI methods remains a large-sample enterprise.

A general problem affecting all UCI methods is that of selecting the correct measurement model for the relationship of Y to W. Nearly all current UCI methods operate within parametric measurement models. Meaningful group comparisons of parameters or response functions require that adequate models are available for the groups under study. If the proposed model is inadequate, tests of UCI are likely to be confounded with tests of fit. Ideally, tests of UCI should be conducted under minimal assumptions about the form of the function relating Y to W. The trend over the last decade toward semiparametric or nonparametric modeling is encouraging and is likely to affect future developments in bias detection.

Additional Issues

One aspect of model choice that is especially important in bias studies is the dimensionality of W. Apart from the factor analytic applications, nearly all UCI investigations to date have been conducted with unidimensional models. Strict unidimensionality is likely to be violated in real data.

Additional dimensions in *W* are possible sources of bias. Recent work has incorporated this idea, distinguishing between the intended or "target" latent variable *W*, and additional "nuisance" latent dimensions whose measurement is not the intended purpose for *Y*. It is important to identify these "nuisance" dimensions to help understand the sources of bias.

There is some confusion in the IRT literature concerning the so-called "invariance" properties of item parameters in item response models. Parameter invariance or UCI should be regarded as an empirical question to be investigated, rather than a mathematical property that can be assumed to hold generally across examinee populations. In the absence of any data on test performance in these multiple populations, there is little that can be said about the invariance properties of parameters in these models. The invariance properties only appear after the fit of the model has been evaluated and UCI has been tested directly.

Most measurement researchers regard measurement bias as an important practical and ethical problem. It should be more widely recognized that measurement bias is an important scientific problem as well. Studies of measurement bias provide empirical tests of construct interpretations. The existence of bias in a given measure indicates that the constructs being measured are not fully understood. Hence, studies of measurement bias should be encouraged as part of the general process of construct validation, if for no other reason. The development of adequate methods for bias detection is, and will continue to be, an important scientific challenge.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67–91.

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Alderman, D. L., & Holland, P. W. (1981). *Item performance across native language groups on the Test of English as a Foreign Language* (ETS Research Rep. No. 81-16). Princeton NJ: Educational Testing Service.

Amemiya, T. (1985). *Advanced econometrics.* Cambridge MA: Harvard University Press.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10,* 95–106.

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

Bentler, P. M. (1983). Some contributions to efficient statistics in structural models: Specification and estimation of moment structures. *Psychometrika, 48,* 493–517.

Berk, R. A. (1982). *Handbook of methods for detecting test bias.* Baltimore MD: Johns Hopkins University Press.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge MA: MIT Press.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Bock, R. D. (1975). *Multivariate statistical methods.* New York: McGraw-Hill.

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

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

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Browne, M. W. (1984). The decomposition of multitrait-multimethod matrices. *British Journal of Mathematical and Statistical Psychology, 33,* 184–199.

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *American Statistician, 36,* 153–157.

Byrne, B. M. (1989). *A primer of LISREL.* New York: Springer-Verlag.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466.

Camilli, G. (1979). *A critique of the chi square method for assessing item bias.* Unpublished manuscript, University of Colorado, Laboratory of Educational Research, Boulder.

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.

Cardall, C., & Coffman, W. E. (1964). *A method for comparing the performance of different groups on the items in a test* (College Entrance Examination Board Research and Development Rep. 64-5 No. 9; ETS Research Bulletin 64-61). Princeton NJ: Educational Testing Service.

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

Cressie, N., & Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika, 48,* 129–141.

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

Donoghue, J. R., & Allen, N. L. (1991, April). *"Thin" versus "thick" matching in the Mantel-Haenszel procedure for detecting DIF.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

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

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 3,* 217–233.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale NJ: Erlbaum.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December, 1977: An application of the standardization approach* (ETS Research Rep. No. RR-83-9). Princeton NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355–368.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29,* 309–319.

Draba, R. E. (1978). *The Rasch model and legal criteria of "reasonable" classification.* Unpublished doctoral dissertation, University of Chicago.

Drasgow, F. (1982). Biased test items and differential validity. *Psychological Bulletin, 92,* 526–531.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19–29.

Drasgow, F., Levine, M. V., Williams, B., McLaughlin, M. E., & Candell, G. L. (1989). Modeling incorrect responses to multiple-choice items with multilinear formula score theory. *Applied Psychological Measurement, 13,* 285–299.

Everson, H. T., Millsap, R. E., & Rodriguez, C. M. (1991). Isolating gender differences in test anxiety: A confirmatory factor analysis of the test anxiety inventory. *Educational and Psychological Measurement, 51,* 243–251.

Ferrara, S., & Walker-Bartnick, L. (1990, April). *Detecting and analyzing differential item functioning in an essay test using the partial credit model.* Paper presented at the meeting of the National Council on Measurement in Education, Boston MA.

Fienberg, S. E. (1980). *The analysis of cross-classified categorical data.* Cambridge MA: MIT Press.

Fraser, C. (1987). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, New South Wales, Australia: The University of New England, Center for Behavioral Studies.

Harvey, A. L. (1990, April). *The stability of the Mantel-Haenszel d-DIF statistic across populations differing in ability.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston MA.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139–164.

Hirsch, T. M., & Miller, T. R. (1991, June). *Evaluation of a multidimensional item response theory procedure for investigating test dimensionality.* Paper presented at the meeting of the Psychometric Society, New Brunswick NJ.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46,* 79–92.

Holland, P. W. (1985). On the study of differential item performance without IRT. *Proceedings of the 27th Annual Conference of the Military Testing Association* (Vol. 1; pp. 282–287). San Diego CA: Navy Personnel Research and Development Center.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14,* 1523–1543.

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. 129–145). Hillsdale NJ: Erlbaum.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology, 67,* 818–825.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood IL: Dow Jones-Irwin.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 117–160). Baltimore MD: Johns Hopkins University Press.

Ironson, G. H., Homan, S., Willis, R., & Signer, B. (1984). The validity of item bias techniques with math word problems. *Applied Psychological Measurement, 8,* 391–396.

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

Jannerone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51,* 357–373.

Jöreskog, K. J. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32,* 443–482.

Jöreskog, K. J. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409–426.

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

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika, 56,* 255–278.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika, 54,* 681–697.

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

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

Kok, F. (1988). Item bias and test dimensionality. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 263–274) New York: Plenum.

Kok, F. G., Mellenbergh, G. H., & Van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement, 22,* 295–303.

Lautenschlager, G. J., & Park, D. G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement, 12,* 365–376.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method.* London: Butterworth.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18,* 109–118.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrup, J. L. (1981). An investigation of 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.

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

Lord, F. M., & Pashley, P. J. (1988). *Confidence bands for the three-parameter logistic item response curve* (ETS Research Rep. No. 88-67). Princeton NJ: Educational Testing Service.

Loyd, B. H. (1984, April). *Evaluation of log-linear models for detection of item bias: A comparison across samples.* Paper presented at the meeting of the American Educational Research Association, New Orleans LA.

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

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58,* 690–700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719–748.

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

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139–160.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49,* 529–544.

Mazur, K. M., Clauser, B. E., & Hambleton, R. K. (1991). *The effect of sample size on the functioning of the Mantel-Haenszel statistic* (Rep. No. 211). Amherst: University of Massachusetts, Laboratory of Psychometric and Evaluation Research.

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.

McKinley, R. L., & Reckase, M. D. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research Rep. No. ONR 83-2). Iowa City IA: The American College Testing Program.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement, 11,* 161–173.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105–118.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127–143.

Meredith, W. (1990, October). *Factorial invariance from a measurement invariance perspective.* Paper presented at the meeting of the Society for Multivariate Experimental Psychology, Newport RI.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57,* 289–311.

Miller, T. R., Spray, J. A., & Wilson, A. (1992, July). *A comparison of three methods for identifying nonuniform DIF in polytomously scored test items.* Paper presented at the Psychometric Society Meeting, Columbus OH.

Millsap, R. E., & Everson, H. T. (1991). Confirmatory measurement model comparisons using latent means. *Multivariate Behavioral Research, 26,* 479–497.

Millsap, R. E., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement, 16,* 389–402.

Mislevy, R. J., & Bock, R. D. (1984). *BILOG: Item analysis and test scoring with binary logistic models.* Mooresville IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1987). *A consumer's guide to LOGIST and BILOG* (Research Rep. No. 87-43). Princeton NJ: Educational Testing Service.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered-categorical and continuous latent variable indicators. *Psychometrika, 49,* 115–132.

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

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

Oshima, T. C., & Miller, M. D. (1990). Multidimensionality and IRT-based item invariance indexes: The effect of between-group variation in trait correlation. *Journal of Educational Measurement, 27,* 273–283.

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

Park, D. G. (1988). *Investigations of item response theory item bias detection.* Unpublished doctoral dissertation, University of Georgia, Athens.

Park, D. G., & Lautenschlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement, 14,* 163–173.

Pashley, P. J. (1992). *Graphical IRT-based DIF analyses* (Research Rep. No. 92-55). Princeton NJ: Educational Testing Service.

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.

Ramsey, J. O., & Winsberg, S. (1991). Maximum marginal likelihood estimation for semiparametric item analysis. *Psychometrika, 56,* 365–379.

Rao, C. R. (1973). *Linear statistical inference and its applications.* New York: Wiley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401–412.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361–373.

Rock, D. A., Werts, C. E., & Flaugher, R. L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research, 13,* 403–418.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49,* 425–436.

Rosenbaum, P. R. (1985). Comparing distributions of item responses for two groups. *British Journal of Mathematical and Statistical Psychology, 38,* 206–215.

Rosenbaum, P. R. (1987a). Probability inequalities for latent scales. *British Journal of Mathematical and Statistical Psychology, 40,* 157–168.

Rosenbaum, P. R. (1987b). Comparing item characteristic curves. *Psychometrika, 52,* 217–233.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53,* 349–359.

Rudner, L. M. (1977). *An evaluation of select approaches for biased item identification.* Unpublished doctoral dissertation, Catholic University of America, Washington DC.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980a). Biased item detection techniques. *Journal of Educational Statistics, 5,* 213–233.

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

Ryan, K. E. (1991). The performance of the Mantel-Haenszel procedure across samples and matching criteria. *Journal of Educational Measurement, 28,* 325–337.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 34.

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

Scheuneman, J. D. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16,* 143–152.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education, 2,* 255–275.

Segall, D. O. (1983). *Test characteristic curves, item bias, and transformation to a common metric in item response theory: A methodological artifact with serious consequences and a simple solution.* Unpublished manuscript, University of Illinois, Department of Psychology, Urbana-Champaign.

Shealy, R., & Stout, W. (1991). *A procedure to detect test bias present simultaneously in several items* (Office of Naval Research Rep. No. 4421-548). Urbana-Champaign: University of Illinois, Department of Applied Statistics.

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

Shealy, R., Stout, W., & Rossi, L. (1991). *SIBTEST manual.* Urbana-Champaign: University of Illinois, Department of Applied Statistics.

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

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

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

Simpson, E. H. (1951). The interpretation of interaction contingency tables. *Journal of the Royal Statistical Society, 13,* (Series B), 238–241.

Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. *The American Statistician, 39,* 33–38.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27,* 229–239.

Spielberger, C. D., Gonzalez, H. P., Taylor, C. J., Anton, W. D., Algaze, B., Ross, G. R., & Westberry, L. G. (1980). *Preliminary professional manual for the Test Anxiety Inventory.* Palo Alto CA: Consulting Psychologists Press.

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

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589–617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to multidimensionality assessment and ability estimation. *Psychometrika, 55,* 293–325.

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

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

Swaminathan, H., & Rogers, H. J. (1990b, April). *A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning.* Paper presented at the meeting of the American Educational Research Association, Boston MA.

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. (1990). *MULTILOG user's guide* (Version 6.0). Mooresville IN: Scientific Software.

Thissen, D., & Steinberg, L. (1986). Taxonomy of item response models. *Psychometrika, 51,* 567–577.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item

bias. *Psychological Bulletin, 99,* 118–128.
Thissen, D., Steinberg, L., & Mooney, J. (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.
Thissen, D., & Wainer, H. (1990). Confidence envelopes for item response theory. *Journal of Educational Statistics, 15,* 113–128.
Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago: University of Chicago Press.
Uttaro, T. (1992). *Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning.* Unpublished doctoral dissertation, Graduate Center, City University of New York.
Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10,* 333-344.
Van der Flier, H., Mellenbergh, G. J., Ader, H. J., & Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement, 21,* 131–145.
Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–201.
Wainer, H., Sireci, S. G., & Thissen, D. (1991). *Differential testlet functioning definitions and detection* (Research Rep. No. 91-21). Princeton NJ: Educational Testing Service.
Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society, 54,* 426–482.
Wang, M. (1986, April). *Fitting a unidimensional model to multidimensional item response data.* Paper presented at the meeting of the Office of Naval Research Contractors, Gatlinburg TN.
Warm, T. A. (1978). *A primer of item response theory* (Technical Rep. No. 941078). Washington DC: U.S. Coast Guard Institute.
Wilder, G. Z., & Powell, K. (1989). *Sex differences in test performance: A survey of the literature* (College Board Rep. No. 89-3). New York: The College Board.
Wilson, D., Wood, R., & Gibbons, R. D. (1987). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Mooresville IN: Scientific Software.
Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide.* Princeton NJ: Educational Testing Service.
Wright, B. D., Mead, R., & Draba, R. (1976). *Detecting and correcting item bias with a logistic response model* (Research Memorandum No. 22). Chicago: University of Chicago, Department of Education, Statistical Laboratory.
Wright, D. J. (1987). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. In A. P. Schmitt & N.J. Dorans (Eds.), *Differential item functioning on the scholastic aptitude test* (ETS RM-87-1). Princeton NJ: Educational Testing Service.
Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185–197.
Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessing differential item functioning in performance tasks. *Journal of Educational Measurement, 30,* 233–251.

Acknowledgments

Preparation of this article was supported in part by a grant from the City University of New York's Professional Staff Congress, the CUNY Research Award Program, to Roger E. Millsap, and in part by a Postdoctoral Fellowship from the Educational Testing Service to Howard T. Everson.

Author's Address

Send requests for reprints or further information to Roger E. Millsap, Department of Psychology, Baruch College, City University of New York, 17 Lexington Avenue, New York NY 10010, U.S.A. Internet: sapbb@cunyvm.cuny.edu.