Inferential Conditions in the Statistical Detection of Measurement Bias

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

William Meredith, University of California, Berkeley

Measurement bias in an observed variable Y as a measure of an unobserved variable W exists when the relationship of Y to W varies among populations of interest. Bias is often studied by examining population differences in the relationship of Y to a second observed measure Z that serves as a substitute for W. Whether the results of such studies have implications for measurement bias is addressed by first defining two forms of invariance—one corresponding to the relationship of Y to the unmeasured W, and one corresponding to the relationship of Y to the observed Z. General theoretical conditions are provided that justify the inference of one form of invariance from the other. The implications of these conditions for bias detection in two broad areas of application are discussed: differential item functioning and predictive bias in employment and educational settings. It is concluded that the conditions for inference are restrictive, and that bias investigations that rely strictly on observed measures are not, in general, diagnostic of measurement bias or the lack of bias. Some alternative approaches to bias detection are discussed. *Index terms: differential item functioning, invariance, item bias, item response theory, measurement bias, predictive bias.*

A commonly asked question concerns the type of evidence that is necessary to identify bias in a measurement process. Statistical methods for the detection of measurement bias have been under development for several decades. Applications of these methods include the detection of item bias, also called differential item functioning (DIF; Berk, 1982; Holland & Thayer, 1988; Ironson, 1982; Marascuilo & Slaughter, 1981; Scheuneman, 1979; Shepard, Camilli, & Averill, 1981); the detection of predictive bias in educational and employment settings (Cleary, 1968; Humphreys, 1986; Linn, 1984; Linn & Werts, 1971; Reilly, 1986); and studies of salary equity (Birnbaum, 1979; Conway & Roberts, 1983; Dempster, 1988; Goldberger, 1984; McFatter, 1982, 1987; Peterson, 1986).

A number of researchers have noted that common features exist in the statistical methods used in these diverse applications (Humphreys, 1986; Linn, 1984; Raju & Normand, 1985). The measurement bias problem is discussed below within a framework that encompasses many detection applications. The DIF and predictive bias applications serve here as illustrations; the same principles can be applied to the salary equity application as well. The theoretical conditions that justify inferences of bias (or lack of bias), given the results of commonly-used detection procedures, are provided. The conditions presented are very general and do not require the ancillary statistical assumptions often invoked in discussions of bias detection. For example, the conditions do not require the additivity, linearity, or homoscedasticity assumptions often used in predictive bias applications. The intent was to provide a more general analysis of the measurement bias problem than has been available previously, and to draw some implications for common approaches to bias detection. A central implication is that detection methods that rely exclusively on observed measures are not generally diagnostic of measurement bias, or the lack of bias.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 4, December 1992, pp. 389-402
© Copyright 1992 Applied Psychological Measurement Inc.
0146-6216/92/040389-14$1.95

Measurement Invariance

Let Y be an observable random variable that is intended to be a measure of, or systematically related to, an unobserved or latent random variable W. All variables are considered discrete, finite, and possibly multivariate. Any continuous distribution can be arbitrarily well-approximated by a discrete distribution, and extensions of the results to the continuous case are generally straightforward. Also, let multiple populations of persons be defined by values of an observable random variable V. Ordinarily, V includes demographic variables such as gender, ethnicity, or age. Broadly speaking, measurement bias in Y as a measure of W can be said to exist when the relationship of Y to W varies among populations defined by V. The unobserved conditional invariance (UCI) for Y with respect to W and V is defined as holding if and only if

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

for all realizations w and v, and $P(\circ)$ denotes probability. When Equation 1 holds, Y and V are statistically independent at all fixed values of W. When Equation 1 does not hold, Y can be said to be biased in relation to W among populations defined by V.

This definition of invariance contains other forms of invariance that have appeared in the literature. Lord's (1980) definition of item bias—in which Y is a dichotomous item score variable and W is a latent trait—corresponds to an inequality in Equation 1. Mellenbergh's (1989) definition of lack of bias in a test item is nearly identical to Equation 1. Factorial invariance (Meredith, 1990) is a weaker form of UCI in which the regression of Y on W is linear, and the conditional covariance matrix for Y is diagonal.

UCI is difficult to evaluate directly because W is unmeasured. Bias detection methods often address this problem by using a second observable random variable Z that is also a measure of, or systematically related to, W. Population differences in the relationship of Y to Z are considered indications of bias in Y (or Z).

Observed conditional invariance (OCI) of Y with respect to Z and V is observed as holding if and only if

$$P(Y|Z = z, V = v) = P(Y|Z = z) , \tag{2}$$

for all realizations z and v. An important limitation in bias detection methods that evaluate Equation 2 is that OCI need not imply, or be implied by, UCI of Y in Equation 1. Hence, although these methods may provide adequate tests of OCI, they may fail to reveal UCI or bias in Y. Two applications illustrate this point.

DIF Detection

In the detection of DIF, Y is a test item score variable, and W is a latent variable that is believed to influence performance on the test item. If W is viewed as a latent trait within an item response theory (IRT) perspective, UCI of Y is equivalent to invariance of the item response function over populations defined by V (Lord, 1980). IRT methods for assessing DIF evaluate UCI directly under various parametric models for the item response function (Thissen, Steinberg, & Wainer, 1988). Other bias detection methods use the total test score Z as a substitute for W. The traditional χ^2 approaches (Ironson, 1982; Marascuilo & Slaughter, 1981; Scheuneman, 1979; Shepard et al., 1981), the Mantel-Haenszel χ^2 method (Holland & Thayer, 1988; Mantel & Haenszel, 1959), standardization approaches (Dorans & Kulick, 1986), and logistic regression methods (Swaminathan & Rogers, 1990) each evaluate OCI in Equation 2. Various authors have questioned whether DIF detection methods based on IRT and those using the observed total score will lead to identical decisions concerning DIF (Ironson, 1982). Zwick (1990) and Meredith & Millsap (1992) presented results showing that UCI of Y and OCI need not be equivalent in DIF detection.

Predictive Bias

In predictive bias, Y is an observable criterion variable, W is an unobservable (possibly multivariate) ''ability'' or ''aptitude'' variable believed to influence Y, and Z is an observable predictor variable, usually a test score. In practice, interest often centers on whether Z is biased as a predictor of Y, given assumptions such as linearity and homoscedasticity in the regression of Y on Z. The Cleary (1968) definition of fairness views Z as free of bias if the linear regression of Y on Z is identical among populations defined by V. OCI in Equation 2 implies fairness under the Cleary definition. Conversely, the Cleary definition implies OCI under the assumptions typically made in predictive bias applications (i.e., Y is conditionally normal given Z, with the regression of Y on Z being additive, linear, and homoscedastic with respect to Z and V). Hence, methods that evaluate bias under the Cleary definition can be viewed as tests of OCI in Equation 2. The difficulties encountered in assessing predictive bias using the regression function of Y on Z have been noted by several authors (Humphreys, 1986; Linn, 1984; Linn & Werts, 1971; Reilly, 1973). A source of difficulty is that OCI is not equivalent to UCI of Y or Z. UCI of Z can be defined by replacing Y with Z in Equation 1. Even assuming UCI for Y, UCI of Z does not require OCI, and OCI may hold when Z is biased as a measure of W.

These examples suggest that it would be useful to know the conditions under which OCI in Equation 2 is equivalent to UCI of Y (or Z), as in Equation 1. The theoretical conditions that justify inferences from empirical tests of OCI of Y are discussed below.

Equivalence Conditions

The following developments assume that no pair of variables among Y, Z, W, and V is marginally independent—that none of the joint bivariate distributions can be factored into a univariate distribution. This assumption avoids some trivial cases (e.g., W and V are independent). Conditional independence is permitted between two variables after conditioning on a third variable. Parametric assumptions about distributional forms, linearity of regressions, homoscedasticity, or homogeneity of variance are not needed in the conditions to be presented. Proofs of inferences based on the conditions are provided in the Appendix.

Condition I

In this condition, the joint distribution of (Z, W) is degenerate because a one-to-one correspondence exists between the values of Z and the values of W. Each value of Z corresponds to only one value of W. Under this condition, $P(Z|W = w) = 0$ or 1 depending on whether Z assumes the value $Z(w)$ that corresponds to $W = w$. If W is considered the classical univariate true score random variable in relation to the univariate observed score Z, then Z is perfectly reliable as a measure of W under this condition. This condition implies UCI for Z because Z depends on W alone, regardless of the value of V.

In DIF applications, Z is usually an unweighted sum of item scores, and W is a univariate latent trait. Under weak assumptions, asymptotically as items that measure W are added to Z, the joint distribution of (Z, W) becomes degenerate in the above sense. Hence, tests of sufficient length should yield scores on Z that asymptotically fulfill Condition I, assuming that the items in Z measure W. The required test length may exceed practical limits and is difficult to specify. Also, Condition I will fail to hold after scores on Z are grouped, which is common in practice (Ironson, 1982). Finally, Condition I is highly restrictive if either Z or W is multivariate. In the multivariate case, a single value of Z may be consistent with several values of W, regardless of the reliability of Z. From this perspective, Condition I is stronger than the condition of "perfect reliability."

The predictive bias problem is trivial under Condition I because Z can be taken as UCI without further evaluation. OCI will or will not hold depending on whether UCI holds for Y. A tacit assumption in most investigations of predictive bias is that UCI holds for Y, implying that OCI holds under Condition I. Frequently, W may be regarded as multivariate, that is, multiple dimensions of ability or aptitude underlie test and/or criterion performance. Condition I may not hold in such cases even when Z is perfectly reliable.

Conditions IIA and IIB

For Conditions IIA and IIB, there are two types of inferences regarding UCI and OCI. Given UCI of Y, Condition IIA justifies the inference that OCI in Equation 2 holds. Conversely, given OCI in Equation 2, Condition IIB justifies the inference of UCI for Y. The two conditions fulfill different roles in practical applications. For example, suppose that the data support OCI. Then, Condition IIB allows UCI of Y to be inferred. If the data suggest that OCI does not hold, then Condition IIA implies that UCI does not hold. Conditions IIA and IIB may both be true. In this case, UCI for Y holds if and only if OCI holds.

Condition IIA holds when the following two equations are true:

$$P(Y \mid Z = z,\ W = w) = P(Y \mid Z = z) \tag{3}$$

and

$$P(Z \mid Y = y,\ W = w,\ V = v) = P(Z \mid Y = y,\ W = w)\ , \tag{4}$$

for all z, w, y, and v. If Equations 1, 3, and 4 are true, then OCI in Equation 2 holds as well. Equation 3 can be denoted "Bayes sufficiency" of Z for W (Lehmann, 1986). When Equation 3 is true, Z behaves much like a sufficient statistic in relation to W: All the information in W that is relevant to Y is captured by Z. Condition IIB holds when the following two equations are true:

$$P(Y \mid Z = z,\ W = w,\ V = v) = P(Y \mid Z = z,\ V = v) \tag{5}$$

and

$$P(Z \mid W = w,\ V = v) = P(Z \mid W = w)\ , \tag{6}$$

for all z, w, and v. If Equations 2, 5, and 6 are true, then UCI of Y also holds. Equation 5 can be viewed as a "within-group" version of Equation 3: Z is Bayes sufficient for W within groups defined by V. Equation 6 indicates that UCI holds for Z, or that Z is unbiased. Note that Equations 4 and 6 are not equivalent. Condition I implies Equations 3, 4, 5, and 6, but the converse is not true. For example, Conditions IIA and IIB do not require Z to be perfectly reliable as a measure of W.

In DIF applications, Bayes sufficiency in Equations 3 and 5 holds when Z is an unweighted sum of item scores that includes Y and all items fit a Rasch model (Lord & Novick, 1968). Holland & Thayer (1988) proved that under certain conditions, the Mantel-Haenszel test of OCI is diagnostic for UCI of Y. The required conditions can be shown to be equivalent to Equations 3, 5, and 4 or Equations 3, 5, and 6. The presence of biased items in Z (other than Y) will generally lead to violations of both Equations 4 and 6. It is still possible for OCI to hold in such cases even if UCI of Y fails. The above choice for Z is unusual in that Z includes Y, and Equation 6 will ordinarily fail when Y is biased. If Y is the only biased item in Z, Equation 4 will hold. Condition IIA can be used in this case to infer failure of UCI from failure of OCI. If Y is omitted in the calculation of Z, Equations 3 and 5 both fail. Local independence holds for Y and Z in this case under standard IRT assumptions (Lord & Novick, 1968). Also, more complex IRT models, such as two- or three-parameter logistic models, generally violate Equations 3 and 5 when Z is an unweighted total test score. Under this choice for Z, violations of the Rasch model assumptions generally lead to violations of Conditions IIA and IIB.

Conditions IIA and IIB will not always be useful in establishing bias, or lack of bias, for Z in predictive bias applications. Suppose that UCI holds for Y and that Bayes sufficiency also holds. Then, under Condition IIA, Equation 4 implies OCI. Therefore, failure of OCI implies failure of Equation 4. Unfortunately, failure of Equation 4 need not imply bias or failure of UCI for Z. Unusual examples can be constructed under which Equation 4 fails, yet UCI holds for Z. Thus, failure of OCI need not imply that Z is biased under Condition IIA. Condition IIB says that given Equation 5 and OCI, UCI in Z implies UCI in Y. In this situation, if Y is known to be biased, it can be concluded that Z also is biased. In practical applications, however, Y is ordinarily assumed to be unbiased; therefore, Condition IIB does not help in establishing either UCI or bias in Z. Given Equations 3 or 5, OCI may hold even though both Y and Z are biased. If neither Equation 3 nor Equation 5 holds, UCI may hold for both Y and Z even though OCI fails. Bayes sufficiency requires that Z capture all aspects of W that are relevant to criterion performance. Discussions of predictive bias often regard W as a common factor, or set of factors, underlying Y and Z (Humphreys, 1986; Linn, 1984). Both Equations 3 and 5 are generally violated under this common factor interpretation.

Conditions IIIA and IIIB

Given UCI of Y, Condition IIIA justifies the inference of OCI. Conversely, given OCI, Condition IIIB justifies the inference of UCI of Y. Hence, Conditions IIIA and IIIB have roles that are analogous to Conditions IIA and IIB, respectively. Given both Conditions IIIA and IIIB, UCI of Y holds if and only if OCI holds.

Condition IIIA holds when the following two equations are true:

$$P(Y|Z = z, W = w, V = \mathrm{v}) = P(Y|W = w, V = v) \tag{7}$$

and

$$P(W|Z = z, V = v) = P(W|Z = z) \ , \tag{8}$$

for all z, w, and v. If Equations 1, 7, and 8 are true, OCI in Equation 2 can be inferred. Condition IIIB holds when the following two equations are true:

$$P(Y|Z = z, W = w) = P(Y|W = w) \tag{9}$$

and

$$P(W|Y = y, Z = z, V = v) = P(W|Y = y, Z = z) \ , \tag{10}$$

for all z, y, w, and v. If Equations 2, 9, and 10 are true, then UCI of Y also must hold. Equation 9 says that Y and Z are conditionally independent given W. In IRT, this property is denoted "local independence" of Y and Z given W. Equation 7 says that local independence holds within groups defined by V. Equations 8 and 10 concern conditional independence between W and V given Z, or given both Z and Y. As shown in the Appendix, Equation 8 implies that UCI fails for Z, or that Z is biased. Similarly, local independence in Equation 9 contradicts Bayes sufficiency in Equation 3. More generally, given UCI of Y, Conditions IIA and IIIA cannot both be true. Given OCI in

Equation 2, Conditions IIB and IIIB cannot both be true.

Local independence in Equations 7 and 9 both hold under nearly all IRT models, provided that the total test score Z is calculated excluding Y. But neither Equation 8 nor Equation 10 will generally hold under any of these models. Conditions IIIA and IIIB do not appear to be realistic equivalence conditions for the DIF application.

Similarly, Conditions IIIA and IIIB do not provide tests for UCI of Z in predictive bias applications. Equation 8 implies that Z is biased, but failure of Equation 8 does not imply UCI for Z. Condition IIIA establishes that when UCI exists for Y, OCI may hold even though Z is biased. Hence, Z would be erroneously declared free of bias in this situation. Local independence in Equations 7 and 9 could hold under a common factor interpretation of W, given additional distributional assumptions (e.g., multivariate normality). The common factor model is generally inconsistent with both Equations 8 and 10, however.

Reverse Regression

In predictive bias applications, researchers have considered reversing the roles of Y and Z in Equation 2 and then exploring the implications of OCI in this case (Birnbaum, 1979; Cole, 1973; Darlington, 1971; Linn, 1984; Wainer & Steinberg, 1991). For example, Wainer & Steinberg demonstrated gender differences in SAT mathematics scores among nearly 12,000 examinees after matching examinees on performance grades within specific college math courses. In this example, Y is defined as the grade in a given math course, and Z is an SAT mathematics score. Reverse OCI can be defined as holding if and only if

$$P(Z \mid Y = y, V = v) = P(Z \mid Y = y) \ , \tag{11}$$

for all y and v. It is easily established that Equations 2 and 11 cannot both be true unless Z and Y are marginally independent. In the above example, reverse OCI would imply that within groups matched on course grades, no gender differences would be found in the distribution of SAT scores. The question of the implications for UCI of Z or Y when reverse OCI holds or fails to hold can be addressed by restating the foregoing conditions—that is, reversing the roles of Y and Z. After reversal, the conditions describe situations in which UCI of Z can be inferred from reverse OCI, or the converse.

First, reverse OCI and UCI of Z are equivalent when there is a one-to-one correspondence between the values of Y and the values of W. This condition will ordinarily be unrealistic in predictive bias applications, for reasons described earlier. Reversed Condition IIA holds when

$$P(Z \mid Y = y, W = w) = P(Z \mid Y = y) \tag{12}$$

and

$$P(Y \mid Z = z, W = w, V = \mathrm{v}) = P(Y \mid Z = z, W = w) \ , \tag{13}$$

for all y, w, and v. Bayes sufficiency of Y holds in Equation 12. Reverse OCI can be inferred, given UCI of Z, Equation 12, and Equation 13. Reversed Condition IIB holds when

$$P(Z \mid Y = y, W = w, V = v) = P(Z \mid Y = y, V = v) \tag{14}$$

and

$$P(Y \mid W = w, V = v) = P(Y \mid W = w) \ , \tag{15}$$

for all y, w, and v. UCI of Y holds in Equation 15, and ''within-group'' Bayes sufficiency of Y holds in Equation 14. UCI of Z can be inferred, given reverse OCI, Equation 14, and Equation 15. Bayes sufficiency of Y requires that Y capture all the information in W that is relevant to Z. In predictive bias applications, Y is often tacitly assumed to be unbiased, fulfilling Equation 15 and possibly Equation 13. If Bayes sufficiency in Equations 12 and 14 also holds, assessment of reverse OCI could be diagnostic for bias in Z. Bayes sufficiency generally fails if W is viewed as a common factor, or set of factors, underlying Y and Z.

Reversed Condition IIIA holds when

$$P(Z \mid Y = y, W = w, V = v) = P(Z \mid W = w, V = v) \tag{16}$$

and

$$P(W \mid Y = y, V = v) = P(W \mid Y = y) \ , \tag{17}$$

for all y, w, and v. Equations 16 and 7 are equivalent. Reverse OCI must hold, given UCI of Z, Equation 16, and Equation 17. Reversed Condition IIIB is equivalent to Condition IIIB in Equations 9 and 10. Given reverse OCI, Equation 9, and Equation 10, UCI of Z must hold. Note that Equation 17 implies that Y is biased, and that among individuals matched on Y, there will be no group differences on W. A common factor interpretation for W is consistent with Equations 9 and 16, but is generally inconsistent with Equations 10 and 17. Under Condition IIIA, if UCI holds for Z, reverse OCI will hold even though Y is biased. In this situation, reverse OCI could lead to an erroneous conclusion of UCI for Y.

Discussion

Most of the commonly used methods for the detection of measurement bias provide tests of OCI. When one of the three sets of conditions presented above is met, these methods also provide tests of UCI for Y. An implicit assumption behind the development of these conditions for inference is that UCI of Y or Z is of ultimate interest. In other words, it was assumed that the question of bias rests on the relationship of the measured variables Y or Z to the unmeasured variable W. The variable W may be conceptualized in various ways depending on the context—as a "latent trait," an "aptitude," or an "ability." In all cases, the investigation of OCI is pursued to provide evidence for the evaluation of UCI of Y or Z. The foregoing results clarify the requirements for inference in specific applications. In developing these results, minimal assumptions have been made about distributional forms, linearity of regressions, homoscedasticity, or causal relationships. More precise results can be reached by invoking such assumptions, and restrictive assumptions may be justified in particular applications. Here, the intention was to explore the fundamental conditions that justify inferences between the two forms of invariance.

An implication of the conditions presented is that tests of OCI can justify inferences regarding UCI for Y even though Z is biased and/or unreliable as a measure of W. Under Condition IIIA, failure of OCI implies failure of UCI for Y, although Z is biased. Perfect reliability in Z is required only in Condition I; Z may be imperfectly reliable in Conditions IIA, IIB, IIIA, and IIIB. Thus, the demonstration of bias and/or unreliability in Z is not sufficient, by itself, to negate the use of Z in tests of OCI and subsequent inference of UCI for Y. Parallel conclusions apply to Y in the case of reverse OCI and UCI of Z.

In most other respects, however, the conditions that justify inferences concerning UCI of Y are quite restrictive. Condition I is unlikely to be fully met in any application. It represents a limiting condition that is only approximately achieved in real data. Conditions IIA and IIB place weaker requirements of Bayes sufficiency and unbiasedness on Z. In DIF applications, the Rasch parametric model leads to Equations 3 and 5. Departures from Rasch model assumptions, such as that produced by substantial guessing on multiple-choice items, will lead to violations of Equations 3 and 5 when Z is an unweighted total score. In such cases, one option would be to expand Z to include additional measures of W that are external to the test under study. Assuming that such measures can be found, it still will be difficult to verify that the resulting multivariate Z satisfies Equations 3 and 5. Also, care must be taken so that the additional measures do not introduce biases in Z that violate Equations 4 and 6. Conditions IIA and IIB are not helpful in predictive bias applications, because the focus in such cases is on testing UCI of Z. Confirmation of OCI has no necessary implication for UCI of Z under either condition. An important feature of the predictive bias application is that the criterion measures included in Y are usually a matter of choice, and different choices lead to different characterizations of W. Improper choices can lead to violations of Bayes sufficiency if Y requires abilities that are irrelevant to actual job or educational performance.

Conditions IIIA and IIIB require local independence properties and conditional independence between W and V given Z, or given Z and Y. Neither condition is realistic in the DIF application, because Equations 8 and 10 are unlikely to hold, and Equation 8 cannot hold if Z is free of bias. Conditions IIIA and IIIB offer no useful way of testing UCI of Z in predictive bias applications. In fact, these conditions represent a situation in which Z could be erroneously declared free of bias using OCI detection methods.

In predictive bias applications, the equivalence conditions between reverse OCI and UCI of Z are equally restrictive. Reversed Condition I will not be fully met in real data. For Conditions IIA and IIB, Equation 15 and/or Equation 13 are often assumed to hold. Under this assumption, the value of reverse OCI in testing UCI of Z will depend on the Bayes sufficiency assumptions concerning Y in Equations 12 and 14. Certain choices for Y may invalidate Bayes sufficiency by failing to capture target abilities that are relevant to performance on Z. This problem is more likely to occur when W is multivariate. Bayes sufficiency is also generally violated if W is viewed as a common factor underlying Y and Z, as noted above. Reversed Conditions IIIA and IIIB impose requirements that are nnearly identical to Conditions IIIA and IIIB. Common-factor interpretations of W are inconsistent with the requirements of conditional independence between W and V given Y. The conditions require the use of a biased criterion measure Y, contrary to the usual assumption in predictive bias applications. Reverse OCI may hold under these conditions even though Y is biased.

The above results suggest that bias detection methods that provide tests of OCI, or reverse OCI, do not provide a rigorous basis for inference about UCI of Y or Z. Furthermore, the minimum requirements for inference involve restrictive assumptions about the unobserved W and its relation to Y, Z, and V. One immediate implication of these results is that efforts to modify either Y or Z to achieve OCI will not guarantee UCI of Y or Z. For example, in predictive bias applications, the item content of Z might be altered to achieve OCI in relation to a fixed criterion Y. Test construction strategies of this sort need not lead to UCI of Z, however, if none of the equivalence conditions are satisfied.

Alternative Approaches to Inferring Bias

If tests of OCI generally do not provide a rigorous basis for the inference of bias, two alternatives are available in practice. First, UCI of Y or Z can be investigated through direct modeling of the Y/W or Z/W relationship. Second, the effects of violations of the foregoing inferential conditions can be explored through simulations under various parametric models. In the second approach, conclusions drawn from tests of OCI should be robust against minor violations of the inferential conditions.

The modeling alternative. This alternative has been used effectively in DIF applications by employing measurement models under IRT. The common approach is to model the Y/W relationship within a parametric model—evaluating both the fit of the model and the invariance of the form of the Y/W relationship over V (Lord, 1980; Thissen et al., 1988). There are at least two difficulties with this approach. First, the sample size requirements for estimation and fit evaluation may exceed practical limitations. Second, the usual reliance on fully parametric models carries restrictions that are not directly relevant to the purpose of testing UCI. For example, unidimensionality assumptions are usually made, but are not a requirement for UCI. Ideally, tests of UCI should be conducted without unnecessary assumptions.

Although the sample size problem remains, recently there has been progress in addressing the second problem. Nonparametric conditions required by general latent variable models were described by Holland & Rosenbaum (1986; Holland, 1981; Rosenbaum, 1984a) who also presented approaches for testing whether a monotone latent variable model is consistent with empirical data. Stout (1987, 1990) developed the concept of ''essential unidimensionality'' as an alternative to strict unidimensionality, along with nonparametric methods of assessing dimensionality. In related work, Shealy & Stout (1990) defined test bias (failure of UCI) within a multidimensional model that posits ''nuisance abilities'' as the source of bias. Kok (1988) applied a similar definition. Within this definition, Shealy & Stout developed an approach to testing UCI that does not require specification of a fully parametric model. In a different direction, Jannerone (1986, 1987) and Junker (1990) each explored models that weaken traditional assumptions of strict local independence among test items. In sum, although at present fully general procedures are not available for direct tests of UCI under minimal assumptions, recent research has widened the scope of available latent variable modeling procedures.

As applied in studies of predictive bias, the modeling approach seeks to model the relationship of the test score Z to W, or the relationship of both Y and Z to W. The most common approach has been to disaggregate items or subtests in Z and study the invariance of the factorial structure of these items or subtests. Confirmatory factor analysis is a useful tool in this situation (Jöreskog, 1971). Factor analytic investigations of invariance should extend to latent means as well as covariance structures (Meredith, 1990; Millsap & Everson, 1991). An obvious alternative approach is to apply item bias detection procedures to the individual test items using the above-mentioned modeling methods. Drasgow (1987) illustrated this approach and suggested that the test response function constructed after fitting items to item models is a useful indicator of test-level bias. Clearly, the two problems mentioned earlier also apply to modeling efforts in predictive bias applications. Finally, it must be remembered that confirmation of UCI for a test Z has no necessary implication for OCI.

The simulation approach. As an alternative to direct modeling, the robustness of conclusions drawn from tests of OCI can be evaluated under violations of the inferential conditions presented earlier. This approach is analogous to sensitivity analysis for causal inference in nonexperimental research designs (Rosenbaum, 1984b; Rosenbaum & Rubin, 1983). The general idea is to demonstrate the impact of varying departures from the inferential conditions on conclusions drawn from tests of OCI. For example, suppose that in the DIF application, a parametric model, such as the two- or three-parameter logistic model, is assumed. Neither model supports Bayes sufficiency of Z when Z is an unweighted sum of item scores. It is still possible that in tests of sufficient length, ''near'' Bayes sufficiency may be achieved, and tests of OCI will have practical value as a basis for decisions about UCI. Ordinarily, simulations would be necessary to determine the minimum test length beyond which inferences could be made safely. This kind of approach has already been used effectively in DIF applications (Donoghue, Holland, & Thayer, 1989; Rudner, Getson, & Knight, 1980; Spray, 1989; Swaminathan & Rogers, 1990; Thissen et al., 1988). Explorations under linear regression or path models have been useful in predictive bias applications (Humphreys, 1986; Linn, 1984). The inferential conditions presented earlier can help structure such investigations by identifying factors that determine the equivalence between OCI and UCI of Y or Z.

In the present development of the equivalence conditions, a variety of practical issues that will influence bias detection in real applications were ignored. The problems of low statistical power and of sample representativeness that are important in finite samples were not discussed. Missing data occur frequently in practice and can affect the results of OCI assessments and of direct assessments of UCI (Linn, 1983; Little & Rubin, 1987; Lord & Novick, 1968; Meredith, 1964; Reilly, 1973). Forms of invariance that are weaker than OCI and UCI also have not been discussed. For example, Equations 1 and 2 could be reformulated in terms of conditional expectation rather than conditional probability. Invariance of conditional expectations may hold in cases in which other features of the conditional probability distribution are not invariant. The specification of equivalence conditions between weaker forms of invariance of this sort is a useful direction for future research.

Appendix

Let Y, Z, W, and V be discrete, possibly multivariate, random variables. Assume that no pair of these variables is marginally independent, but conditional independence is allowed. Let y, z, w, and v denote realizations of the respective random variables. The subscripts g, h, and i denote distinct values of the realizations.

UCI of Y given W and V holds if and only if

$$P(Y \mid W = w, V = v) = P(Y \mid W = w) \ , \tag{18}$$

for all w and v. UCI of Z is defined by replacing Y with Z in Equation 18. OCI of Y given Z and V holds if and only if

$$P(Y \mid Z = z, V = v) = P(Y \mid Z = z) \ , \tag{19}$$

for all z and v.

Conditions IIA and IIIA each represent situations in which OCI in Equation 19 can be inferred given UCI of Y in Equation 18. As proof, begin with Condition IIA and assume that Equation 18 holds. This condition says that Equation 19 will be true when the following two equations are true:

$$P(Y \mid Z = z, W = w) = P(Y \mid Z = z) \tag{20}$$

and

$$P(Z \mid Y = y, W = w, V = v) = P(Z \mid Y = y, W = w) \ , \tag{21}$$

for all y, z, w, and v. Then Equations 18, 20, and 21 imply

$$\begin{aligned} P(Y, Z \mid W = w, V = v) &= P(Z \mid Y = y, W = w) P(Y \mid W = w) \\ &= P(Y, Z \mid W = \mathrm{w}) = P(Y \mid Z = z) P(Z \mid W = \mathrm{w}) \ . \end{aligned} \tag{22}$$

Also,

$$\begin{aligned} P(Z \mid W = w, V = v) &= \sum_g P(Y_g, Z \mid W = w, V = v) \\ &= \sum_g P(Y_g \mid Z = z) P(Z \mid W = w) = P(Z \mid W = w) \ , \end{aligned} \tag{23}$$

and hence,

$$P(Y \mid Z = z, V = v) = \frac{\sum_i P(Y, Z, W_i, V)}{\sum_i P(Z, W_i, V)} = \frac{\sum_i P(Y, Z \mid W_i = w_i, V = v) P(W_i, V)}{\sum_i P(Z \mid W_i = w_i, V = v) P(W_i, V)} = P(Y \mid Z = z) \ . \tag{24}$$

Thus, Condition IIA is established. Assuming Equation 18, Condition IIIA says that Equation 19 will hold if the following two equations are true:

$$P(Y \mid Z = z,\ W = w,\ V = v) = P(Y \mid W = w,\ V = v) \tag{25}$$

and

$$P(W \mid Z = z,\ V = v) = P(W \mid Z = z)\ . \tag{26}$$

As proof, note that Equations 18 and 25 imply

$$P(Y \mid Z = z,\ W = w,\ V = v) = P(Y \mid W = w)\ . \tag{27}$$

Then, using Equation 26,

$$\begin{aligned} P(Y \mid Z = z,\ V = v) &= \sum_i P(Y \mid Z = z,\ W_i = w_i,\ V = v)P(W_i \mid Z = z,\ V = v) \\ &= \sum_i P(Y \mid W_i = w_i)P(W_i \mid Z = z) = P(Y \mid Z = z)\ , \end{aligned} \tag{28}$$

and Condition IIIA is established. Note also that Equations 18 and 25 imply

$$\begin{aligned} P(Y \mid Z = z,\ W = w) &= \sum_j P(Y \mid Z = z,\ W = w,\ V_j = v_j)P(V_j \mid W = w,\ Z = z) \\ &= P(Y \mid W = w)\ . \end{aligned} \tag{29}$$

Conditions IIB and IIIB represent situations in which UCI of Y in Equation 18 can be inferred from OCI in Equation 19. Given Equation 19, Condition IIB says that Equation 18 will hold if the following two equations are true:

$$P(Y \mid Z = z,\ W = w,\ V = v) = P(Y \mid Z = z,\ V = v) \tag{30}$$

and

$$P(Z \mid W = w,\ V = v) = P(Z \mid W = w)\ . \tag{31}$$

As proof, Equations 19 and 30 imply that

$$P(Y \mid Z = z,\ W = w,\ V = v) = P(Y \mid Z = z)\ . \tag{32}$$

Then, using Equation 31,

$$P(Y,Z \mid W = w,\ V = v) = P(Y \mid Z = z)P(Z \mid W = w) = P(Y,Z \mid W = w) \tag{33}$$

and

$$\begin{aligned} P(Y \mid W = w,\ V = v) &= \sum_h P(Y,Z_h \mid W = w,\ V = v) \\ &= \sum_h P(Y,Z_h \mid W = w) = P(Y \mid W = w)\ . \end{aligned} \tag{34}$$

Thus, Condition IIB is established. Note that Equations 19, 30, and 31 also imply that

$$P(Y \mid Z = z,\ W = w) = \frac{P(Y,Z \mid W = w,\ V = v)}{P(Z \mid W = w,\ V = v)} = P(Y \mid Z = z)\ . \tag{35}$$

Condition IIIB says that Equation 19 and the following two equations imply Equation 18:

$$P(Y \mid W = w,\ Z = z) = P(Y \mid W = w) \tag{36}$$

and

$$P(W|Y = y, Z = z, V = v) = P(W|Y = y, Z = z) \ . \tag{37}$$

As proof, Equations 19 and 37 imply

$$\begin{aligned} P(Y, W|Z = z, V = v) &= P(W|Z = z, Y = y)P(Y|Z = z) \\ &= P(Y, W|Z = z) = P(Y|W = w)P(W|Z = z) \ . \end{aligned} \tag{38}$$

Then,

$$P(W|Z = z, V = v) = \sum_g P(Y_g, W|Z = z, V = v) = P(W|Z = z) \ , \tag{39}$$

and, using Equation 36,

$$P(Y|W = w, V = v) = \frac{\sum_h P(Y, Z_h, W, V)}{\sum_h P(Z_h, W, V)} = \frac{\sum_h P(Y, W|Z_h = z_h, V = v)P(Z_h, V)}{\sum_h P(W|Z_h = z_h, V = v)P(Z_h, V)} = P(Y|W = w) \ . \tag{40}$$

Thus, Condition IIIB is established.

Condition I implies that Equations 20, 21, 30, and 31 all hold. Hence, the above proofs also establish that if Condition I is true, UCI of Y holds if and only if OCI in Equation 19 holds. Proofs for the reversed conditions can be obtained by reversing the roles of Y and Z in the above proofs.

Given the restrictions on marginal independence, Conditions IIA and IIIA cannot both hold. As shown above, Condition IIIA implies that $P(Y|Z = z, W = w) = P(Y|W = w)$, and Equation 20 holds under Condition IIA. Assume both conditions are true. Then $P(Y|W = w) = P(Y|Z = z)$, and

$$P(Y, W) = P(Y|Z = z)P(W) \tag{41}$$

and

$$\sum_i P(Y, W_i) = P(Y) = P(Y|Z = z) = P(Y|W = w) \ , \tag{42}$$

which implies marginal independence between Y and Z and between Y and W. Similarly, Condition IIA implies that $P(Z|W = w, V = v) = P(Z|W = w)$, and Equation 26 is true under Condition IIIA. Assume that both conditions hold. Then

$$\begin{aligned} P(Z, W, V) &= P(Z|W = w)P(W, V) = P(W|Z = z)P(Z, V) = P(V|W = w)P(W, Z) \\ &= P(V|Z = z)P(W, Z) \ , \end{aligned} \tag{43}$$

which implies that $P(V|W = w) = P(V|Z = z)$. Using reasoning identical to that used in the previous proof (i.e., replacing Y with V), V is marginally independent of both Z and W. Hence, Conditions IIA and IIIA cannot both hold. Completely analogous arguments establish that Conditions IIB and IIIB will not both hold given OCI in Equation 19.

References

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

Birnbaum, M. H. (1979). Procedures for the detection and correction of salary inequities. In T. R. Pezzullo & B. E. Brittingham (Eds.), *Salary equity* (pp. 121–144). Lexington MA: Lexington Books.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement, 10,* 237–255.

Conway, D. A., & Roberts, H. V. (1983). Reverse regression, fairness, and employment discrimination. *Journal of Business and Economic Statistics, 1,* 75–85.

Darlington, R. B. (1971). Another look at "cultural fairness." *Journal of Educational Measurement, 8,* 71–82.

Dempster, A. P. (1988). Employment discrimination and statistical science. *Statistical Science, 3,* 149–185.

Donoghue, J., Holland, P. W., & Thayer, D. T. (1989, March). *A Monte Carlo study of factors that affect the Mantel-Haenszel measure of differential item functioning.* Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco.

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.

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

Goldberger, A. S. (1984). Reverse regression and salary discrimination. *Journal of Human Resources, 19,* 293–318.

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

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.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology, 71,* 327–333.

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

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

Jannerone, R. J. (1987). *Locally dependent models for reflecting learning abilities* (Report No. 87-67). Columbia: University of South Carolina, Center for Machine Intelligence.

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

Junker, B. W. (1990, June). *Essential independence and structural robustness in item response theory.* Paper presented at the annual meeting of the Psychometric Society, Princeton NJ.

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

Lehmann, E. L. (1986). *Testing statistical hypotheses.* New York: Wiley.

Linn, R. L. (1983). Predictive bias as an artifact of selection procedures. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A festschrift for Frederic M. Lord* (pp. 27–40). Hillsdale NJ: Erlbaum.

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement, 21,* 33–47.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8,* 1–4.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

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.

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 χ^2 statistics. *Journal of Educational Measurement, 18,* 229–248.

McFatter, R. M. (1982). On detecting sex bias in salaries. *American Psychologist, 37,* 1144–1146.

McFatter, R. M. (1987). Use of latent variable models for detecting discrimination in salaries. *Psychological Bulletin, 101,* 120–125.

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

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29,* 177–185.

Meredith, W. (1990, October). *Factorial invariance from a measurement invariance perspective.* Paper presented at the annual 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.

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

Peterson, D. W. (1986). Measurement error and regression analysis in employment cases. In D. H. Kaye & M. H. Aicken (Eds.), *Statistical methods in discrimination litigation* (pp. 107–131). New York: Dekker.

Raju, N. S., & Normand, J. (1985). The regression bias method: A unified approach for detecting item bias and selection bias. *Educational and Psychological Measurement, 45,* 37–54.

Reilly, R. R. (1973). A note on minority group test bias studies. *Psychological Bulletin, 80,* 130–132.

Reilly, R. R. (1986). Validating employee selection procedures. In D. H. Kaye & M. H. Aicken (Eds.), *Statistical methods in discrimination litigation* (pp. 133–158). New York: Dekker.

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

Rosenbaum, P. R. (1984b). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association, 79,* 41–48.

Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study. *Journal of the Royal Statistical Society, 45,* (Series B), 212–218.

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

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

Shealy, R., & Stout, W. F. (1990, June). *A new model and statistical test for psychological test bias.* Paper presented at the annual meeting of the Psychometric Society, Princeton NJ.

Shepard, L. A., 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.

Spray, J. A. (1989). *Performance of three conditional DIF statistics in detecting differential item functioning on simulated tests* (Research Rep. Series No. 89-7). Iowa City IA: American College Testing Program.

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.

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

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

Wainer, H., & Steinberg, L. S. (1991, June). *Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A bidirectional validity study.* Paper presented at the annual meeting of the Psychometric Society, New Brunswick NJ.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185–197.

Acknowledgments

Preparation of this article was supported in part by PSC-CUNY grant No. 661282 to Roger E. Millsap.

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 Ave., New York NY 10010, U.S.A.