DOCUMENT RESUME

ED 33C 726 TM 016 333

AUTHOR Palomares, Ronald S.; Friedrich, Katherine R.

TITLE Conceptual Underpinnings and Historical Perspectives on Evaluating Test Bias.

PUB DATE Jan 91

NOTE 17p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 24-26, 1991).

PUB TYPE Information Analyses (070) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Educational History; *Evaluation Methods; Literature Reviews; *Mathematical Models; *Predictive Measurement; Regression (Statistics); *Test Bias; Test Validity

ABSTRACT

Var:ous classical approaches to evaluating test bias are reviewed, and their conceptual underpinnings are discussed. Bias is examined as it relates to predictive validity, rather than construct validity. One of the first test bias models was developed by T. A. Cleary (1968). Cleary's work with test bias and regression stimulated others to develop related methods. One of the first to do so was R. L. Thorndike. Thorndike actually developed a selection model, not a definition of test bias, because his model imposes a method to ensure the fair use of a test in selection procedures. The restatement of the models of Cleary and Thorndike by R. B. Darlington (1971) is also appropriate for ensuring fair selection, but not for the determination of test bias. The definition of bias of A. R. Jensen (1980) is inclusive and gives a clear statistical definition of bias in that it is, in fact, the definition of a biased predictor. A test is considered biased with respect to predictive validity when there is a significant difference between the majority and minority groups as found between the slopes, intercepts, or standard errors of estimates. The emphasis is on the predictive ability of the test, and not some internal property of the test itself. Five graphs illustrate the text. (SLD)

CONCEPTUAL UNDERPINNINGS AND HISTORICAL PERSPECTIVES ON EVALUATING TEST BIAS

Ronald S. Palomares & Katherine R. Friedrich

Texas A&M University 77843-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, January 24, 1991.

ABSTRACT

Test bias studies date back several decades (cf. Taylor & Russell, 1939). The traditional models of evaluating test bias use regression analyses with the test in question as a predictor and a criterion of the researchers choice (Cleary , 1968). The present paper reviews various classical approaches to evaluating test bias (e.g. Darlington, 1071; Thorndike, 1971) with a view toward explaining the conceptual underpinnings of these methods.

CONCEPTUAL UNDERPINNINGS AND HISTORICAL PERSPECTIVES ON EVALUATING TEST BIAS

The controversy over the issue of test bias has been present since the introduction of testing, and originated specifically within intelligence testing. Confusion results from the lack of a clear understanding of the statistical definition of test bias and it's application to testing. This paper discusses the evolution of Cleary's (1968) original definition of test bias and how it applies to evaluating the predictive validity of a test. From Cleary's definition (1968) three additional theories have been derived, i.e., the theories of Darlington (1971), Thorndike (1971), and Jensen (1980). Each of these four theories will be reviewed. The review will encompass the history of the definition of test bias, as it refers to predictive validity, and will conclude with a restatement of Cleary's original definition as it is now employed in contemporary measurement practice.

The emotional aspect of this issue sometimes confounds the understanding of the term bias and it's application to testing. Much of this confusion results from the various ways in which bias can be defined. As Reynolds (1982a) states:

> Discussions of test bias are frequently accompanied by emotionally laden polemics decrying the use of tests with any minority group member and considering all tests to be inherently biased against individuals who are not members of the white middle class. (p. 178)

This paper does not attempt to address the emotional and social issues which surround this aspect of testing. Instead, the focus will be placed upon one definition of the empirical evaluation of bias, as put forth by Reynolds (1982a, 1982b) and Jensen (1980). This widely recognized, distinct statistical view of bias defines bias as "constant or systematic error, as opposed to chance or random error, in the estimation of some value" (Reynolds, 1982b, p. 199).

When examining the test as whole, in relation to test bias, the validity of the particular test must be considered. Some researchers distinguish construct validity and predictive/criterion related validity, and emphasize the importance of construct validity. Construct validity refers to the extent to which a test measures what it purports to measure (Allen & Yen, 1979; Annastasi, 1988). In terms of bias, Reynolds (1982a) has offered the following definition:

> Bias exists in regards to construct validity when a test is shown to measure different hypothetical traits (psychological constructs) for one group than another or to measure the same trait but with differing degrees of accuracy. (p. 194)

The issue of bias, in terms of construct validity, is of major importance. A test that is biased is not measuring the constructs, and only those, that it is purported to measure. Evidence of nonbiasedness in this area must be established before the question of predictive validity can be addressed.

Since the main focus of the present paper is to define test bias as it relates to predictive validity, methods for examining bias in terms of construct validity will not be addressed here. The reader is referred to Berk (1982) for an in-depth review of the issue of bias as it relates to construct validity and methods for assessing it. The paper is grounded on assumptions that (a) bias can be defined in purely statistical terms and (b) before employing the methods discussed here the researcher has already investigated bias as it relates to construct validity.

Predictive validity refers to the relationship of the test in question to some external criterion, and how that test predicts an individuals' performance on that criterion (Annastasi, 1988; Crocker & Algina, 1986). A problem exists when examining bias and predictive validity in that the criterion measure is not inherently reliable, and may often be subjective (Jensen, 1980; Reynolds, 1982b). The researcher should be aware of this problem and take into consideration the properties of the particular criterion being utilized in a given case before drawing final conclusions concerning the extent to which the test may or may not be biased.

The relationship between a criterion (such as future performance in school) and a predictor, i.e., the test in question (an example would be the SAT), is statistically expressed in the form of a regression equation. This equation is typically written as

$$Y \;\texttt{<--}\; Y^* = bX + a$$

where Y* represents an individual's predicted score on the criterion (Y, e.g., future school performance) and X is the score of the individual on the test one wishes to predict from (e.g., the SAT). This relationship can be expressed graphically, as depicted in Figure 1, where b is the slope of the line and a is the y-intercept (Ott, 1988; Pedhazur, 1982).

Insert Figure 1 about here

Cleary's Model

Various test bias models have been based upon the linear regression of the criterion (Y) on the predictor (X) (Jensen, 1980). One of the first such models was developed by Γ. Anne Cleary in her research of black-white differences in college performance (grades) based upon SAT scores (Cleary, 1968). According to Cleary, the regression lines within each subgroup of a population must be equal (identical regression equations) in order to consider the test to be non-biased. That is, the predictor is biased when consistent non-zero errors of predictions are made for members of a sub-group of a population. This approach implies two hypotheses, first testing for equality of slopes, and secondly, testing for equality of intercepts.

Cleary (1968) suggests that the researcher first tests for the equality of the slopes to determine if the test is a valid predictor for both groups. If the slopes are not equal, then the test is not able to predict future performance, as measured by the criterion, to the same degree of accuracy within each subgroup. Given that the slopes are equal, the next step is to test for the hypothesis of equality of intercepts. If this hypothesis is rejected, Cleary suggests that the researcher can then conclude that the test is a biased predictor for one of the subgroups.

Hunter and Schmidt (1976) elaborate Cleary's definition in their discussion of test bias. When the regression lines differ, the test being used will result in bias because use of a single regression line can result in consistent over- or under-prediction for one or both groups. An example of this situation, when the regression lines have equal slopes, but still involves bias, is presented in Figure 2. If the regression line for group A is used as the prediction line for the entire population, the researcher will consistently over-predict for Group B, the group with the smaller Y-intercept (the regression "a" weight). The use of the common regression line, found between the two regression lines, will result in constant under-prediction for Group A and over-prediction for Group B.

Insert Figure 2 about here

Thorndike's Model

Cleary's work with test bias and regression stimulated others to develop related methods. One of the first to elaborate Cleary's work was Thorndike (1971). Thorndike began his theory with the assumption that the slope of the two regression lines were equal (Hunter & Schmidt, 1976;

Thorndike, 1971). When this assumption isvalid, Thorndike then proposed three cases that were possible. In the first case, the regression lines are the same, i.e., also have the same "a" weight or Y-intercept, thus satisfying Cleary's (1968) definition, so the test can be considered non-biased. The second possibility is that the regression line for one group (group B - minority) is higher than the regression line for another group (group A - majority), as shown in Figure 3, i.e., the "a" weights in the regression equations of the two groups differ. In this case, using the majority group regression line (A), there is bias due to the consistent under-prediction of the minority group (B) (Thorndike, 1971).

Insert Figure 3 about here

The third possibility is the reverse of the second, where the majority regression line is above that of the minority and each line is used for it's own particular group as the regression equation for prediction. Throughout an extended argument, Thorndike finally concludes that this situation is also unfair to the minority group and states that a single regression line (i.e., the majority line) should be used (Hunter & Schmidt, 1976; Thorndike, 1971). Some authors (Jensen, 1980; Reynolds, 1982b) have noted that selection models are often construed as methods of determining bias, whereas they are really only models to be used for fair selection. Thorndike' is in actuality a selection model, not a definition of test bias, because it imposes a method to ensure the fair use of a test in selection procedures.

Darlington's Model

At the same time Thorndike was developing his model, Darlington proposed his own theory (Darlington, 1971; Petersen & Novick, 1976). This model was seen as a restatement of both Cleary's and Thorndike's models, expressing them in the form of a correlation coefficient rather than a regression equation (Darlington, 1971; Hunter & Schmidt, 1976). Once again, the assumption of equal slopes was made, restating it as the assumption of equal standard deviations on both predictor and criterion variables and equal validity for both groups (Hunter & Schmidt, 1976). Given these assumptions, Darlington proposed four separate definitions of the test bias. Each of these definitions uses c as the applicant's group membership (1 = majority and 2 = minority) in the various calculations (Petersen & Novick, 1976). These are as follows:

1. $r_{cx} = r_{cy} / r_{xy}$
2. $r_{cx} = r_{yy}$
3. $r_{cx} = r_{cy} r_{xy}$
4. $r_{cx} = 0$

Petersen and Novick (1976) elaborate on each of these definitions, describing the first to be equivalent to the Regression Model, having a common regression line, the second to be the same as Thorndike's Constant Ration Model, while the third is a special case of Cole's Probability Model and the fourth is the same when subpopulations have equal means on the test. Furthermore, Petersen and Novick (1976) state that the results from these four equations are contradictory except when there is perfect validity. Darlington (1971) also claims that the definitions are

> all based on the false view that optimum treatment of cultural factors in test construction or test selection can be reduced to completely mechanical procedures. If a conflict arises between the two goals of maximizing a test's validity and minimizing the test's discrimination against certain cultural groups, then a subjective, policy-level decision must be made concerning the relative importance of

the two goals. (p. 71)

Darlington (1971) urges that the term "culture fairness" be replaced in public discussions by the concept of "cultural optimality" (p. 79). The question of a test's "cultural optimality" can then be divided up between a "subjective, policy-level question" and a "purely empirical question concerning the test's correlation with culture-modified variables" (p. 80). Thus, a test user must first decided upon whether there is a subpopulation that has a certain "subjective value" and then use (Y - kC) to be the predictive criterion variable, with k being the "subjective value" given to that subpopulation. In this approach Darlington's model can be viewed to be similar to Thorndike's models, which is appropriate for ensuring fair selection, but not in the determination of test bias.

Jensen's Model

Jensen (1980) states a theoretical definition of bias using the framework developed by Cleary (1968). His definition is all inclusive and provides a clear statistical definition of bias. This definition corresponds directly to the definition of bias stated previously in this paper, i.e., "constant or systematic error, as opposed to chance or random error, in the estimation of some value" (Reynolds, 1982b, pg 199). This definition, however, is not applicable in practice due to the stringent requirements placed upon the predictor, i.e.. perfect reliability (Jensen, 1980). Thus, Jensen (1980) has proposed a more realistic definition which states that if there is a significanct difference between the majority and minority groups, as found between the slopes, or the intercepts, or standard error of estimates, then the test may be considered a biased predictor.

At this point, it should be noted that this definition of bias is based upon the predictive power of the test. It is a definition of a biased predictor, not of a biased test in and of itself. The approach is designed to determine if the test is useful in the prediction of a certain criterion across a populations that has identifiable subgroups. The prediction of that criterion across the various subgroups should be the same in order for the test not to be considered bias, according to Jensen (1980). When a test is found to be lacking in predictive bias, a single regression equation can then be used for all subgroups within the population, otherwise separate regression equations should be computed for each subgroup (Jenson, 1980; Reynolds, 1982a).

Jensen (1980) also argues that three concepts of test bias should be recognized as being inadequate. They are the egalitarian fallacy, the culture-bound fallacy and the standardization fallacy. To fully understand the predictive bias that Jensen is defining, an understanding of these three concepts and why they fail is essential. The first, the egalitarian fallacy, is based upon the assumption that all subgroups have an equal distributions of the trait or ability that the specific test is measuring (Jensen, 1980; Reynolds, 1982b; Thorndike, 1971). This assumption rests upon the valid statement that there is no a priori conclusion that subgroups should differ in levels of the trait or ability being measured (Reynolds, 1982b). However, it is also usually true that there is no a priori reason to believe the subgroups do not differ. Therefore, the inference that tests are biased, based upon the assumption that the distribution between subgroups do not differ, is not scientifically justifiable (Reynolds, 1982b; Jensen, 1980; Thorndike, 1968).

According to Jensen (1980), the culture-bound fallacy states that the content or face validity of a test judged to be "culture bound" and unfair results in bias towards the subgroups that are not members of that culture. The assumption is that subgroups within the population have different cultural experiences, which will lead to the test becoming harder for other subgroups that did not have the same cultural experiences as the majority subgroup, and creates a biased test. The reason this view is fallacious is because the determination of bias is based soley on subjective criteria. The assessment of test bias "must be based on objective psychometric and statistical criteria" (Jensen, 1980, p. 371).

The standardization fallacy states that when a test is standardized only on the majority population, the test is then biased towards the minority population (Jensen, 1980). The fact that a test is standardized only upon one group within the population does not in itself constitute evidence of bias. The fact that a test that is reliable and valid for one subgroup of a population does not mean that it is reliable and valid for another. It is necessary that the researcher first determine the reliability and validity of the test in the non-standardization subgroup before any assessment of bias can be made (Jensen, 1980).

Summary

The issue of test bias has been present since intelligence testing was first introduced in the early 1900's (Jensen, 1980). It was not until 1968 that a statistical definition was proposed by Cleary (1968). Since then, several authors have augmented her initial work in pursuit of a comprehensive definition of test bias (Darlington, 1971; Jensen, 1980; Thorndike, 1971). The evolution of Cleary's original definition was stymied initially by a lack of clear understanding of the term bias and it's application to psychometrics. This problem was further confounded by the difficulty in distinguishing between methods for assessing bias and methods for assuring fair selection practices. As Hunter and Schmidt (1976) and others (Jensen, 1980; Reynolds, 1982a, 1982b) have argued, many of the definitions proposed by other researchers, such as Thorndike and Darlington, are in actuality models of selections and not a definition of test bias.

In recent years, Jensen (1980) and others (Reynolds 1982a, 1982b) have returned to Cleary's original work and refined her definition to provide researchers with a more theoretical and complete definition of test bias and its determination. It is as follows:

> A test is considered biased, with respect to predictive validity when the inference drawn from the test score is not made with the smallest feasible random error or if there is constant error in an inference or prediction as a function of membership in a particular group. (Reynolds, 1982b, p. 216)

One should note that the emphasis in this definition is placed upon the predictive ability of the test and not some internal property of the test itself. The controversy and confusion over test bias should no longer exist, as a clear definition of bias, with respect to the predictive validity, and has been provided by Jensen (1980) and further elaborated upon Reynolds (1982b) in their treatments of this issue.

References

Allen, M.J. & Yen, W.M. (1979). Introduction to measurement theory. CA: Brooks/Cole Publishing Company.

Annastasi, A. (1988). Psychological testing. New York: Macmillan.

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

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

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

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

Hunter, J.E. & Schmidt, F.L. (1976). Critical analysis of the statistical and ethical implications of various definitions of test bias. Psychological Bulletin, 83, 1053-1071.

Jensen, A.R. (1980). Bias in Mental Testing. New York: The Free Press.

Ott, L. (1988). An introduction to statistical methods and data analysis. Boston: PWS-Kent Publishing Company.

Pedhazur, E.J. (1982). Multiple Regression in Behavioral Research. Fort Worth: Holt, Rinehart and Winston.

Petersen, N.S. & Novick, M.R. (1976). An evaluation of some models for culture-fair selection. Journal of Educational Research, 13, 3-29.

Reynolds, C.R. (1982a). The problem of bias in psychological assessment. In C.R. Reynolds & T.B. Gutkin (Eds.), The handbook of school psychology (pp. 178-208). New York: Wiley. .pa

Reynolds, C.R. (1982b). Methods for detecting construct and predictive bias. In R.A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 199-

277). Baltimore Johns Hopkins University press.

Taylor, H.C., & Russell, J.T. (1936(). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. Journal of Applied Psychology, 23, 565-578.

Thorndike, R.L. (1971). Concepts of culture-fairness. Journal of Educational Measurement, 8, 63-70.

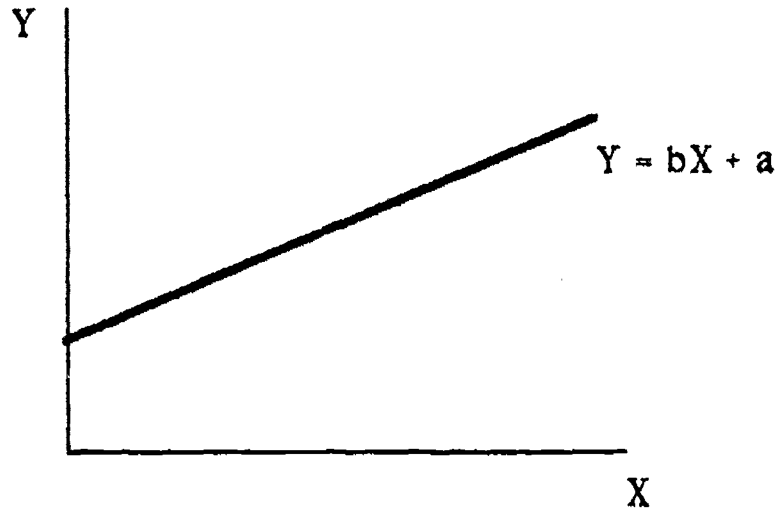


Figure 1: Graph of the regression of Y (the criterion variable) on X (the predictor variable) where b represents the slope of the line and a is the y-intercept.

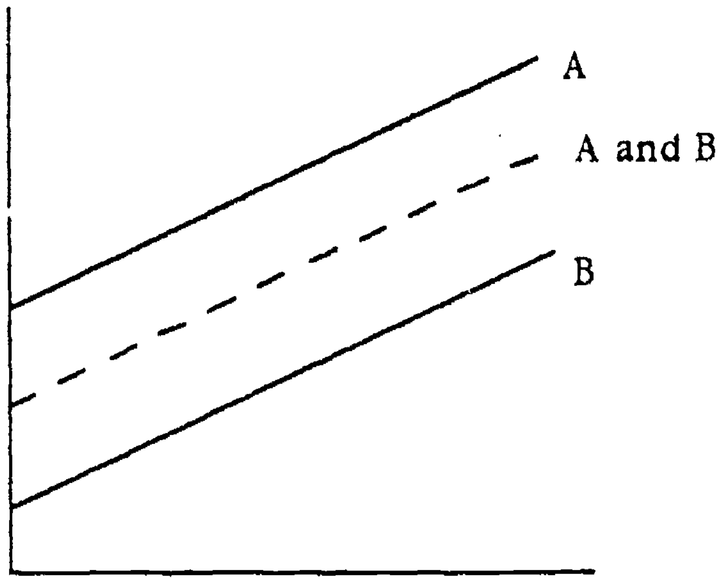
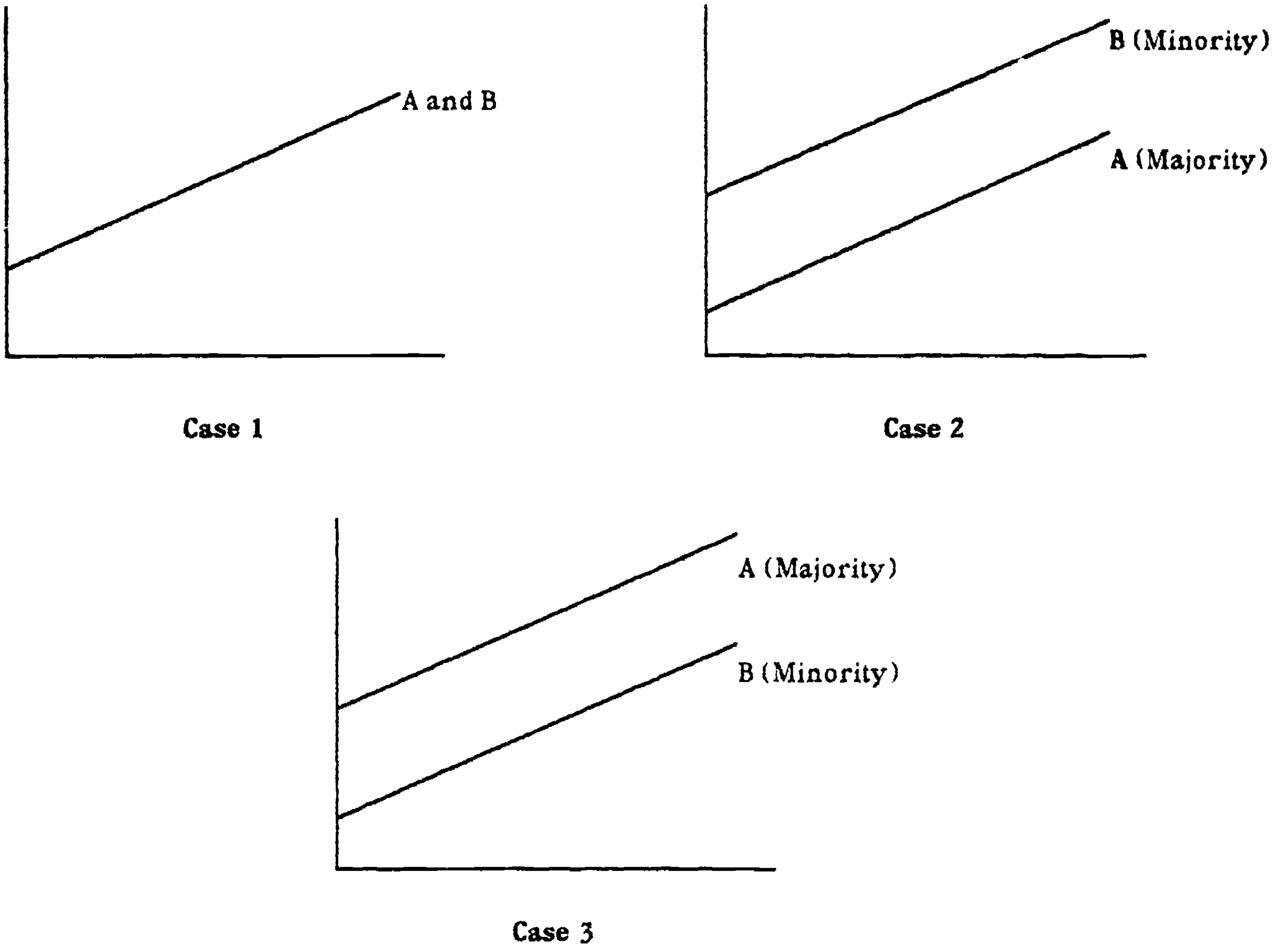


Figure 2: Graphical depiction of Cleary's definition of test bias. When the slopes of the regression eguations for the two subgroups are found to be unequal, use of a single regression line, as opposed to separate lines, will result in bias. The dashed line represents the common regression line for the combined sample.



Case 1

Case 2

Case 3

Figure 3: Three different senarios as described by Thorndike in his discussion of test bias. In case 1, both subgroups A and B have equal regression lines, resulting in a nonbiased test. For case 2, bias exists when the majority line is used for both groups, instead of separate lines. Thorndike argues in case 3 that a single regression line, the majority line, should be used for both groups to avoid bias against the minority group.