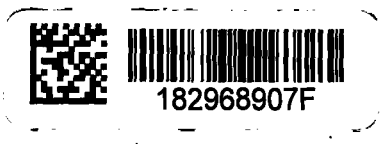
Office Report 1988-09



Bias, comparability, fairness, and utility in cross-cultural testing: Implications in the South African context

Elmien Spies-Wood

Western Cape Regic
Human Sciences Research Council
1988

Elmien Spies-Wood, M.A., Senior Researcher
Project leader: Dr J.M. Verster

Human Development Division
National Institute for Personnel Research
Executive Director: Dr G.K. Nelson

Printed and published by HSRC
134 Pretorius Street
Pretoria

ACKNOWLEDGEMENT

Dr G. K. Nelson

Dr Y. H. Poortinga

Dr J. M. Verster

Mr M. A. Coulter

Mrs R. Lombard

Mrs J. Rosinger and other staff members of the NIPR library.

CONTENTS

Abstract

Hypotheses, methodologies, and research findings from contemporary literature on testing across cultures are critically appraised and systematized. A distinction is made between attempts to explain the causes of test bias and the application of statistical procedures to determine whether test results can be compared across groups. Selection fairness (authoritative selection) is contrasted to the meanings attached to utility by industrial psychologists, statistical economists, and cross-cultural psychologists. The practical implications of select/reject, success/fail for management, the applicants, and disadvantaged groups in a plural society such as South Africa are outlined.

Ekserp

'n Kritiese waardering van hipoteses, metodologië en navorsingsbevindings oor kruiskulturele toetsing word aan die hand van kontemporêre literatuur gemaak en die gegewens word gesistematiseer. 'n Onderskeid word getref tussen pogings om die oorsake van toetssydigheid te verduidelik en statistiese metodes om vas te stel of toetsresultate tussen groepe vergelyk kan word. Keuringsredelikheid (gesaghebbende keuring) word gestel teenoor die betekenisse wat bedryfsielkundiges, statistiese ekonome en tussenkulturele sielkundiges aan keuringsnut heg. Die praktiese implikasies van keur/afkeur, slaag/misluk vir die bestuur, die aansoekers en die onbegunstigde groepe in 'n plurale samelewing soos Suid-Afrika word geskets.

1. HISTORICAL TRENDS

Since the first experimental work that can be regarded as cross-cultural psychological comparison was carried out (Rivers, 1901, 1905), through the period of intracultural research (Biesheuvel, 1952) to today's emphasis on the statistical evaluation of cross-cultural data (Berk, 1982), it has become evident that a number of important issues have not yet been resolved. It can be shown that there have been three main trends in interpreting cross-cultural research findings.

First, although it was realised that the test scores of different cultural groups could not be compared directly, tests were still needed for applied purposes in industry and education. Putting the question of data equivalence temporarily to one side, the problem of assessment when dealing with non-Western groups was handled pragmatically and testing was carried out intraculturally in applied settings (Biesheuvel, 1952). Other intracultural projects included factor analyses of tests results within one ethnic group (Murray, 1956) and decentered research (Irvine, 1966). In the latter case an answer was sought to the question "How (well) can we measure how they do their tricks?" (Wober, 1969, p. 488).

Second, a group of researchers placed emphasis on the differences between populations. In this tradition there have been several approaches. It started with the exploratory work of Rivers (1901, 1905) and the naive interpretation of test score differences as "real" or "true" in the early cross-cultural application of mental tests (Rowe, 1914). Jensen's (1969, 1973, 1981) and Eysenck's (1971, 1981) controversial interpretations of such differences as indicative of differences in the underlying genotype of populations also belong to this tradition. Thereafter followed the search for environmental factors responsible for these differences (Garth, 1921), and then the quest to demonstrate the principle of "radical cultural relativism" (Berry, 1972) as an explanation for such differences.

Finally, a group of researchers endeavoured to demonstrate similarities in test scores between populations. Whereas in the case of intracultural research as well as the studies emphasising differences between populations, the various approaches had very little in common, a clear developmental thread can be discerned in the search for cross-cultural similarities. In response to the differences found in test scores when mental tests were first applied across

cultures, this movement started with the search for a culture-free test (Goodenough, 1926; Porteus, 1924) and later a culture-fair test (Cattell, 1965; Davis & Eells, 1953). When this approach proved to have only limited success, the emphasis was shifted from culture-fair to equivalent tests (Frijda & Jahoda, 1966; Sears, 1961; Straus, 1969). After equivalent tests also failed to explain differences in test scores across cultures adequately, two new developments took place. In the first place the methodology of comparison was questioned. Measuring instruments were refined and new techniques for determining comparability of data spawned a number of studies to examine whether tests can be used justifiably across cultures (Berk, 1982; Osterlind, 1983). Secondly, the constructs measured were reconsidered, mainly with regard to the universals and specifics of attributes measured. The controversial "emic-etic" dichotomy (simplistically regarded as an analogy of cultural specifics and cultural universals) introduced into cross-cultural literature by Berry (1969) and the application of Generalizability Theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963) to cross-cultural problems (Van de Vijver & Poortinga, 1982) can be mentioned in

The ideas of Biesheuvel and his colleagues dominated South African cross-cultural research between the years 1946 and 1962 while he was the director of the National Institute for Personnel Research (NIPR). During this period psychological theories in South Africa developed not in parallel, but in opposition to the political climate in the country. This leads to a questioning of the thesis proposed here. Scientific theories do not always reflect the political climate in which they are developed. Sometimes other trends in thinking exist in opposition to the mainstream "Zeitgeist". Furthermore, the fact that an idea is different from the generally accepted norm does not necessarily mean that it is without its own bias. Even this historical outline is "ethnocentric". The quotations were carefully chosen to prove a point. An in-depth study of Andor's bibliographies will reveal that even in the early years some researchers voiced their doubts about the genetic explanation of the discrepancy in test scores between black and white subjects. Even today not all psychologists are convinced of the environmental explanation of these differences. However, this caution does not invalidate the argument that a clear developmental thread can be found in the history of cross-cultural studies: the blatantly racist

statements made in the early years would not be accepted in present day scientific journals.

Similar to the mistake of attributing causation when a correlation is found, a mistake can also be made by attributing causation to the "Zeitgeist". Scientific theories can just as well influence the political and social beliefs of the time, as vice versa. This places a real responsibility on the psychologist to avoid mistakes that can have negative consequences for anyone involved.

The fact that different ethnic groups obtain different scores on psychological tests, together with disagreement among researchers on the reasons for these differences, has caused the application of tests across cultures to become a problematic issue. This question is sometimes connected to the debate on genetic and environmental influences on test results and consequently has many emotional overtones. Before certain issues central to the question of intergroup test score interpretation is defined it is useful to establish what is meant by (dis)advantagement.

Here (dis)advantagement is defined as the differences between groups, independent of the construct of intent, which influence group measurements on that construct. (Dis)advantagement in testing is determined by the assets or handicaps groups of people experience in the testing situation. "Handicap" is not understood in the sense of being disabled, but in a relative sense as used in horse racing or in a game of golf. The same argument applies to "asset", namely, that it is subject to change.

When test results of an advantaged and a disadvantaged group are compared the following issues, which have come under the attention of test developers and users have to be distinguished: (a) ideas about the causes of discrepancies in test results between cultural groups, (b) the statistical detection of systematic error in test data between cultural groups, (c) how tests can be used in practice to select members of advantaged and disadvataged groups for employment and education, and (c) the benefit to the organisation arising from the use of tests for selection.

2. BIAS AT THE CONCEPTUAL LEVEL

The first distinction to be made is between a conceptual basis on which differences between groups are postulated before the test data are collected, and the statistical inference of the (in)comparability of data after the tests have been applied. The former will be referred to as conceptual bias, and the latter as (in)comparability or statistical bias.

In current cross-cultural literature the term "bias" has acquired many different meanings (Flaugher, 1978; Reynolds & Brown, 1984). Here the use of the term will be restricted to instances where measurement is influenced by group-related systematic error irrelevant to the construct or attribute under investigation.

When extraneous factors that are irrelevant to the construct being measured systematically influence test results for a certain group (or groups), these results are considered to be biased. Statistical bias is defined as constant or systematic error in test scores, as opposed to chance error. "Error" implies that the assets or handicaps of the group(s) are irrelevant to the measured construct.

Therefore bias has to do with intergroup differences in test performance that can be attributed to factors extraneous to the construct studied. A test is regarded as biased if any aspect of testing, which is irrelevant to what the construct is supposed to measure is reflected in the test scores of one or more groups. This means that a variable is regarded as biased if it inflates or suppresses the scores of certain subgroups on the measured construct. This holds true whether or not there are real differences in performance between the various groups in the absence of a biasing factor. In practice bias is usually taken to mean the suppression of test scores of disadvantaged ("minority") groups, but the general definition implies the decrease or increase in scores of any group.

Bias can be demonstrated, whereas the absence of bias never can be proven. As long as a good case can be made out for adverse elements in the testing situation the possibility of bias has to be acknowledged.

The sources of bias can be classified with reference to the "test takers", the "test", the interaction between "test taker" and "test", and "testing".

this regard. Progress in the measurement of cultural universals is not only a question of developing new methods, but more importantly, a realization that a debate on performance discrepancies, based mainly on level differences, is fruitless without clarity regarding the functional equivalence of tests across cultures.

It is often assumed that accepted scientific theories are based on facts; these facts being derived from astute observation and carefully conducted experiments, followed by critical and logical reasoning. However, an examination of the history of science shows that scientific thought is not independent of the prevailing "Zeitgeist". Scientific theories reflect the social, economic, political, and religious feelings of the time.

Psychology as a science is often guilty of this "ethnocentrism", approaching a problem from one particular point of view. The historic trends in testing across cultures outlined in this section bear this out. If psychological theory in general, as opposed to test score interpretation in particular, is considered this becomes even clearer. Summaries of

historic trends with particular reference to South Africa can be found in Kendall, Verster and Von Mollendorf (in press), Retief (in press), and Verster (in press). A few quotations from the two bibliographies compiled by Andor (1966, 1983) provide illustrations of this point.

> . . . the eastern and western coasts of Africa are inhabited by stupid and unenlightened hordes; . . . ferocious as their own congenial tigers, nor in any respect superior to these rapacious beasts in intellectual advancement but distinguished only by a rude and imperfect organ of speech . . . (Slavery no oppression, circa 1788, as quoted by Andor, 1966, p. 31).

> The Negro child is intellectually precocious up to puberty when a radical change takes place: his development stops suddenly or even slightly retrogresses. The white man's intellectual ability begins to broaden at the very moment when the Negro's reaches the stationary period (Cureau, 1915, as quoted by Andor, 1966, p. 37).

> . . . native children are retarded by at least five years at the age of fourteen . . . . Native children are not likely therefore to proceed with their education beyond European school standard II or III (Fick, 1934, as quoted by Andor, 1966, p. 59).

> Practical and scientific evidence does not warrant the conclusion that there are quantitative and qualitative differences between the potential intellectual and industrial capacities of European and African peoples. Studies of inter-racial differences are complicated and largely invalidated by ignoring differences in culture and social environment . . . (Biesheuvel, 1950, as quoted by Andor, 1966, p. 75).

> Cross-cultural analysis reveals that item difficulties change from culture to culture, and that test scores approach Western patterns as the groups adopt Western value systems . . . (Irvine, 1969, as quoted by Andor, 1983, p. 139).

The test takers. All variables associated with the person taking the test are included in this category, for instance, nutritional status, education, motivation, cognitive style, and test sophistication.

The test. The test is the instrument with which the psychologist has to measure constructs. Unfortunately tests are often biased with respect to different groups. Test language, test content, the meaning of symbols and the relevance of the construct measured form important sources of variation for the different target groups.

The interaction between test takers and test. The well-known study of French (1965) on cognitive styles can be used to illustrate how a single test can tap different aspects of cognitive behaviour in a group of test takers. French made use of a group of subjects who could be regarded as coming from a common culture, yet he was able to form two subgroups on the basis of the way in which they solved the problems: some used an "analytical" style, whereas others used a more "global" style. Tests purported to measure so called "general intelligence", in which the items require mainly, say, analytical methods to solve problems, will

favour those test takers who normally utilise this style, etc.

Testing. Much has been written about the experimenter-subject relationship. The effect of characteristics such as age, sex, and race of the experimenter on the test scores have been investigated, invariably with contradictory findings (Garcia & Zimmerman, 1972; Jensen, 1974). It has also been said that experimenter expectation has the effect of a self-fulfilling prophesy, especially in the case of young children (Rosenthal & Jacobsen, 1968). Furthermore, experimenter bias can have a direct effect on a test taker's score, particularly when the evaluation has an element of subjectivity. Other situational variables include the test instructions (Crawford-Nutt, 1976; Godsell, 1976, 1979), computerised versus manual environments, test atmosphere, and test format (Johnson & Mihal, 1973; Samuel, 1977; Williams, Davis, Anderson & Favor, 1978).

Bias at the conceptual level can be determined using experimental or judgemental methods. When using experimental methods, the researcher has an hypothesis

beforehand about those aspects of testing that cause inequality in test results between groups. A "treatment effect" is built into an experimental design with the ethnic groups in order to test the hypothesis. Two of the most obvious variables that can be investigated in this way are test language and test content. In two studies on black and white Americans, Schmeiser (1982) varied the test content between the black and white groups using materials reflecting black or white culture. Although the results of this experiment were inconclusive, the method does warrant further attention, possibly in populations where the cultural differences are larger than between black and white Americans.

Conceptual bias is also detected by giving the tests to experts to evaluate the items. These judgemental methods are in their infancy, relying almost totally on subjective evaluation. Hilliard III (1984) has pointed out that cultural linguists and anthropologists are in general not consulted when judgements about bias are made. Only in the case of equivalent translations has the researcher advanced techniques available, such as back translation and decentering. In the case of back-translation the first translator translates the

text from the source to the target language. The second translator, who does not see the original, uses the translated text to translate back to the source language. The two versions are then compared to see if there are any discrepancies. If necessary, the procedure is repeated. Decentering refers to the condition where the source and target languages are considered of equal importance and changes are made to both to ensure equivalence. This is in contrast to the practice of keeping the original version of the source language unchanged and making all the changes to the target language only (Brislin, Lonner & Thorndike, 1973; Werner & Campbell, 1973).

Judgemental methods are important with regard to the public acceptance of tests, especially in the case of minority groups, for whom acceptance often depends on whether the tests are perceived to be biased or not.

Both judgemental and experimental methods are based on ideas formed before the data are collected. In the next section methods are discussed that are used after the data have been gathered.

3. COMPARABILITY: BIAS AT THE STATISTICAL LEVEL

Having discussed the conceptual methods of determining test bias, the next group of methods are those set to detect differences between groups by analysing the results after the data have been collected. These methods are not based on hypotheses about the causes of the score differences but rely solely on statistical techniques.

They were initiated in work such as that of Irvine (1966) who correlated item difficulties for Raven's Progressive Matrices between British school children and African pupils. The resulting rank correlation (0,695) was used as an index of the "comparative validity" of the test. Other researchers soon followed suit and objective indications of the identity of scale properties of the psychological measures themselves were sought. These indications took the form of psychometric checks, mainly on the item parameters of the test.

These statistical checks on score distributions, previously carried out without properly substantiated reasons, were placed in a test-theoretic framework by

Poortinga (1971) who used the term "comparability". He argued that before test scores can be compared across cultural groups the researcher must establish whether the test measures qualitatively and quantitatively the same attributes of behaviour in the (two) populations. No notions regarding the nature of cross-cultural differences are attached to the concept of comparability; it is a psychometric concept and can be analysed in terms of various statistically testable conditions. Statistical controls are now widely applied in cross-cultural psychological studies (Berk, 1982; Osterlind, 1983).

Stated in this way bias and incomparability are compatible; both concepts refer to the same issue. The only difference being that bias is defined in terms of the factors that have an undesirable effect on the test scores, whereas (in)comparability is defined in terms of the (dis)similarity of test scores. Bias is postulated <u>à priori</u> on a conceptual level, whereas incomparability of test results is demonstrated by statistical methods.

Comparability (or more precisely incomparability) therefore can be seen as the psychometric counterpart of bias. Consequently, (in)comparability can also be called bias at the statistical level, although it has to be remembered that comparability also implies equivalence of the attribute measured. A problem in this regard is that the two approaches for determining equivalence, namely conceptual methods and statistical methods do not necessarily yield the same results. The items judged by experts to be biased, are not always the same as the items identified by various statistical procedures as being incomparable (Burrill, 1982). This can be explained by the inadequacy of both methods. In the case of judgemental methods, Hilliard III (1984) points out that generally the expertise of cultural linguists and anthropologists is not consulted when judgements are made about the differences in test scores of different groups.

Analogous to the four widely recognised levels of measurement, namely, nominal, ordinal, interval, and ratio scales, Van de Vijver and Poortinga (1982) propose four levels on which psychological universals can be represented along a dimension of experimental

rigour, namely, conceptual universals, functionally equivalent, metrically equivalent, and scalar equivalent universals. Conceptual equivalence refers to constructs at a high level of abstraction, that are not operationally defined. It is not possible to make empirical comparisons of conceptual universals across cultures.

For each of the other categories conditions of scale identity are imposed on the data to investigate comparability. However, comparability can never be proven, only a lack of comparability can be demonstrated if certain conditions are not met. These conditions have to be specified beforehand, and if a single relevant condition is not met, the test in question is not regarded as comparable. Criticism of Poortinga's definition is generally directed at the severity of these requirements (Van der Flier & Drenth, 1980). However Poortinga (1983) points out that the choice of conditions has to be guided by theoretical considerations; only relevant statistical tests are appropriate tests.

> Essential improvement in our understanding and explanation of cross cultural differences requires first and foremost theories which enable us to make very specific statements about the (expected) interrelationships between phenomena. As long as the desired theories are not available, psychometric techniques to analyze comparability deserve serious attention (Poortinga, 1983, p. 251).

For the actual analysis of comparability several requirements have been proposed. The techniques will be classified under the headings functional equivalence, metric equivalence, scalar equivalence, and item equivalence.

3.1 Qualitative or functional equivalence

When a test satisfies conditions of qualitative scale identity it is regarded as functionally equivalent. Concepts regarded as functionally equivalent are generalizable in a qualitative sense, but not necessarily in a quantitative sense. An analogy is the measurement of temperature with a Celsius as compared to a Fahrenheit scale (Van de Vijver & Poortinga, 1982). Techniques used to investigate this definition

are similar to those used for construct validation. In fact construct validation and the analysis of comparability are sometimes identified with each other (Reynolds, 1982). But although the concepts have much in common, it is important to distinguish between them: "In an analysis of comparability the main question is whether the same construct is being measured rather than which construct" (Poortinga, 1983, pp. 245-246).

Functional equivalence is generally analysed by means of correlational techniques such as similarity of patterns of intercorrelations (Poortinga, 1971; Reynolds, 1982) and similarity of factor structures (Jensen, 1980; Poortinga, 1983; Reynolds, 1982), but other methods are possible, for instance confirmatory maximum likelihood factor analysis based on covariance rather than correlation matrices (Poortinga & Foden, 1975; Poortinga, 1983).

3.2 Metric equivalence

In order to be metrically equivalent, test scores should represent the same metric or unit of measurement across cultures, but could have a different origin in each culture. Measurement on a Celsius and Kelvin scale can serve as an analogy. The instrument measures

correctly within the cultures, but no intercultural comparison can be made (Van de Vijver & Poortinga, 1982). In order to analyse metric equivalence, the regression of test scores from subjects in different populations on a common criterion can be studied. This procedure has to be considered carefully, however, the problem being that the group mean is used for estimating the predicted score of a person (Poortinga, & Foden, 1975). For example, when two variables measure the same construct, there is as much reason to stipulate equality of regression lines for different populations of the first variable on the second as vice versa. However, when the regression lines are the same in both cases, the mean scores are also equal (Van de Vijver & Poortinga, 1982; Poortinga, 1983). In 1982 Van de Vijver and Poortinga proposed use of Generalizability Theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963) as a coherent framework for the analysis of comparability.

3.3 Scalar equivalence

Measures considered to be scalar equivalent should not only have the same metric across cultures, but also the same origin. In practice scalar equivalence nearly

always implies distributional identity across cultures. At this stage of cross-cultural comparison strict universals can be found for only a few concepts: speed of processing of simple visual and auditory stimuli being possible examples (Van de Vijver & Poortinga, 1982). But even with these simple measurements there are doubts about this claim.

3.4 Item equivalence

A test is considered to be item equivalent if the items of that test, taken as separate measurements, satisfy the requirements of metric equivalence. A separate category of item equivalence is distinguished to investigate whether items have a common bias with regard to the underlying construct (Poortinga, 1983). A distinction between unconditional item bias detection methods and item bias detection methods conditional on the subject's ability level was made by Mellenbergh (1981). He defined conditional methods as follows:

> An item is unbiased if the probability of a correct response is the same for all subjects with the same ability level (Mellenbergh, 1981, p. 294).

Methods to detect differential item performance conditional on the underlying ability level are, for example, item characteristic curve methods and methods based on contingency tables, such as chi-square and logit methods.

Statistical procedures to examine the psychometric properties of deviant test items unconditional of ability level, include analysis of variance, transformed item difficulties, and distractor response analysis.

4. FAIRNESS: INEQUITIES IN THE PROCESS OF SELECTION

The term "culture fair" was first introduced into cross-cultural psychological literature to refer to tests in which cultural variables, such as language and specific content have been reduced to a minimum. Other terms that have related meanings are "culture free" and "culture reduced" (Jensen, 1980). Fairness is also a non-psychological undefined term as used by laymen, namely, just or equitable. Finally, fairness may refer to interculturally fair decisions on selection. The latter interpretation of fairness is discussed in this section. The concept of selection fairness in

cross-cultural testing can be traced back to Thorndike (1971) who was the first to make the distinction between the efficiency of the test as a measuring instrument (bias and comparability) and the adequacy of decisions based on test results. The latter concerns test fairness. By 1976 this distinction was so well accepted by psychologists and educationalists in the United States that an entire issue of the Journal of Educational Measurement was devoted to test fairness. A number of "models of fairness" were in use with no general agreement as to the applicability of each. In order to distinguish between this narrowly defined meaning of fairness and other interpretations of the term, it might be worthwhile to conceive of fairness as authoritative selection.

Given a bivariate distribution of test on criterion for two or more ethnic groups, the main characteristic of a model of fairness (authoritative selection) is that it is a rule applied when tests are used for selection, primarily for employment purposes, or for admission to educational institutions. A selection rule is considered fair or not by the impact it has on the candidates, the particular concern being about the lower scoring (disadvantaged) group.

Initially this procedure was regarded as appropriate. Soon, however, it became clear that the models had to be evaluated not only on a technical level, but also on a philosophical level. If selection fairness (authoritative selection) were to be considered an absolute quality, the term fairness would be a misnomer; inherent in each decision rule is a value judgement or philosophy about acceptance and rejection of potentially successful candidates in different ethnic groups, or about the desirability of maximising achievement and productivity. Examples of such philosophies are: equality of opportunity for all individuals, achievement or productivity maximization, or extending preferential treatment to groups disadvantaged by past discrimination (Anastasi, 1982).

According to the philosophy of group parity, the test score deficit for members of certain ethnic groups is due to their being disadvantaged. This gap can be bridged only by taking positive steps such as the introduction of educational enrichment programmes to provide disadvantaged groups with the opportunity to work in certain crafts and professions, together with special efforts to recruit members of disadvantaged groups, and policies such as numeric goals and quotas,

that will allow more members of disadvantaged groups to be accepted into educational or employment situations in which places are limited. Extending preferential treatment to groups disadvantaged by past discrimination is sometimes also referred to as affirmative action.

Hunter and Schmidt (1976) make a distinction between unqualified individualism

> . . . this means that an organization should use whatever information it possesses to make a scientifically valid prediction of each individual's performance and always select those with the highest predicted performance (Hunter & Schmidt, 1976, p. 1053).

and qualified individualism

> . . . this [is] an ethical imperative to refuse to use race, sex, and so on, as a predictor even if it were in fact scientifically valid to do so (Hunter & Schmidt, 1976, p. 1054).

On first examination it seems as if group parity and unqualified individualism are essentially similar

philosophies. If, however, a test predicts differently for members of different groups, the proponent of qualified individualism is in a quandary: he cannot use group membership to improve decision making and therefore has to rely on strategies which give under certain conditions unfavourable outcomes to exactly those people he wants to protect (Jensen, 1980).

5. UTILITY: THE OUTCOME (PRACTICAL EFFECTIVENESS) OF TESTS IN SELECTION

Bias, comparability, and fairness (authoritative selection) have application in a cross-cultural context only. The concept of utility, although originating in industrial psychology, can be applied cross-culturally (Cole, 1973; Hunter, Schmidt, & Rauschenberger, 1977).

5.1 Utility according to industrial psychologists

The problem of selection utility was first treated by Taylor and Russel (1939) in measuring the "practical effectiveness of tests in selection" (p. 565). More contributions followed until the publication in 1957 of Cronbach and Gleser's book on personnel decisions based on tests.

The question of test utility (sometimes also referred to as selection utility) usually applies to prediction and selection problems in industry, but can be applied to problems in education as well. The industrial psychologists working in the earlier years (Berkson, 1947; Brogden, 1946; McClelland, 1942; and Taylor & Russell, 1939) based their test's utility models on objective indices. Criterion related yardsticks, for example job performance or university examinations, determine the utility of selection decisions in output or probability of success. For instance, a test is useful to an organization if the mean criterion performance of test-selected individuals is higher than the mean criterion performance of unselected individuals. Similarly, a test is useful if the probability of success of test-selected individuals is higher than the probability of success of unselected individuals. By "unselected" candidates is meant candidates selected by existing practices, such as the interview, or criteria based on qualifications, previous achievements, testimonials, and references. It can even refer to selection by testing, the assumption being that the new test replacing the existing ones has a higher validity.

The utility of a test is the extent to which using that test as a method of selection constitutes an actual gain to management over alternative selection procedures previously used, expressed in cost-accounting terms.

5.2 Utility according to cross-cultural psychologists

Since the early 1970s utility has come under fresh scrutiny by psychologists. Not only were the older models of utility reassessed in an intracultural context (see Schmidt & Hoffmann, 1973), but the utilities of outcomes under various fairness models (i.e., in a cross-cultural context) were also calculated (Hunter, Schmidt, & Rauschenberger, 1977; Gross & Su, 1975; Petersen & Novick, 1976).

Utility applied to cross-cultural problems involves information from all ethnic groups. In this context utility refers to the practical significance or payoff of selection decisions including information from members of all groups. Models of utility are traditionally applied to industrial problems. The outcome of such decisions can be calculated and is stated in cost-effectiveness terms. For industrial

organisations this is usually expressed in terms of money, but other standards are possible, for example pass rates at educational or training institutions. Utility, therefore, is the outcome of criterion related decisions. Utility can be analysed irrespective of how the test is evaluated in terms of bias, comparability, or fairness (authoritative selection). The choice of a utility model is usually pragmatic in that it has to satisfy the requirements of the rational-economic manager as opposed to those of the applicants or society.

5.3 Expected utility according to statistical economists

Statistical economists ascribe a different meaning to utility than that used in industrial or cross-cultural psychology. In decision theoretic terms a value (or weight) is stated quantitatively and assigned to each possible outcome. Expected utility is the product of this value and the probability of the outcome summed over all possibilities. The value is usually an economic yardstick but it can also be different weights (Cronbach, 1976; Gross & Su, 1975; Petersen & Novick, 1976; Sawyer, Cole, & Cole, 1976). For example, Cronbach (1976) assigns 0 - 1 values to each of the

four possibilities of the select-reject, success-fail matrix. A weight of 1 can be assigned either to being selected or only if being selected is associated with being successful as well. This implies not only that utilities are not "real" in the sense of representing criterion yardsticks any more, but also that they now represent the interest of each of the three parties, namely, the employer, the individual applicant or the group. Weights assigned to "success" represent the interest of the organization, to "select" the interests of the applicants, and higher weights to "select" in some groups compared to other groups, the interests of these groups. An example of different weights is found in the work of Gross and Su (1975), where they assign utilities of 2, -3, 1, and -1 to the advantaged group and 3, -5, 1, and -1 to the disadvantaged group for the four categories select-success, reject-success, reject-fail and select-fail.

6. OVERVIEW: IMPLICATIONS OF THE ISSUES OF BIAS, COMPARABILITY, FAIRNESS, AND UTILITY IN THE SOUTH AFRICAN CONTEXT

The task of evaluating and integrating the diverse concepts covered in the previous sections is no easy one. Not only are various branches of psychology involved, but also different disciplines, for instance, industrial psychology, psychometrics, political philosophy, and educational measurement. Furthermore, psychological testing across cultures is developing rapidly. New theories and hypotheses are being created and tested. Final answers may never be found. Nonetheless, in a post-Rubicon South Africa in search of a new dispensation for different cultural groups, it is politically and economically necessary to make informed judgements. At present the researcher can provide the practitioner with knowledge about what questions to ask and suggest approaches to some of the solutions.

6.1 Bias

The interpretation of test results is related to the environmentalist-geneticist controversy. Much of the displeasure of black people with tests can be attributed to "overinterpretation" bias (Flaugher,

1978). This bias occurs when an individual is judged on a measure of, say, "general intelligence", when the test used measures but a specific part of the broad spectrum of human competence. The definition of bias proposed in section two is an attempt to eliminate this form of bias. Bias is defined as any aspect of the testing situation which is irrelevant to the construct measured but which influences the test results systematically. The more broadly a construct is defined, the greater the possibility of overinterpretation bias. Constructs that are operationally defined are much less subject to this kind of bias.

The present situation with regard to the detection of biased items in tests is unsatisfactory. Hambleton and Rogers (1986) draw attention to the fact that the item bias detection methods presently used are inappropriate: items presenting cultural stereotypes instead of deviant items are identified.

Eliminating these items from a test does, however, serve a useful purpose. Shepard (1982) lists three reasons for subjecting test items to subjective reviews:

1. Any aspect of testing that is offensive to any group, should be eliminated as a matter of principle.

2. An item may not be detected by psychometric methods as being deviant, but may still have a delayed negative influence on subsequent items.

3. Subjective reviews of test items may generate hypotheses about the nature of bias.

A fourth reason can be added to the three mentioned above that:

4. Items that are subjectively experienced as biased may lead to the impression that tests discriminate against a particular group, this in turn can leading to dissatisfaction with tests.

6.2 Comparability

The exploration of properties of deviant items from a purely psychometric point of view is perhaps the area of cross-cultural research that has developed most in recent years. Osterlind (1983) and Jensen (1984) remark on the very recent dates of the references in the deviant item detection literature. Yet, in spite of prolific publications on this topic, the detection methods suffer from a number of shortcomings, for example: in one-parameter latent trait models,

model-data fit problems are confounded with deviant item detection, and also, when significance tests are used the size of examinee groups confounds the results (Hambleton & Rogers, 1986). The methods of distinguishing between group differences attributable either to the instrument (`bias') or to the subjects (`real differences') are more complicated than usually suspected (Van de Vijver, 1986). Most statistical procedures to detect incomparability of items are "relative" methods in that they are dependent on a particular pool of items and identify those items that do not "fit in", and finally, it is very hard to eliminate group differences in ability as a confounding factor (Burrill, 1982).

6.3 Fairness

In a previous section the different models of fairness (models of authoritative selection) have been associated with different philosophies regarding testing across cultures.

6.3.1 The moratorium philosophy

Anti-test movements argue in particular against unfairness of test-based decisions. There are, however, a number of arguments in favour of testing:

1. Open admission is impractical, because employers and educational institutions can only accommodate a limited number of new intakes. Some form of selection has to take place.

2. Alternatives to testing are loaded with inequalities. School examinations and personal recommendations are even more dependent on socially undesirable factors and favouritism than tests. Groups previously discriminated against do not have access to schools that prepare them specifically for entrance examinations, neither do they know many influential people to give them the necessary recommendation for available positions.

3. Evaluation that cannot be accounted for by objective indices, is open to corruption.

Tests, then, represent an attempt, albeit an imperfect one to counteract this trend in selection strategy.

6.3.2 The autonomy philosophy

Decentered research still has a place in cross-cultural psychology in order to identify skills that have never been developed in Western civilization. The best results are achieved when incorruptibleness methodological techniques are used. For instance,

Klich (1986) drew on the expertise of ethnographers and anthropologists in the study of contextual variables that mediate human behaviour. Specific areas of inquiry in this study were the route finding and geographical orientation skills of Australian Aboriginal people.

Thus, although decentered research is important, test taking across cultures needs another approach. Tests designed for use in only one culture sidestep the problem of equivalence. This use of tests, although successfully applied by Biesheuvel and Hudson (1949) when he constructed the General Adaptability Battery (GAB) for classifying black mining recruits, is unsuitable for the present problem of finding an equitable solution for black and white people in the mainstream of society. Questions that need to be addressed are how to advance black people in an organization or how to choose the cutoff points on a predictor when applicants from different population groups apply for the same position.

6.3.3 Group parity models

The most important models falling under this label are those based on quotas and on ratios.

6.3.3.1 The quota models

These models are based on the ratio of the number of selectees to (a) the proportion of that group (ethnic, race, sex, or socioeconomic) in the total population; or to (b) the number of applicants. The first option is unrealistic because not all members of a certain group are in the workforce, for example, many married women may choose not to work. Neither are all those who are in the workforce suitable for a specific position. In India where quotas for government jobs and college places are reserved for the untouchables and tribal peoples in proportion to their numbers in the population, only seven percent of university places are taken up by these groups as compared to the fifteen percent of places set aside. This is ascribed to "poverty, the lack of education and the use of child labour" ("India's Castes," 1986, p. 54). Therefore, the second option is the better one, that quotas should be set according to the number of people applying for the position. However, since the applicant pool depends on education and training as well as recruitment the long term effects of these inputs should be kept in mind.

Gross, A. L. & Su, W. (1975). Defining a "fair" or "unbiased" selection model: A question of utilities. Journal of Applied Psychology, 60, 345-351.

Hambleton, R. K. & Rogers, H. J. (1986). Approaches for identifying and understanding bias in test items. In S. E. Newstead, S. H. Irvine, & P. L. Dann (Eds.), Human assessment: Cognition and motivation (pp. 398-399). Dordrecht: Nijhoff.

Hilliard III, A. G. (1984). IQ testing as the emperor's new clothes: A critique of Jensen's Bias in Mental Testing. In C. R. Reynolds & R. T. Brown (Eds.), Perspectives on bias in mental testing (pp. 139-169). New York: Plenum Press.

Humphreys, L. G. (1973). Statistical definitions of test validity for minority groups. Journal of Applied Psychology, 58, 1-4.

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.

Garcia, A. B. & Zimmerman, B. J. (1972). The effect of examiner ethnicity and language on the performance of bilingual Mexican American first graders. Journal of School Psychology, 87, 3-11.

Garth, T. R. (1921). White, Indian and Negro work curves. The Journal of Applied Psychology, 5, 14-25.

Godsell, G. (1976). Cross-cultural differences in cognitive flexibility. Part one: Theories and hypotheses for proposed research (Special Report PERS 242). Johannesburg: National Institute for Personnel Research, Council for Scientific and Industrial Research.

Godsell, G. (1979). An investigation into cognitive flexibility among educated black and white groups (Special Report PERS 272). Johannesburg: Council for Scientific and Industrial Research, National Institute for Personnel Research.

Goodenough, F. L. (1926). Measurement of intelligence by drawings. Terrytown-on-Hudson, NY: World Book Company.

Eysenck, H. J. (1971). Race, intelligence and education. London: Temple Smith.

Eysenck, H. J. (1981). H. J. Eysenck. In S. Raby, (Ed.), Intelligence: The battle for the mind. H. J. Eysenck versus Leon Kamin (pp. 11-83). London: MacMillan.

Flaugher, R. L. (1978). The many definitions of test bias. American Psychologist, 33, 671-679.

Franks, P. E., Ngwane, T. L. S., & Rheeder, S. (1986). Preliminary survey of human relations in a retail organization (Contract Report C/PERS-372). Pretoria: Human Sciences Research Council, National Institute for Personnel.

French, J. W. (1965). The relationship of problem-solving styles to the factor composition of tests. Educational and Psychological Measurement, 25, 9-28.

Frijda, N. & Jahoda, G. (1966). On the scope and methods of cross-cultural research. International Journal of Psychology, 1, 109-127.

Crawford-Nutt, D. H. (1976). Are black scores on Raven's standard progressive matrices an artifact of method of test presentation? Psychologia Africana, 16, 201-206.

Cronbach, L. J. & Gleser, G. C. (1957). Psychological tests and personnel decisions. Urbana: University of Illinois Press.

Cronbach, L. J. (1976). Equity in selection: Where psychometrics and political philosophy meet. Journal of Educational Measurement, 13, 31-41.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements. New York: John Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. British Journal of Mathematical and Statistical Psychology, 16, 137-163.

Davis A. & Eells, K. (1953). Davis-Eells games: Manual. Yonkers-on-Hudson: World Books.

Brogden, H. E. (1946). On the interpretation of the correlation coefficient as a measure of predictive efficiency. Journal of Educational Psychology, 37, 65-76.

Burrill, L. E. (1982). Comparative studies of item bias methods. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 161-179). Baltimore, MD: John Hopkins.

Cattell, R. B. (1965). The IPAT culture fair intelligence scales 1, 2 and 3 (3rd ed.). Champaign, IL: Institute for Personality and Ability Testing.

Chemel, C. S. (1985). Fairness in selection with an emphasis on fairness in testing. Unpublished thesis in part fulfilment for the M.B.L., University of South Africa, Pretoria.

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

Berry, J. W. (1969). On cross-cultural comparability. International Journal of Psychology, 4, 119-128.

Berry, J. W. (1972). Radical cultural relativism and the concept of intelligence. In L. J. Cronbach & P. J. D. Drenth (Eds.), Mental tests and cultural adaptation (pp. 77-88). The Hague: Mouton.

Biesheuvel, S. & Hudson, W. (1949). Aptitude tests for native labour on the Witwatwersrand gold mines: Parts I and II. Johannesburg: South African Council for Scientific and Industrial Research, National Institute for Personnel Research.

Biesheuvel, S. (1952). The study of African ability. African Studies, 11, 45-58, 105-117.

Brislin, R. W., Lonner, W. J., & Thorndike, R. M. (1973). Cross-cultural research methods. New York: John Wiley.

REFERENCES

Anastasi, A. (1982). Psychological testing: Fifth edition. New York: MacMillan.

Andor, L. E. (Compiler). (1966). Aptitudes and abilities of the Black man in Sub-Saharan Africa 1784-1963: An annotated bibliography. Johannesburg: South African Council for Scientific and Industrial Research, National Institute for Personnel Research.

Andor, L. E. (Compiler). (1983). Psychological and sociological studies of the black people of Africa, South of the Sahara 1960-1975: An annotated select bibliography. Johannesburg: Council for Scientific and Industrial Research, National Institute for Personnel Research.

Berk, R. A. (Ed.). (1982). Handbook of methods for detecting test bias. Baltimore, MD: John Hopkins.

Berkson, J. (1947). "Cost-utility" as a measure of the efficiency of a test. Journal of the American Statistical Association, 42, 246-255.

Thus, no hard and fast rules can be set, the "best" policy can only be decided upon by taking the real meaning of the criterion and the selection ratio (number of applicants to number of candidates) into account. Furthermore, selection policies always must be evaluated and updated against the background of education, training, and recruitment.

principles, where merit is associated with performance on the criterion as opposed to performance on the test.

3. Group parity: If the selection ratio is favourable (i.e. only a few employees have to be selected out of a large number of applicants), it is likely that many applicants that would have been successful have to be rejected. In such a case a kind of two-stage (merit based/group based) selection procedure can be applied. In the first stage all applicants who are likely to succeed are identified, whereas group membership is taken into account in the second stage. This would be an example of affirmative action, i.e. group-based decisions on selection and promotion to remedy the effects of prior discrimination, but taking merit also into account.

4. Quotas: Rigid quotas can be set, even if in contradiction to test scores. This strategy is an example of reverse discrimination and is sometimes referred to as "puppeteering" or "tokenism" and often means compromising standards and setting people up for failure.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. M. (1977). Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring. Journal of Applied Psychology, 62, 245-260.

India's castes: Old fissures in a new landscape. (1986, September 6). The Economist, p. 54.

Irvine, S. H. (1966). Towards a rationale for testing attainments and abilities in Africa. The British Journal of Educational Psychology, 36, 24-32.

Irvine, S. H. (1969). Figural tests of reasoning in Africa. Studies in the use of Raven's Matrices across cultures. International Journal of Psychology, 4, 217-228.

Jensen, A. R. (1969). How much can we boost I.Q. and scholastic achievement? Harvard Educational Review, 39, 1-123.

Jensen, A. R. (1973). Educational differences. London: Methuen.

Jensen, A. R. (1974). Effect of race of examiner on the mental test of black and white pupils. Journal of Educational Measurement, 11, 1-14.

Jensen, A. J. (1980). Bias in mental testing. London: Methuen.

Jensen, A. R. (1981). Straight talk about mental tests. New York: The Free Press.

Jensen, A. R. (1984). Test bias: Concepts and criticisms. In C. R. Reynolds, & R. T. Brown (Eds.), Perspectives on bias in mental tests (pp. 507-586). New York: Plenum Press.

Johnson, D. F. & Mihal, W. L. (1973). Performance of blacks and whites in computerized versus manual testing environments. American Psychologist, 28, 694-699.

Kendall, I. M., Verster, M. A., & Von Mollendorf, J. W. (in press). Test performance of blacks in Southern Africa. In J. W. Berry & S. H. Irvine (Eds.), Human ability in cultural context. Cambridge: Cambridge University Press.

Klich, Z. L. (1986). Australian Aboriginal cognition in context. In S. E. Newstead, S. H. Irvine, & P. L. Dann (Eds.), Human assessment: Cognition and motivation (pp. 75-79). Dordrecht: Nijhoff.

Mellenbergh, G. H. (1981). Conditional item bias methods. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors (pp. 293-302). New York: Plenum Press.

Mercer, J. R. (1984). What is a racially and culturally nondiscriminatory test? A sociological and pluralistic perspective. In C. R. Reynolds & R. T. Brown (Eds.), Perspective on bias in mental testing (pp. 293-356). New York: Plenum Press.

McClelland, W. (1942). Selection for secondary education. London: University of London Press.

Murray, C. O. (1956). The structure of African intelligence: A factorial study of the abilities of Africans. Unpublished master's thesis, University of Natal, Durban.

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

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

Poortinga, Y. H. (1971). Cross-cultural comparison of maximum performance tests: Some methodological aspects and some experiments with simple auditory and visual stimuli. Psychologia Africana Monograph Supplement, 6.

Poortinga, Y. H. (1983). Psychometric approaches to intergroup comparison: The problem of equivalence. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors (pp. 237-257). New York: Plenum Press.

Poortinga, Y. H. & Foden, B. I. M. (1975). A comparative study of curiosity in black and white South African students. Psychologia Africana Monograph Supplement, 8.

Porteus, S. D. (1924). Guide to Porteus Maze test. Vineland, NY: The Training School.

Reynolds, C. R. (1982). Methods for detecting construct and predictive bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 199-227). Baltimore, MD: John Hopkins.

Reynolds, C. R. & Brown, R. T. (1984). Bias in mental testing: An introduction to the issues. In C. R. Reynolds, & R. T. Brown (Eds.), Perspectives on bias in mental testing (pp. 1-39). New York: Plenum Press.

Rivers, W. (1901). Introduction and vision. In A. C. Haddon (Ed.), Reports of the Cambridge anthropological expedition to the Torres Straights (Vol. 2, Pt. 1). Cambridge: Cambridge University Press.

Rivers, W. (1905). Observations on the senses of the Todas. British Journal of Psychology, 1, 321-396.

Retief, A. L. (in press). Theoretical and methodological problems in cross-cultural psychological assessment. Pretoria: Human Sciences Research Council.

Rosenthal, R. & Jacobsen, L. (1968). Pygmalion in the classroom. New York: Holt, Rinehart, & Winston.

Rowe, E. C. (1914). Five hundred forty-seven white and two hundred sixty-eight Indian children tested by Binet-Simon tests. The Pedagogical Seminary and Journal of Genetic Psychology, 21, 454-468.

Sawyer, R. L., Cole, N. S., & Cole, J. W. L. (1976). Utilities and the issue of fairness in a decision theoretic model for selection. Journal of Educational Measurement, 13, 59-76.

Samuel, W. (1977). Observed IQ as a function of test atmosphere, tester expectation, and race of tester: A replication for female subjects. Journal of Educational Measurement, 69, 593-604.

Schmeiser, C. B. (1982). Use of experimental design in statistical item bias studies. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 64-95). Baltimore, MD: John Hopkins.

Schmidt, F. L. & Hoffmann, B. (1973). Empirical comparison of three methods of assessing utility of a selection device. Journal of Industrial and Organizational Psychology, 1, 13-22.

Sears, R. R. (1961). Transcultural variables and conceptual equivalence. In B. Kaplan (Ed.), Studying personality cross-culturally (pp. 445-455). Evanston, IL: Row Peterson.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 9-30). Baltimore, MD: John Hopkins.

Shuey, A. M. (1969). The testing of Negro intelligence (2nd ed.). New York: Social Science Press.

Sprigle, J. E. & Schaefer, L. (1985). Longitudinal evaluation of the effects of two compensatory preschool programs on fourth- through sixth-grade students. Developmental Psychology, 21, 702-708.

Straus, M. A. (1969). Phenomenal identity and conceptual equivalence of measurement in cross-national comparative research. Journal of Marriage and the Family, 31, 233-239.

Taylor, T. R. (1987). The future of cognitive assessment (Special Report PERS 420). Pretoria: Human Sciences Research Council, National Institute for Personnel Research.

Taylor, H. C. & Russell, J. T. (1939). The relationship of validity coefficients in the practical effectiveness of tests in selection: Tables and discussions. Journal of Applied Psychology, 23, 565-578.

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

Van de Vijver, F. J. R. (1986). Group differences in structured tests. In S. E. Newstead, S. H. Irvine, & P. L. Dann (Eds.), Human assessment: Cognition and motivation (pp. 399-400). Dordrecht: Nijhoff.

Van de Vijver, F. J. R. & Poortinga, Y. H. (1982). Cross-Cultural generalization and universality. Journal of Cross-Cultural Psychology, 13, 387-408.

Van den Berg, A. R. (in press). Using the Junior South African Individual Scales (JSAIS) (1979) for test takers from South African population groups who were not included in the norm population. Paper presented at the annual test liaison meeting held at the National Institute for Personnel Research, Johannesburg, April 2, 1986.

Van der Flier, H. & Drenth, P. J. D. (1980). Fair selection and comparability of test scores. In L. J. T. van der Kamp, W. F. Langerak, & D. N. M. de Guijter (Eds.), Psychometrics for educational debates (pp. 85-101). New York: Wiley.

Verster, J. M. (in press). Cross-cultural cognitive research: Some methodological problems and prospects. In K. F. Mauer & A. I. Retief (Eds.), Psychology in context: Cross-cultural research trends in South Africa. Pretoria: Human Sciences Research Council.

Verster, J. M. & Prinsloo, R. J. (in press). Test performance of English-speaking and Afrikaans-speaking South Africans. In J. W. Berry & S. H. Irvine (Eds.), Human abilities in cultural context. Cambridge: Cambridge University Press.

Werner, O. & Campbell, D. T. (1973). Translating, working through interpreters, and the problem of decentering. In R. Naroll & R. Cohen (Eds.), A handbook of method in cultural anthropology (pp. 398-420). New York: Columbia University Press.

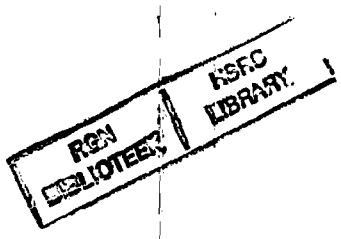
Williams, R. L., Davis, W., Anderson, P., & Favor, K. (1978). Test format as a form of bias for black students. Journal of Non-white Concerns in Personnel and Guidance, 6, 141-147.

Wober, M. (1969). Distinguishing centri-cultural from cross-cultural tests and research. Perceptual and Motor Skills, 28, 488.

BRN. 409694

ORDER No. 98.7698

Copy No. PB99039



R15.00

6.3.3.2 The ratio models

The various ratio models are all based on the four categories indicated in the following figure:

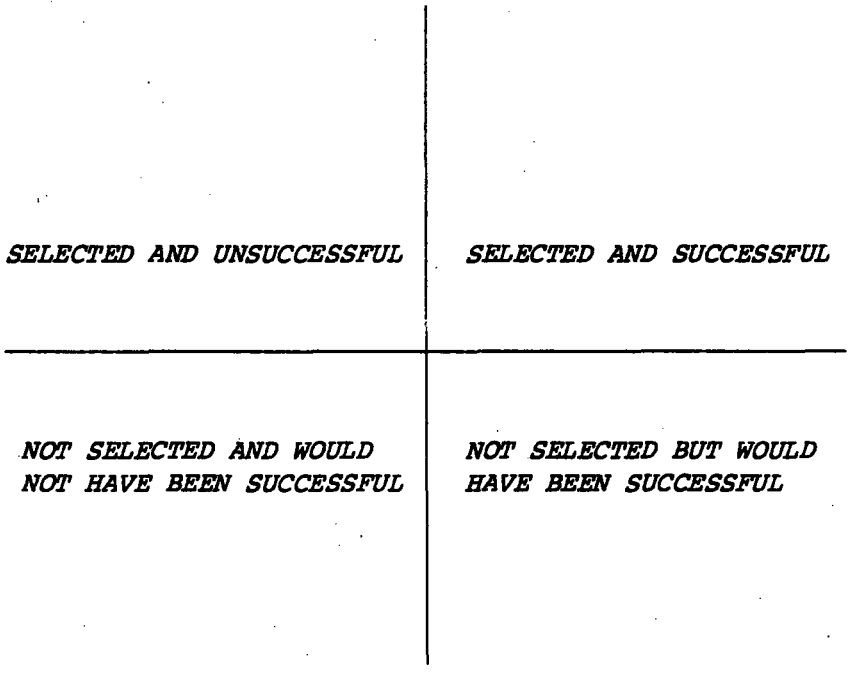


Figure 1. The meaning of the four categories defined by the cutoff points select-reject, success-fail.

It is necessary to consider these four categories in terms of what they really mean for both the applicants and the management:

1. Selected and successful: For the applicant this means personal achievement and for the employer productivity.

2. Selected and unsuccessful: For the applicant this means the trauma of being accepted only to fail subsequently. For the employer this means making investments in time and money in the training of people who are unsuitable for the job. (This is a very important category for decision making and will be further discussed at the end of this section with regard to the willingness to take the risk of failing, and also in the section about utility with regard to the importance of the criterion performance.)

3. Not selected and would not have been successful: Applicants in this category are rightly not given the job which they wanted. The employer also attaches little value to this category. He has to pay for the cost of testing, but does not enter into further relationships with the applicants concerned.

4. Not selected, but would have been successful: For the applicant this means unnecessary unemployment; for the employer, vacancies in the workforce. When the selection ratio is low (only a few candidates have to be selected from a large number of applicants), it may also be that many more candidates would have been successful than the number the organization can accept.

In this discussion on ratio models it is interesting to note that there is considerable agreement between the employer and the applicant as to whether a positive or a negative value should be attached to a category. The main controversy, however, turns on the category "selected and unsuccessful". The arguments put forward against tokenism and affirmative action are the following:

(a) If many members of the disadvantaged group are accepted who will eventually fail, the effect of second stage failure might be more pronounced than first stage rejection: the quotas are undone and there is considerable anguish for those who do not make it (Hunter & Schmidt, 1976).

(b) If the disadvantaged group in the workforce is less competent than the advantaged group, the situation will perpetuate unfavourable stereotypes. Thus, if a group as a whole gives a low performance, it has a bad effect on the image of all its members and even the competent workers of the disadvantaged group have to pay the price of lowered prestige (Jensen, 1980).

(c) If the institution adjusts to the performance level of employees and students selected by quotas, there will be an inevitable drop in standard (Hunter & Schmidt, 1976).

At the time of selection the criterion outcome is not known, therefore the employer will always be biased towards "success" on the criterion, whereas the applicant will prefer "selection", i.e. the applicants want to be given a chance even if it involves the risk of failure. The viewpoint of a black worker given in a survey of human relations in a retail organization serves as an example:

> Others have been here a long time. I don't really believe they are not capable. Those who are productive, who manage . . . they must be given a chance. It is better if someone fails at an attempt rather than being condemned beforehand (Franks, Ngwane, & Rheeder, 1986, p. 51).

6.3.4 Unqualified individualism versus qualified individualism

As long as norm referenced tests are used in decision making, the most accurate prediction will be made when all available information is taken into account, thereby placing the person in his/her correct norm group. The relative nature of norms is borne out in a study by Verster and Prinsloo (in press). In comparing the test scores obtained for Afrikaans-speaking and English-speaking South Africans in previous studies, including published and unpublished work as well as new analyses of the data, they were able to demonstrate a gradual reduction of the initial deficit of the Afrikaans speaking group. The oldest cohorts included in the sample were born as early as the 1890s and the youngest were contemporary preschool children. Verster and Prinsloo attribute the gradual convergence of mean scores of the two samples to a corresponding cultural

convergence between the two subpopulations. Van den Berg (in press) argues that because of the cultural convergence, not only between Afrikaans and English speakers, but also between white, "coloured" and Indian people, socioeconomic rather than language or ethnic groupings should be used. The formation of groups and their composition is a controversial issue. Outdated norms will work against members of the population whose norm group had the highest score when the test was constructed and this can be regarded as reverse discrimination. On the other hand, not recognising subgroups in a population will mean that bias in tests is ignored. Yet, the formation of groups can start a vicious circle, especially when applicants are selected on different norms and tested subsequently on tests with similar biases. The main problem is that correlations between certain variables and test scores have been found, but no satisfactory causal explanation of these relationships has been given. Definition of the mechanisms which mediate test performance is lacking, although suggestions including language, schooling, radio and television programmes, and child rearing practices have been made (Humphreys, 1973; Verster & Prinsloo, in press). "Broad cultural environment" is, contrary to what Van den Berg suggests

not enough to ensure equal measures of advantagement. Not enough proof for singling out socioeconomic status in preference to variables such as urban-rural, ethnicity, sex, malnutrition, and age, or, even for that matter birth order, or some weighted composite of some or all of these, has been given.

The use of criterion-related decisions and criterion-referenced tests seem to be defensible interim options. Equally attractive options are proficiency testing (non-normative mastery measurement), real life measures (performance appraisals), measures of learning potential (see Taylor, 1987), etc. None of these require a priori classification of subjects into groups as in case of normative psychometric testing.

6.4 Utility

In the initial work done on utility by industrial psychologists (Berkson, 1947; Brogden, 1946; McClelland, 1942; and Taylor & Russell, 1939) the evaluation of the efficiency of selection decisions was based on criterion related yardsticks, for example, job performance. But profit is often not linearly related to performance. Four possibilities are distinguished.

1. A monotonic increasing relationship exists between the criterion performance and its value. This is the situation in most easily measurable skills, for example, the skill of a sewing machinist in the number of pieces of work completed in an hour.

2. The criterion performance has an all or none effect. Some performances have dichotomous outcomes: hit-miss, life-death, freedom-imprisonment and often the consequences of failure are severe. Selection for such tasks must be undertaken with great care. Examples of careers in which failure has to be minimised are those of the airline pilot, air traffic controller, surgeon, judge.

3. Further improvement above a certain minimum requirement does not have a noticeable effect on the end product. For instance, on an assembly line the ability to do a task a little faster may be of little value to the organisation, because the output depends on the rate at which the objects pass that particular point.

4. In some instances the criterion performance has no quantifiable consequences. For example in survey

research, differences in responses of the participants are of equal merit.

Other relationships are possible, for instance, when success does not only depend on the quality of the job, but also on factors such as the ability to work with the public, as in the work of a general practitioner. The relationship of criterion performance and profit is thus not as clear-cut as it might appear to be at first. Furthermore, the viability of companies in South Africa will in future not only depend on profit, but more importantly on outside pressures, which will affect their selection policies.

Chemel (1985) indicates how different organizations might be pressurised into accepting different selection policies. He makes a distinction between government-controlled companies, universities and colleges, small companies, large South African companies, and big multinationals.

Privately owned companies and partnerships can select their staff with almost no influence from outside, but large South African-owned organisations will experience pressure from unions for a selection policy that does

not take race into account, i.e., group-blind selection. Government services and statutory bodies may go the same way, whereas foreign owned companies that are experiencing pressure of disinvestment as well as pressures exerted by codes of conduct, and perhaps even mandatory legislation in the home country will be forced to employ a selection policy which not only aims to be equitable but also to favour blacks at the expense of whites. Universities and colleges will be forced to face up to the dilemma of maintaining standards while at the same time upgrading the standards of underprivileged communities.

7. CONCLUSION

Dissatisfaction with tests in the United States stems from the fact that they are seen as a means of excluding certain groups from the mainstream of society. For instance, blacks are disproportionately represented in special programmes at school. This objection may seem incomprehensible, but when the consequences of classification are examined the objection becomes clear. It is aimed against the erroneous classification and labelling of children as handicapped, with the concomitant social consequence of

educational tracking, when they are non-English speaking, or of lower socioeconomic classes, or members of minority groups (Mercer, 1984; Reynolds & Brown, 1984). This can be contrasted to the placement of children in compensatory preschool programs, such as Learning to Learn and Head Start (Sprigle & Schaefer, 1985). In the one instance tests can be seen to exclude certain children from the mainstream realm of education, while in the other, facilities for entrance into the mainstream are provided.

In view of the situation which prevailed in South Africa in the past, where whites and blacks seldom competed for the same position in education or employment, tests were generally not used to exclude blacks from certain positions and a dissatisfaction with tests have not developed.

It is therefore very important for test users not to lose the trust of test takers and a number of safeguards should be built into the use of tests. Factors such as bias, comparability, the principle of inclusion and the application of appropriate models of fairness (authoritative selection) should be taken into account. Non-normative tests as discussed in section

6.3.4 should be given preference. If, however, norm-referenced tests are used, valid tests to select the most suitable candidates from each basic group together with an accountable policy of (black) advancement seem indicated.

The policy of advancing/not advancing blacks (or members of certain other identifiable groups such as women) can be applied at different levels:

1. Qualified individualism: The same test norms and cutoff points can be used for all applicants, with no reference to variables such as race, sex, or socioeconomic status. This means that, at present with tests being biased towards the advantaged groups, this strategy will result in discrimination against disadvantaged groups.

2. Unqualified individualism: As much information about an individual as possible can be taken into account to make the most accurate prediction possible about criterion performance. This can be seen as a policy of equal opportunity or criterion maximization, that is, merit based selection and promotion