Fairness and Bias

Fairness

Fairness is a social rather than a psychometric concept. Its definition depends on what one considers to be fair. Fairness has no single meaning and, therefore, no single definition, whether statistical, psychometric, or social. The *Standards* notes four possible meanings of "fairness."

The first meaning views fairness as requiring equal group outcomes (e.g., equal passing rates for subgroups of interest). The *Standards* rejects this definition, noting that it has been almost entirely repudiated in the professional testing literature. It notes that while group differences should trigger heightened scrutiny for possible sources of bias (i.e., a systematic error that differentially affects the performance of different groups of test takers), outcome differences in and of themselves do not indicate bias. It further notes that there is broad agreement that examinees with equal standing on the construct of interest should, on average, earn the same score regardless of group membership.

The second meaning views fairness in terms of the equitable treatment of all examinees. Equitable treatment in terms of testing conditions, access to practice materials, performance feedback, retest opportunities, and other features of test administration, including providing reasonable accommodation for test takers with disabilities when appropriate, are important aspects of fairness under this perspective. There is consensus on a need for equitable treatment in test administration (although not necessarily on what constitutes equitable treatment).

The third meaning views fairness as requiring that examinees have a comparable opportunity to learn the subject matter covered by the test. However, the *Standards* notes that this perspective is most prevalent in the domain of educational achievement testing and that opportunity to learn ordinarily plays no role in determining the fairness of employee selection procedures. One exception would be settings where the organization using the tests purposely limits access to information needed to perform well on the tests on the basis of group membership. In such cases, while the test itself may be unbiased in its coverage of job content, the use of the test would be viewed as unfair under this perspective.

The fourth meaning views fairness as a lack of predictive bias. This perspective views predictor use as fair if a common regression line can be used to describe the predictor-criterion relationship for all subgroups of interest; subgroup differences in regression slopes or intercepts signal predictive bias. There is broad scientific agreement on this definition of predictive bias, but there is no similar broad agreement that the lack of predictive bias can be equated with fairness. For example, a selection system might exhibit no predictive bias by race or gender, but still be viewed as unfair if equitable treatment (e.g., access to practice materials) was not provided to all examinees.

Thus, there are multiple perspectives on fairness. There is agreement that issues of equitable treatment, predictive bias, and scrutiny for possible bias when subgroup differences are observed, are important concerns in personnel selection; there is not, however, agreement that the term "fairness" can be uniquely defined in terms of any of these issues.

Bias

The *Standards* notes that bias refers to any construct-irrelevant source of variance that results in systematically higher or lower scores for identifiable groups of examinees. The effect of such irrelevant sources of variance on scores on a given variable is referred to as measurement bias. The effects of such sources of variance on predictor-criterion relationships, such that slope or intercepts of the regression line relating the predictor to the criterion are different for one group than for another, is referred to as predictive bias. The *Standards* notes that, in the employment context, evidence of bias or lack of bias generally relies on the analysis of predictive bias. Both forms of bias are discussed below.

Predictive Bias

While fairness has no single accepted meaning, there is agreement as to the meaning of predictive bias. Predictive bias is found when for a given subgroup, consistent nonzero errors of prediction are made for members of the subgroup (Cleary, 1968; Humphreys, 1952). (Another term used to describe this phenomenon is differential prediction. The term "differential prediction" is sometimes used in the classification and placement literature to refer to differences in predicted performance when an individual is classified into one condition rather than into another; this usage should not be confused with the use of the term here to refer to predictive bias.) Although other definitions of bias have been introduced, such models have been critiqued and found wanting on grounds such as lack of internal consistency (Petersen & Novick, 1976).

Testing for predictive bias involves using moderated multiple regression, where the criterion measure is regressed on the predictor score, subgroup membership, and an interaction term between the two. Slope and/or intercept differences between subgroups indicate predictive bias (Gulliksen & Wilks, 1950; Lautenschlager & Mendoza, 1986; Nunnally & Bernstein, 1994).

Predictive bias has been examined extensively in the cognitive ability domain. For White–African American and White–Hispanic comparisons, slope differences are rarely found; while intercept differences are not uncommon, they typically take the form of overprediction of minority group performance (Bartlett, Bobko, Mosier, & Hannan, 1978; Hunter, Schmidt, & Rauschenberger, 1984; Schmidt, Pearlman, & Hunter, 1980). In some other domains, there has been little to no published research on predictive bias,

though work in the personality domain is now beginning to appear. Saad and Sackett (2002) report findings parallel to those in the ability domain in examining predictive bias by gender using personality measures (i.e., little evidence of slope differences and intercept differences in the form of over-prediction of female performance). Given the limited research to date, broad conclusions about the prevalence of predictive bias for many constructs are premature at this time.

Several important technical concerns with the analysis of predictive bias are noted here. The first is that an analysis of predictive bias requires an unbiased criterion. Confidence in the criterion measure is a prerequisite for an analysis of predictive bias. It is important to note, though, that while researchers should exercise great care in the development and collection of criterion data, investigations of criterion bias are limited by the lack of a true score against which criterion measures can be compared. The second is the issue of statistical power to detect slope and intercept differences. Small total or subgroup sample sizes, unequal subgroup sample sizes, range restriction, and predictor unreliability are factors contributing to low power (Aguinis, 1995; Aguinis & Stone-Romero, 1997). A third is the assumption of homogeneity of error variances (Aguinis, Peterson, & Pierce, 1999); alternative statistical tests may be preferable when this assumption is violated (Alexander & DeShon, 1994; DeShon & Alexander, 1996; Oswald, Saad, & Sackett, 2000).

Some perspectives view the analysis of predictive bias as an activity contingent on a finding of mean subgroup differences. In fact, however, subgroup differences and predictive bias can exist independently of one another. Thus, whether or not subgroup differences on the predictor are found, predictive bias analysis should be undertaken when there are compelling reasons to question whether a predictor and a criterion are related in a comparable fashion for specific subgroups, given the availability of appropriate data. In domains where relevant research exists, generalized evidence can be appropriate for examining predictive bias.

Measurement Bias

Measurement bias, namely, sources of irrelevant variance that result in systematically higher or lower scores for members of particular groups, is a potential concern for all variables, both predictors and criteria. Determining whether measurement bias is present is often difficult, as this requires comparing an observed score to a true score. In many domains, such as performance appraisal, such a standard for comparison is generally unavailable.

An approach to examining measurement bias in the domain of multi-item tests is to perform a differential item functioning (DIF) analysis. DIF refers to analyses that identify items for which members of different subgroups with identical total test scores (or identical estimated true scores in

item response theory [IRT] models) have differing item performance. Such analyses are uncommon in the employment domain. First, they require data on large research samples prior to operational use, as DIF analyses are often part of the predictor development process. Second, empirical research in domains where DIF analyses are common has rarely found sizable and replicable DIF effects (Sackett, Schmitt, Ellingson, & Kabin, 2001). Third, such analyses require unidimensional tests, and many employment tests are not factorially pure unidimensional tests, and the unidimensionality assumption is often untested in DIF research (Hunter & Schmidt, 2000). Fourth, for cognitive tests it is common to find roughly equal numbers of differentially functioning items favoring each subgroup, resulting in no systematic bias at the test level (Hunter & Schmidt, 2000). As a result of these factors, DIF findings should be viewed with caution. DIF analysis is not likely to become a routine or expected part of the test development and validation process in employment settings; however, researchers may choose to explore DIF when data sets appropriate for such analysis are available.

Linked to the idea of measurement bias in terms of conducting analysis at the item level is the concept of an item sensitivity review, in which items are reviewed by individuals with diverse perspectives for language or content that might have differing meaning for members of various subgroups and language that could be demeaning or offensive to members of various subgroups. Instructions to candidates and to scorers or assessors also may be reviewed in a similar manner. The value of such analysis will vary by test content, and the need for and use of such information is a matter of researcher judgment in a given situation.