DOCUMENT RESUME

ED 426 081 TM 029 308

AUTHOR Taylor, Terence R.
TITLE Are You Testing Fairly?
INSTITUTION Human Sciences Research Council, Pretoria (South Africa).
REPORT NO BIZ-1
PUB DATE 1990-00-00
NOTE 18p.; A summary in Afrikaans is included.
PUB TYPE Reports - Descriptive (141)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Culture Fair Tests; Equal Education; *Equal Opportunities (Jobs); Ethnic Groups; Foreign Countries; Minority Groups; *Occupational Tests; *Personnel Management; Prediction; Selection; *Test Bias; Test Use
IDENTIFIERS *Fairness; *South Africa

ABSTRACT

Because intelligence and culture are inextricably intertwined, the measurement of ability is difficult. Test material inevitably reflects the culture of the test developer. As a consequence of this culture-boundedness, a test might not measure the same thing in another culture, and scores might not be comparable across cultures. There are three main types of test incomparability (or bias): construct, item, and predictive. From the test user's point of view, the most important is predictive bias, for this has a direct bearing on processes of selection or placement of individuals in an organization. Predictive bias and fairness are not synonymous because fairness rests in the use of an assessment instrument and not in the instrument itself. There is no universally agreed-on conception of what is fair and what is unfair in an organization's relations with employees or prospective employees. It is incumbent on each organization to develop a clear fairness policy with regard to recruitment, selection, placement, and job advancement and to develop practical procedures to implement this policy. Various fairness models relevant to selection are discussed. It is recommended that organizations that do not have the skills to frame and implement a fairness policy either acquire these skills or engage the services of a consultant. (Author/SLD)

Report BIZ 1

1990

# Are You Testing Fairly?

Terence R Taylor



Pretoria
Human Sciences Research Council
1990

**ARE YOU TESTING FAIRLY?**

T R TAYLOR

## EXECUTIVE SUMMARY

Intelligence and culture are inextricably intertwined. This makes measurement of ability difficult: test material inevitably reflects the culture of the psychologist who devised the test; as a consequence of this culture–boundedness, a test might not measure the same thing in another culture and scores might not be comparable (i.e., have the same meaning) across cultures.

There are three main types of test incomparability (or bias): construct, item, and predictive. Their characteristics are described in the report. From the test user's point of view the most important is predictive bias – for this has a direct bearing on processes of selection and placement of individuals in the organization. Although predictive bias is relevant to the fairness of selection procedures, predictive bias and fairness are not synonymous. The reason for this is that fairness rests in the use of an assessment instrument and not in the instrument itself. A biased test can be used fairly and an unbiased test can be used unfairly.

There is no universally agreed–upon conception of what is fair and what is unfair in an organization's relations with its employees or prospective employees; this is not surprising when one bears in mind that there is no ethical system in the world that is universally agreed upon as the "best". Nevertheless, just as it is incumbent on each mature individual to choose a set of ethical standards for himself, so it is incumbent on each "mature" organization to develop a clear fairness policy with regard to recruitment, selection, placement, and job advancement – and to develop practical procedures to implement this policy.

Various fairness models relevant to selection are discussed in the report. It is recommended that organizations that do not have the skills to frame and implement a fairness policy either acquire these skills or engage the services of a consultant such as the NIPR.

## OPSOMMING VIR HOOFBESTUUR

Intelligensie en kultuur is onlosmaaklik verbind. Dit bemoeilik die meting van vermoëns. Toetsmateriaal reflekteer noodwendig die kultuur van die sielkundige wat die toets opgestel het. As gevolg van hierdie kultuurgebondenheid meet 'n toets dalk nie dieselfde verskynsel in twee verskillende kulture nie, wat die vergelykbaarheid (i.t.v. dieselfde betekenis) van toetstellings oor kulture heen, benadeel.

Daar is drie hooftipes toetssydigheid: konstruk, item en voorspelling. Die eienskappe word in hierdie verslag beskryf. Van die toetsgebruiker se oogpunt af is voorspellingsydigheid die belangrikste, want dit het 'n direkte effek op die seleksie/keuring en plasing van individue in die organisasie. Hoewel voorspellingsydigheid relevant is met betrekking tot die billikheid van keuringsprosedures, is voorspellingsydigheid en billikheid nie sinoniem nie. Die rede hiervoor is dat billikheid berus op die gebruik van 'n metingsinstrument eerder as op die instrument self. 'n Sydige toets kan billik gebruik word en 'n onsydige toets kan op 'n onbillike wyse gebruik word.

Daar is geen universeel aanvaarde konsepsie van wat billik in 'n organisasie se verhouding met werknemers of voornemende werknemers is nie. Dit is ook geen verrassing as mens in aanmerking neem dat daar geen etiese sisteem in die wêreld bestaan wat universeel as die "beste" beskou kan word nie. Nietemin, net soos dit by elke volwasse individu berus om 'n stel etiese standaarde vir homself te kies, so berus dit by die "volwasse" organisasie om 'n duidelike billikheidsbeleid te ontwikkel met betrekking tot werwing, keuring, plasing en bevordering en om praktiese prosedures te ontwikkel vir die implimentering van hierdie beleid.

Verskillende billikheidsmodelle wat van toepassing op keuring is, word in hierdie verslag bespreek. Daar word aanbeveel dat organisies wat nie die vaardigheid het om 'n billikheidsbeleid op te stel en te implimenteer nie, óf hierdie vaardighede ontwikkel, óf van die dienste van 'n konsultant soos die NIPN gebruik maak.

Intelligence cannot be separated from culture. Culture is a web of meaning through which the individual interprets the events of the world. It breaks up reality into "foreground" and "background" and distinguishes various concepts and relationships in the foreground. Each culture does this somewhat differently: the foreground and background are not quite the same and the "figures" – concepts – in the foreground are not identical in different cultures. Due to these differences, cultures have different ideas on what constitutes an intelligent breaking up – or interpretation – of facts and phenomena.

Take for instance this problem, which requires the individual to identify an odd–man–out from a set of five words:

DRUM FLUTE DANCE MUSIC SONG

When this item was administered to Dutch and central African subjects, the "smart" Dutch subjects selected the "DANCE" alternative, which was also the keyed right answer: it was keyed as right because it is the only one that is not directly to do with music. The more intelligent African subjects, however, chose "FLUTE" – presumably because this is the only member of the set that is not found in traditional celebrations. Here we can clearly see how culture – the web of meaning – determines the "intelligent" answer even a simple exercise such as this.

If the world is broken up in a different way by different cultures, do test scores have the same meaning in different cultures? There is a field of research in psychology, known as comparability research, which is devoted to studying this issue; sometimes it is also referred to as bias research, although the latter term has a rather more restricted meaning than the latter. Comparability refers to the degree to which (test) performance in one culture or group can be qualitatively and quantitatively compared with performance in another group. Clearly in the item cited above, performance does not have the same meaning, and if we regard the "FLUTE" response as wrong, we are being unfair to the African subjects.

Comparability is a fairly multidimensional issue and can be divided into at least three aspects. There are two comparability issues that are relevant to the measuring instrument itself:

(1) Is the psychological construct measured the same in the different cultures or groups, and
(2) Are there items that are anomalously easy or hard for certain groups?

The third issue involves not only the test but a criterion – such as job performance or training course performance:

(3) Does the test predict performance on such criteria in a similar way in different groups?

These three aspects are known as Construct Comparability, Item Comparability (or Item Bias), and Predictive Comparability (or Predictive Bias) respectively.

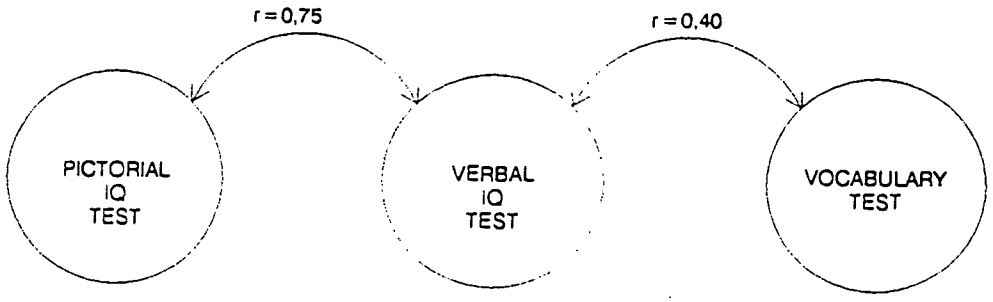
The three types of comparability, or incomparability, are related to one another. If the test measures different constructs in different groups, there is sure to be item bias and predictive bias. If the test does measure the same thing in different groups, but item bias is present in some of the items, then it is almost certain that there will be predictive bias.

Let us look at each of these three concepts in a little more detail, starting with construct comparability. Suppose that we have a verbally presented test of IQ or general intellectual ability. The verbal material is a vehicle for the measurement of IQ. In a group of mother tongue speakers (say, English–speaking whites), all the words that appear in the test are familiar; but in a group of blacks — for whom English is a second, third, or fourth language — some of the words are not known, or are only poorly understood. The blacks therefore have two hurdles to overcome to solve an item: a language hurdle and the conceptual hurdle of the item itself.

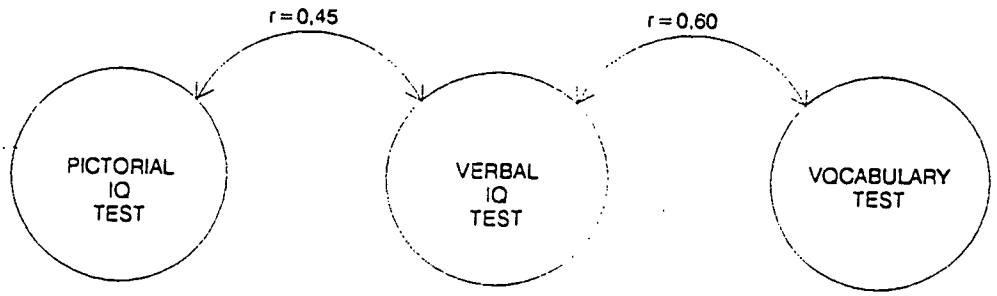
Now suppose that we administer three tests to the groups of whites and blacks: the verbal test of intelligence, a pictorial or nonverbal test of intelligence (such as the Ravens, FCT, or Pattern Relations), and a vocabulary test. The results that we might obtain are presented in the diagram that appears overleaf.

# CONSTRUCT EQUIVALENCE

WHITES




BLACKS

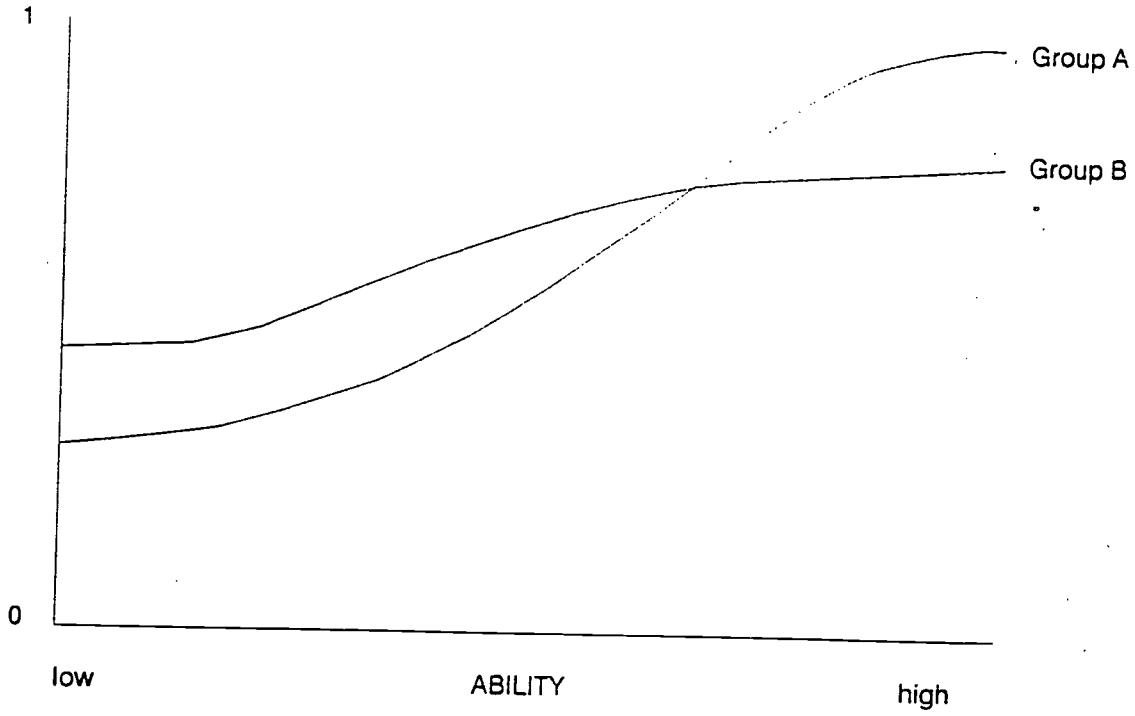



**Relationship among psychological variables in black and white groups**

For whites, the two types of intelligence tests are highly correlated, as one would expect, and the vocabulary test is correlated only modestly with the verbal intelligence test. In the black group, the picture is rather different: here the two IQ tests are not very highly related, but the verbal IQ test is highly related to the vocabulary test. This has happened because for blacks, the verbal IQ test is also something of a vocabulary test; consequently it does not measure intelligence that well – as is reflected in the lowish correlation with the nonverbal IQ test. The "bottom line" of this is that the verbal IQ test does not measure exactly the same construct in the two groups, and therefore the scores from this test do not have the same meaning and cannot be validly compared.

Item bias is at a more micro level than construct incomparability. As groups may differ in a certain ability, bias is not immediately indicated by group differences in the proportions of individuals who get a given item right. However, after taking these overall differences into account, certain items may be anomalously hard or easy for a certain group. For instance, an item in an IQ test that requires knowledge of the workings of an engine is likely to be biased against females – because knowledge of how cars work has little or nothing to do with general intelligence. In some cases, the nature of the bias might be quite complicated, as is shown in the following diagram.

## ITEM BIAS




**Example of complex item bias**

Here low ability members of Group A do worse than low ability members of Group B, but the opposite holds for individuals with high ability levels. An item is unbiased only if the curves for the various groups lie over one another. Once there are items that favour one or other group, scores are not comparable across groups, and the more items there are of this nature, the worse the situation is.

Predictive bias is probably the most important kind of bias for users of tests because it has direct implications for selection decisions and the fairness of employment practices. Tests are used in selection because they correlate (or should correlate) with relevant criteria, such as job performance. The relationship between test scores and criterion scores can be represented by a line called the regression line, which is the best–fitting line through the points representing subjects' test and criterion scores. This line is used to predict what criterion performance an individual is likely to display, given his test score. If, on the basis of this exercise, an individual appears to have sufficient ability or aptitude to do the job in question, a decision is then normally made to hire him; if he fails to make the grade, he is usually sent away. Test scores, and the way they are interpreted, therefore have a critical impact on people's lives.

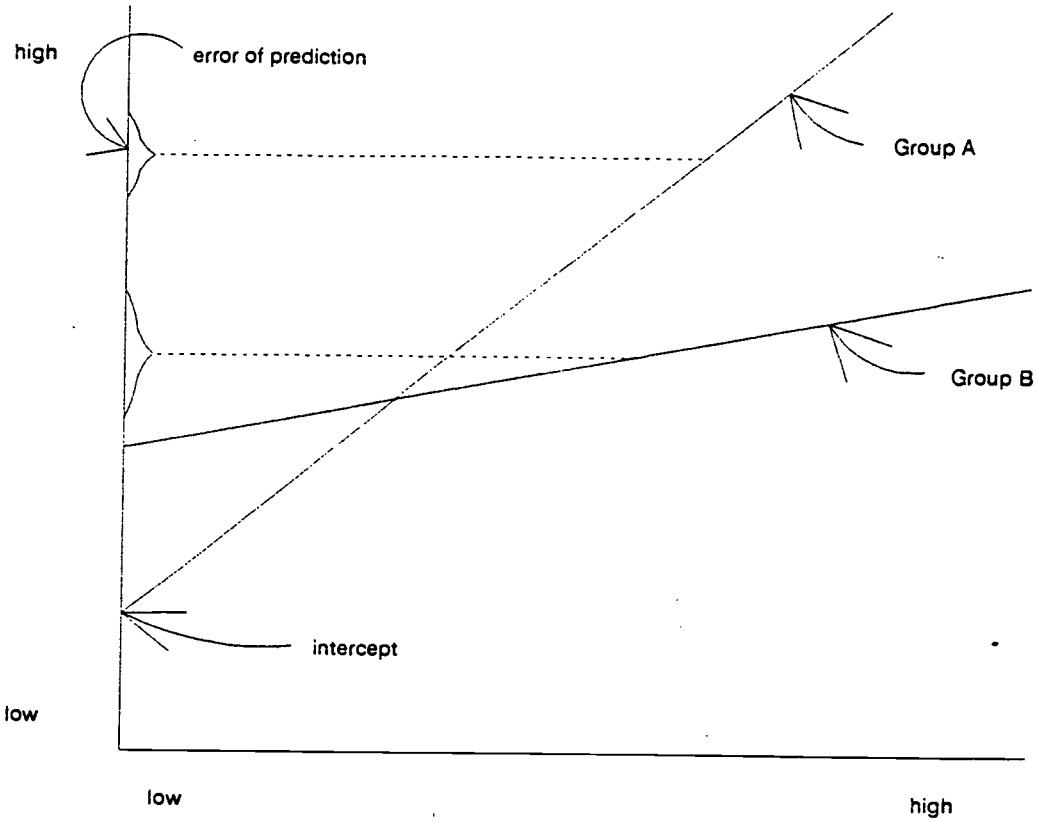
Now, for various reasons, it might be the case that the regression lines are not the same for different groups. The use of a single line (possibly based on data from only one group) to select all individual is then unfair, for members of certain groups will be advantaged while members of other groups will be disadvantaged by this practice. Predictive bias is absent only if the following conditions are met:

1. The slopes of the regression lines are the same in different groups;
2. The intercepts of the regression lines with the performance axis are the same; and
3. The error of prediction in the different groups is the same.

A graphical illustration of predictive bias is given in the next diagram. Here all three conditions are not met, and the use of a single regression line to select all individuals would be unfair.

We must now look at the idea of fairness in more detail. The first thing to know is that fairness is not the same as bias or incomparability: *a biased test can be used fairly and an unbiased test can be used unfairly*. Fairness rests in the use that the test is put to and not in the test itself. The phrase "the use that the test is put to" should be interpreted very broadly, to embrace all activities, from the manner in which the test is administered to the interpretation of scores and the selection decisions that are made. It is therefore the user who is primarily responsible for fairness in testing: tests themselves are not inherently unfair, although they

# BIAS IN PREDICTION




CRITICAL FACTORS:

Slope
Intercept
Error of prediction

**Predictive bias**

may be biased, and thus more prone to be used unfairly. It is the test publisher's responsibility to investigate tests for construct incomparability and item bias, to report on the results obtained, and ultimately to create tests that are effectively free of these kinds of bias and incomparability. (The NIPR is currently undertaking an exercise of this kind.) But the test publisher cannot establish unbiasedness of prediction, for there are many variables that are specific to each test user's application; however by eliminating item bias, the test publisher increases the probability that there will not be predictive bias. The publisher also cannot guarantee that a test be used fairly, and cannot even prescribe a single method of how to go about testing in a fair way.

Why can't the test publisher prescribe a strategy of fair testing? *Because there is no universally agreed upon conception of fairness*. Just as people cannot agree on matters of politics, they cannot agree on what constitutes fairness: one man's fair is another man's foul. This relativity, however, does not excuse organizations from stating their conception of fairness – clearly, exhaustively, and in practical terms. Vague, high–sounding phrases like "We are an equal opportunity company" are insufficient – almost meaningless – because they do not unambiguously declare a particular conception of fairness and how it is to be implemented. (Possibly the organization itself has not thoroughly thought through the issue and does not really know what it stands for.) Phrases such as the one given above are of more value as a public relations exercise than as the basis of a usable selection and placement procedure.

Let's take a closer look at fairness. Although the number of fairness conceptions is potentially limitless, they can be classified into various categories. There are two main varieties of fairness models:

1. Those based on regression equations, and
2. Those based on quotas or proportions.

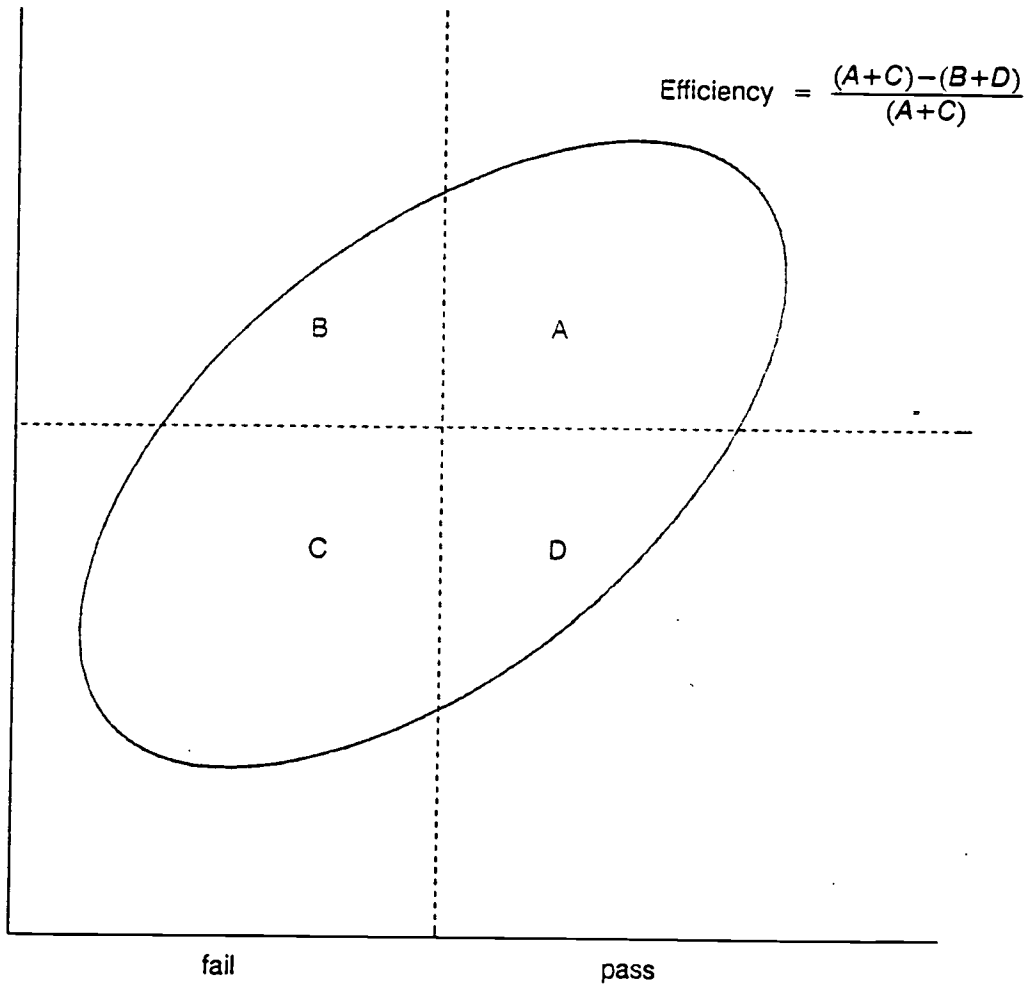
Implementing a fairness model based on regression equation is more demanding than implementing one based on quotas because a great deal more psychometrics is required. For a start, the three aspects of predictive bias – slopes, intercepts, and error of prediction have to be investigated. If bias is present, the prediction equation has to be set up to take account of this and eliminate its effects. Decisions also have to be made about what variables to put in the prediction equation: more than one test can be included, as well as biographical variables, even group membership. This last point, which is highly contentious, divides regression–based fairness models into two sub–categories, known under the names of unqualified individualism and qualified individualism. In an unqualified individualism approach, all variables that improve prediction are regarded as admissible: if group membership (such as sex, race, or

language group) increases predictive power, it is included in the equation. The qualified individualism approach, on the other hand, regards the inclusion of a variable reflecting group membership as inadmissible (for ethical reasons) and uses only information that does not require the knowledge of group membership. The approach is therefore "group blind". Unfortunately, there are many variables that fall in a "grey area" and require evaluation for the admissibility. These grey area variables can also be used to subvert the spirit of the qualified individualism approach: for instance by asking applicants their height and using this variable in the prediction equation, the sex variable is introduced through the back door.

There are many quota or proportion models, each based on a different conception of fairness. These models can be understood more easily by referring to the diagram which is presented overleaf.

## QUOTA MODELS

The quadrants of the predictor–criterion distribution on which they are based




Two popular models:

Proportional Representation requires that $\frac{(A+D)}{(B+C)}$ be the same in all groups

Constant Ratio requires that $\frac{(A+D)}{(A+B)}$ be the same in all groups

**Proportion models**

This diagram depicts the relationship between the predictor – some appropriate test – and the criterion – job or training performance. The oval ring represents the area occupied by the test–criterion points for each subject: the more elongated this oval, the more effectively the test predicts the criterion. We assume that a level can be identified on the criterion dimension, such that performance above that level is acceptable and below that level unacceptable. This level on the criterion can then be used to set a cut–off on the predictor. These two cut–offs (on the predictor and criterion) allow us to divide the plot into four areas:

A. Those individuals who satisfy the predictor cut–off and who also succeed on the criterion;

B. Those who do not satisfy the predictor cut–off but who nevertheless succeed on the criterion;

C. Those who do not satisfy the predictor cut–off and who fail on the criterion (or who would fail if they were selected); and

D. Those who do satisfy the predictor cut–off but who fail on the criterion.

If the test is any good as a predictor, the proportion of people in A and C should be substantially greater than the proportion of people in B and D; for A and C are "hits" (correct assignments) whereas B and D are "misses". The efficiency of the selection exercise is given by the formula:

$$\frac{(A+C)-(B+D)}{(A+C)}.$$

All quota fairness models make use of this four–way classification of subjects. Quite a few of these have been proposed in the literature, of which six or eight have gained some prominence. The conception of fairness underlying each model can be inferred from the mathematical formula that it proposes as the basis of the selection decision.

The main interested parties in the selection exercise are: the applicant as an individual; advocates of group rights and advancement; and the employing organization. Fairness models always embody a view of fairness as seen from the point of view of one or other of these parties. Let us take what is possibly the simplest and best known of the quota models – the proportional representation model. This model states that the proportions of individuals selected from various groups should reflect the proportions in the applicant population: in mathematical terms, the ratio (A + D)/(B + C) should be the same for all groups. This model clearly reflects a conception of fairness as seen from the point of view of groups that have been previously discriminated against in the workplace – usually disadvantaged groups. Applying the model often constitutes a form of affirmative action.

Now let us consider another model, the constant ratio model. It specifies that selection test cut–offs should be set for different groups so that the ratio of accepted individuals to successful individuals is the same in all groups. In this model the ratio $(A+D)/(A+B)$ must be the same in all groups. The model is attractive to the managements of some organizations (hence is regarded by them as fair) because it is minimally disruptive. Essentially it says: previously excluded, disadvantaged or minority groups are welcome in the organization as long as their employment does not alter the proportion of individuals who "make it" on the job; if a *higher* cut–off on the selection test has to be set to achieve this end, so be it. Clearly, disadvantaged or previously excluded (or discriminated against) individuals are unlikely to regard this model as favourably as management.

We now come to the critical question: How is an organization to apply selection tests in such a way as to be acceptable in a country that is trying to throw off its discriminatory past? As we have seen, there is no single way and the approach adopted depends on what is valuable – maximum efficiency, a workforce representative of the population, social upliftment, etc. It is not unlikely that Government will in future years also play a greater role in setting criteria for fair employment, in which case the State's political agenda, group aspirations, and the organization's self–interest will all have an impact on the choice of a fairness model.

Organizations therefore are not completely free agents: their employment practices will be influenced by union pressure, pressure by various groups who may or may not be aligned with unions, and Government policy. These factors will have to be taken into account along with the organization's own interests when deciding on employment practices. As different organizations are subject to different patterns of pressures, it is likely that they will come to somewhat different conclusions as to what route to take in this regard. Organizations should think through the issues carefully and thoroughly, taking their situation and the possible consequences of various courses of action into account before deciding on their basic orientation. Then they should frame a policy on selection that is explicit and can be unambiguously implemented in the selection situation. In some cases the organization may benefit by making its policy public, but it should not be blind to the possibility that certain groups may not like the policy and may engage in action (legal or illegal) to pressure the organization into changing its policy..

Once an organization has developed a preference for a certain model, it should, before finally settling for the model, consider the amount of effort, expense, and expertise required to implement it. Problems of bias are usually taken more seriously in the regression based models, but this adds complexity to these models. It is probably true to say that regression models require greater statistical sophistication than quota models; but these models are probably the best for optimizing the quality of the workforce – although this efficiency can be

at the expense of certain desirable states of affairs such as a socio–culturally balanced workforce. Quota models can be easier to put in place; but, like regression models, most of them require that valid and reliable criterion data be collected. In addition, a fair amount of work has to be done on setting appropriate cut–offs. A drawback from some points of view is that group membership has to be taken into account and that cut–offs are liable to differ from group to group: this practice might be regarded as unfair and challenged by certain groups.

Not all organizations, of course, have valid and reliable data on criteria, of course: in fact, possibly only a minority do. How is one to proceed then? The only reasonable practice under these circumstances, it seems to me, is to use separate norms for each group and the same normalized cut–off for all groups. (A computer program is available from the NIPR which runs on a IBM–compatible micro–computers and which makes it easy for an organization to compile sets of norms on data from its own test population.) The cut–off will be decided, "semi–scientifically", on the basis of the judged the complexity of the job in question and the abundance of job applicants. Although the normalized cut–off will be the same for all groups, the raw–score cut–offs will vary from group to group, due to the fact that different norm tables are being applied. This practice overcomes the problem of item bias to a large degree and might even favour disadvantaged individuals (in that more of them would be selected than would be the case of a pure optimization of work–performance model were applied). The separate norms approach is therefore likely introduce a degree of affirmative action – which some organizations might welcome. There are, however, disadvantages to making selection decisions in this rather scientifically naive manner, and the adoption of a model (regression or quota) that takes criterion performance into account is preferable. If the organization lacks the expertise to install one of these models, it should consider contracting the NIPR or some other organization with the necessary skills to undertake it.

TM029308

**U.S. Department of Education**
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

# NOTICE

## REPRODUCTION BASIS

☑ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").