Test – Item Bias: Methods, Findings and Recommendations

Report ED-15

Test – Item Bias: Methods, Findings and Recommendations

K. Owen

Human Sciences Research Council
Pretoria
1992

ISBN 0 7969 1218 1

K. Owen, D.Litt. et Phil., Senior Research Specialist

Division for Psychology in Education
Group: Education
General Manager: Dr S.W.H. Engelbrecht

Published by:

Human Sciences Research Council
134 Pretorius Street
Pretoria
0002

ACKNOWLEDGEMENT

Special thanks are due to Mrs W.G. Verster who typed the manuscript and prepared it for publication, and Mrs Jennifer Folwaczny who revised the language.

TABLE OF CONTENTS

EKSERP

Die vraagstuk van toetssydigheid is tans uiters aktueel. Meeste van die psigometriese toetse wat in Suid-Afrika ontwikkel is, is vir 'n bepaalde groep opgestel. As gevolg van die veranderende politieke en nywerheidsopset in die land het situasies ontstaan, byvoorbeeld mededinging tussen groepe, wat nie deur die gebruik van afsonderlike toetse hanteer kan word nie. Die gebruik van gesamentlike of gemeenskaplike toetse bring egter 'n nuwe probleem na vore, naamlik dié van sydigheid.

Die doel met hierdie verslag is om die toetsopsteller en -gebruiker in 'n groter mate bewus te maak van die vraagstukke betrokke - veral dié van sydigheid - wanneer 'n psigometriese instrument wat vir een groep ontwikkel is by 'n ander groep gebruik word. Met hierdie doel voor oë word aandag geskenk aan metodes vir die opspoor van sydigheid, navorsingsbevindinge in verband met sydigheid en stappe wat deur toetsopstellers geneem kan word om sydigheid in toetsing te verminder.

Daar word vertrou dat hierdie verslag tot 'n beter begrip sal lei van die talle fasette wat betrokke is by toetssydigheid.

ABSTRACT

The question of test bias is a very important one at present. Most psychometric tests developed in South Africa were developed with a specific group in mind. Changing political and industrial conditions in this country are now giving rise to situations such as intergroup competition that cannot be handled by separate tests. However the use of joint or common tests poses a new problem, namely that of bias.

The aim of this report is to promote in the test developer and test user a greater understanding and awareness of the issues involved, notably that of bias, when a psychometric instrument that has been developed for one group is used for another. With this purpose in mind, attention is given to methods for detecting bias, research findings regarding bias, and steps that could be taken by test developers to reduce bias in testing.

It is hoped that this report will make some contribution towards an understanding of the many facets involved in test bias.

1.0 INTRODUCTION

Since the earliest measurement of mental abilities, it has become clear that such tests might be class or culture bound. For example, soon after the appearance of Binet and Simon's intelligence test in 1905 it was noticed that children from the higher socio-economic groups obtained higher test scores on average than children from the lower socio-economic groups.

The suitability of a psychometric test for a group that did not form part of the standardization group is closely related to the problem of bias. According to Reynolds and Brown (1984) this matter came to light mainly because of the nature of psychological processes and the measurement of such processes. Psychological processes cannot be observed or measured directly and the nature of a process therefore has to be deduced indirectly from behaviour. Psychologists have reached consensus on very few of these deductions or hypothetical constructs and it is therefore understandable that intelligence or intellectual ability, which represents what is probably one of the most complex processes in psychology, can stimulate the interest of both experts and laymen. The criticisms - including that of bias - levelled against psychometric tests such as intelligence and aptitude tests should also be seen against this broad background. Complaints about test bias are particularly heard from minority groups in the USA and elsewhere. According to them certain tests are most suited to the group that formed the largest component of the standardization sample. Whether such tests are indeed biased and harmful to the interests of minority groups is one of the questions that have to be answered empirically.

To limit the potential influence of cultural factors on test performance Cattell in 1940 suggested a "culture free" intelligence test. Owing to the misunderstanding and misrepresentation caused by the term "culture free" the name was later changed to "culture fair" intelligence tests. At that stage it was not clear whether there really was a link between culture and test performance. The first systematic investigation in the USA with regard to cultural bias in psychometric tests was con-

ducted in 1945 by Allison Davies, a sociologist, and Kenneth Eells, a psychologist (Jensen 1980). These researchers regarded cultural bias in test items as only one of several possible factors responsible for average IQ differences between culture groups. Other aspects such as hereditary traits, developmental factors, motivational factors, work habits and test-taking skills were also taken to play a role.

According to Jensen (1980) the confusion in the literature regarding terms such as "bias" and "unfairness" that developed after the pioneer work of Eells can be attributed to a lack of conceptual clarity regarding the meaning of these terms. For example, "cultural bias" became a popular cliché for explaining racial and social differences in intelligence test performance. Although there was much conjecture on the subject, researchers did not try to determine the real nature of test score differences or to establish objective criteria for determining bias in a specific test.

During the sixties the situation in the USA became more pressing and concern developed over the fact that the test scores of blacks and other minority groups (the "culturally deprived") were generally lower than those of white Americans. These differences were found in a wide variety of tests, for example intelligence, scholastic aptitude and performance tests. Because the tests revealed differences they were condemned as being culturally biased against blacks by sociologists, anthropologists, educationists and other critics who were actually outside the field of psychometrics. The outcome of this was the formation of pressure groups advocating the total abolition of psychological tests. However responsible persons and institutions realized that rather than abolishing tests attention should be paid to the fair use of tests on the one hand and the possible bias of such tests on the other.

During the seventies psychometricians embarked on a more systematic study of concepts such as test bias and test fairness. Not only did this result in greater clarity in definitions and terminology, it also led to a great deal of research.

The term *bias* has been used quite a few times so far in this report and a brief definition of the concept is probably in order (it will be discussed in more detail later). In mathematical statistics the term bias refers to a systematic underestimate or overestimate of a population parameter by

a statistic that is based on a sample from the population. In psychometrics the term refers to systematic errors in the *predictive validity* or the *construct validity* of the test scores of individuals that are associated with the group membership of those individuals (Jensen 1980).

These two aspects of test scores, namely predictive validity and construct validity, represent the two main fields of research on test bias. At first glance the two fields appear to be completely divergent and to have little in common in terms of methods of investigation. However the two approaches or fields do *not* represent different concepts of bias and, as Raju and Normand (1985) have indicated, in certain cases the same method of analysis can be used by both approaches.

Research in South Africa regarding test bias is still in its infancy. There has been no particular need for such research since the tests developed for the various ethnic groups have largely been separate. Changing political and industrial conditions in South Africa are now giving rise to situations such as intergroup competition that cannot be handled by separate tests. However the use of joint or common tests poses a new problem, namely that of bias. The question can rightly be asked whether a group that was not included in the standardization of a specific test is not at a disadvantage when taking that test. If the intention is to introduce unbiased common tests for the various population groups (so that test scores reflect true and not artificial differences), attention will have to be given to the extent of bias in tests as well as to the nature, operation and causes of bias. Although it is important to know what types of items are often biased, it is even more important to determine why these items are biased. It is more important that the test developer promote insight and understanding than that he or she merely identify and eliminate biased test items.

The aim of this report is firstly to promote in the test developer (item writer) and test user a greater understanding and awareness of the issues involved, notably that of bias, when a psychometric instrument that has been developed for one group is used for another. With this purpose in mind, the main findings regarding test and item bias contained in reports P-66, P-96, P-106 and 1991/1 (Owen 1986, 1989a, 1989b and 1991a respectively) are summarized and integrated with relevant overseas findings. Secondly, attention is drawn to steps that could be taken by test developers to

reduce bias in testing. Of special importance in this regard are the steps initiated by prominent test publishers in the USA.

To familiarise the reader with methods and techniques that could be used for determining bias in tests, a brief overview of the more important of these is given and references to a more detailed description provided.

Apart from features of the test itself, there are also certain background factors that may contribute to test bias. Some of these are now briefly discussed.

2.0 SOME BACKGROUND FACTORS REGARDING TEST BIAS

There are various background factors which play a role in understanding why divergent groups (in terms of culture and other characteristics) perform differently on specific tests or test items. Although many such factors can be distinguished, for instance the quality of school education, motivation, test anxiety and test sophistication, only four of the most important ones will be briefly discussed, namely culture, socio-economic status, language and cognitive style.

2.1 Culture

The issue of *cultural differences* is probably the most common criticism levelled at standardized tests. For example, the Council of the Society for the Psychological Study of Social Issues states that:

> *We must also recognize the limitations of present day intelligence tests. Largely developed and standardized on white, middle-class children, these tests tend to be biased against black children to an unknown degree* (quoted by Jensen 1974:190).

It seems fair to assume that a specific culture stimulates a specific form of cognitive development, in other words that intellectual ability is *linked to culture* (compare Lesser, Fifer & Clark 1965; Scarr 1981; Anastasi 1970; Sundberg & Gonzales 1981). According to Lesser *et al* (1965: 3) the study of *cultural differences* with regard to intellectual ability is becoming a fundamental issue in education and psychology: *How do we provide valid psychological evaluation of children from widely dissimilar groups?*

One reason why no definite answers have yet been found to this question is, according to Anastasi (in Lesser *et al* 1965:11), that a concept such as intelligence is defined by Western technological culture:

> *It is not so much that tests are unfair to lower-status groups, as that lower-class environment is not conducive to the effective development of 'intelligence' as defined in our culture.*

Various researchers agree that every culture encourages and promotes the development of certain abilities or types of behaviour while discouraging and suppressing others (Anastasi 1976). The possible relation between culture and cognitive development becomes clear from the prerequisite set by Ferguson (1954) for the development of an ability, namely that the opportunity for "overlearning" the activity should exist within the relevant culture. From this it follows that if a certain activity is not found in a culture, the ability concerned is also absent. Irvine (1970) points out that non-Western communities in which skills are traditionally transferred orally prescribe their own rules for intelligent behaviour. Such behaviour may use knowledge in a manner which differs completely from the individualistic competitive and industrialized manner of the West. It also appears that people in a rural environment who have little school experience tend to be unable to accept the logical assumptions of syllogisms and to draw conclusions on the basis of these assumptions (Cole & Scriber 1974). However it is clear from an experiment conducted by Brislin (1983) that black pupils are able to deal with syllogisms formulated in their mother tongue.

According to Triandis and Brislin (1984) the relation between culture and cognitive development can be viewed from three perspectives, namely the *universal* (all cultures show certain similarities with respect to cognition), the *evolutionary* (differences in cognitive functioning are sought in the activities of the group concerned) and the *relativistic* (the focus falls on differences between cultural groups). These writers maintain that all three approaches are valid to a certain extent. For example, the existence of universal characteristics cannot be denied: all people categorize, use opposites and classify. However several writers point out that, although the components of the cognitive system, in other words memory, categories, associations, syllogisms, coding and decoding, semantic integration and verbal explanation, are found in most cultures, they are related

in an extremely complex way and deviations occur as a result of specific characteristics of the culture concerned (for instance literacy) (Triandis & Brislin 1984; Bond 1981; Irvine 1970). In this regard Anastasi (1985:122) states the following:

> *These differences are reflected in the development of abstract thinking and in the nature and breadth of the concepts that are formed.*

Ebel's (1979) statement that what a person learns depends on the culture in which he finds himself is illustrated very well by Medley and Quirk (1972, 1974). These writers come to the conclusion that if the 65 items comprising the Common Examinations of the National Teacher Examination were to be replaced by "black" items (i.e. items dealing with the activities of American blacks in the field of social life, literature and the arts, but that are not so esoteric that an informed person from another population group cannot answer them too), blacks would improve their test performance by 27% in relation to that of the whites. These findings correspond with those of Nancy Abrams (1979) who found that blacks performed about one and a half standard deviations better than whites in a "black" test.

The effect of culture on intellectual ability can take many forms. In addition to direct learning situations in and out of school the typical behaviour code of the community also exerts a subtle influence. Guthrie in 1963 posed the question whether members of a community based on authority and tradition develop a reasoning ability that is organized "differently". Almost twenty years later Ghuman (1980) found indications that this could be the case. According to him young people from the Punjab displayed a convergent style of thought that could be ascribed to the rigid form of authority maintained by the teachers in their schools and encouraged by the community as a whole.

To summarize it can be said that cognitive development and the acquisition of certain abilities are to a considerable degree linked to culture. However on the other hand the statement by Buss (1977) could also be true, namely that cultural differences have their origin in the situations in which cultural groups apply their skills rather than in the skills themselves. This view corresponds with that of certain modern anthropologists:

...the belief systems and cultural premises of traditional people may differ from those in industrialized societies, but they embody the same logical processes and concern with relation of cause and effect (Cole & Scribner 1974:25).

From the above one may surmise that although the influence of culture is extremely subtle in many ways and difficult to observe directly, it could still be an important source of item bias, particularly in plural communities with divergent cultural backgrounds.

2.2 Socio-economic status (SES)

There is apparently little doubt that socio-economic circumstances affect test performance in some way or other; that better circumstances put certain individuals and groups at an advantage when taking an intelligence test, that children of a higher social status generally perform better than children of a lower social status (Havighurst & Breese 1947; Lorge 1952-53). In recent years the question of the role of socio-economic factors in test bias has become very important to psychologists (Oakland & Feigenbaum 1979).

What are the characteristics of an underprivileged person? According to Novick (1980) our knowledge regarding what constitutes a disadvantage or deficit is still incomplete. However certain factors that contribute to educational deficits can be identified, for example poverty, ill-health, malnutrition, inadequate school facilities, maladjustment, absence of cultural stimulation, illiterate and superstitious parents, lack of books, pictures, furniture, toys, television and all other items of modern society that contribute towards shaping the thoughts of the Western child; lack of discipline and particularly the absence of a stable relationship with a supportive adult, or feelings of hostility and helplessness regarding the social and educational establishment (Vernon 1968; Scarr 1981; Thorndike & Hagen 1969). The cultural level at home, the parents' level of education and the extent of parental encouragement are some of the most important factors that influence a child's performance in intelligence and verbal tests (Radford & Burton 1972; Vernon 1968; Eells *et al* 1951).

Children from underprivileged areas often perform poorly in psychological tests because these tasks are strange to them and they regard them as unimportant (Lesser, Fifer & Clark 1965). If a

child's environment were to be improved, it could lead to improved test performance (Walberg & Haertel 1984). There is little doubt that the IQ's of schoolchildren correlate with the socio-economic status of their parents (Jensen 1969; Humphreys & Taber 1973). Wigdor and Garner (1982) estimate the correlation between SES and scores in ability tests to be approximately 0,30. The extent of the differences in performance between low and high SES groups is illustrated by the fact that the average test scores of pupils from families in the top 20% of the socio-economic distribution lie on approximately the 65th percentile with respect to the general population, while the average scores of children whose families fall in the bottom 20% of the socio-economic distribution lie approximately on the 35th percentile (Wigdor & Garner 1982).

According to Jensen (1969) children with low SES perform equally well in Jensen's Level I abilities (that is, associative processes such as memory) regardless of their ethnicity. Further to this Horn (1976) also found that low and high SES groups performed equally well in Level I abilities. As far as Level II abilities are concerned (that is, abilities involving conceptual and reasoning processes) MacKenzie (1981) report a stronger association with SES than in the case of Level I abilities.

The different ways in which SES-defined groups approach and solve the problems contained in test items is instructive not only in that they illustrate differences between the groups, but also because they contribute to a better understanding of the possible causes of item bias. The following are some of the most important findings with regard to the relationship between SES and performance in test items:

- Children of low SES complete the test more quickly and often select an answer option at random if they do not understand the principle involved, for instance series completion (Wolf 1978; Eells *et al.* 1951). Similar observations have been made with Peurto Rican schoolchildren in New York and children in Hawaii (Marmorale & Brown 1979).

- If the structure of the item is complicated, for example in numerical problems, children of a low SES find it more difficult to solve (Wolf 1978).

- The distractor which is most similar to the correct answer is almost equally popular among low and high SES groups. However low SES group members more often select less probable answers (Wolf 1978) and their answers are also distributed more uniformly across all the answer options (Eells *et al.* 1951).

- Certain types of items such as analogies, antonyms and classification items can be revised or rewritten to eliminate the effect of middle-class social status more easily than others, such as syllogisms (Haggard 1954).

- The greatest mean difference between SES groups are found on verbal items and the smallest on items comprising pictures and geometric figures (Eells *et al.* 1951).

- Almost all items on which there are unusually large differences between SES groups consist of verbal symbolism; a large number of these items involve academic or bookish vocabulary (Eells *et al.* 1951).

- Items on which SES group differences are small consist almost without exception of either non-verbal symbolism or simple, common words (Eells *et al.* 1951).

- Motivated children of a low SES perform significantly better than non-motivated children of the same status (Haggard 1954).

- If the aim is to narrow the considerable gap in performance between low and high SES groups, it is necessary in the first place not only to increase the lower status group's familiarity with the test material, but also to increase their motivation to perform well (Haggard 1954).

To summarize it can be said that socio-economic status has a bearing on item difficulty, the typical response pattern to an item, familiarity with the type of material an item comprises and the motivation to perform well. There is also an interaction between SES and item type or format. When compiling a test to be used with different SES groups the above matters should be kept in mind.

2.3 Language

As far as test language is concerned Oakland (1977) points out that it should be of such a nature that every testee can understand what is expected of him and can respond freely and with ease. If this cannot be achieved the language in the test may contribute to cultural bias (Schultz & Fortune 1981: 119). Although language is extremely important - in many cases it is responsible for the gap between verbal and nonverbal scores - its effect is indirectly proportional to the amount of instruction given to testees in the language medium of the test (Jensen 1980).

The translation of a test results in many problems of conceptualization, cultural interpretation and connotation (Merryfield 1985; Olmedo 1981). Translation into non-standard dialects is no solution; this resulted for instance in only a minor improvement in the performance of American black children (Oakland 1977). Elements in the test material such as a word or a sentence, a diagram or a figure, can stimulate different associations in different cultures (Vernon 1969).

When tests are taken in a language which is not the dominant language of the testee - as is often the case in Africa - a test that is supposed to measure intelligence, for example, may also provide unwanted measurements, such as language proficiency (Guthrie 1963; Hale 1982).

The language usage of some blacks in Africa is characterized by a unique style, involving for instance a tendency to give elaborate descriptions rather than exact definitions of concepts. Direct statements or standpoints are considered rough and unimaginative, while camouflage by means of constantly changing definitions is regarded as intelligent behaviour (Hale 1982). It is not yet clear at this stage to what extent such traditional language usage is also characteristic of black primary and high school pupils in South Africa - owing to the influence of Western-type school instruction it is suspected that this might not play a major role in their test scores.

2.4 Cognitive style

Research findings suggest that the typical cognitive style of individuals contribute to group differences in test performance and therefore also to test and item bias. Of all the different cognitive styles that have been identified (see for example Coop & Singel 1871; Guilford 1980; Federico 1983)

the bipolar field-dependent (FD)/field-independent (FI) style has not only been the subject of the most research, it also has the widest application in psychology and education (Templer 1972; Witkin *et al.* 1967, 1977).

Field-dependence and field-independence (FD and FI) refer to an individual's tendency to organize his experiences in either an analytical or a global way. A FI person is able to focus on a specific stimulus despite the presence of a number of attractive irrelevant stimuli. This ability is not just a mental approach, it represents a general analytical orientation. Such individuals are able to overcome or restructure the organization or form of a given field. FD individuals on the other hand accept the field as it is and find it difficult to distinguish the parts of the field from the overall context.

As regards the development these cognitive styles is, it appears that individuals reared in a climate that emphasizes autonomy and performance tend to be field-independent, while those who have grown up in a protective and conformist environment are generally more field-dependent. On the basis of this one might expect some cultures to be more field-independent and others more field-dependent.

In general FD testees' performance has been found to be poorer than that of FI testees (Globerson. *et al.* 1985), particularly in the presence of misleading cues to the solution of a task. However in the absence of such misleading information, FD testees have performed just as well and even better than FI testees (Witkin *et al.* 1977). According to the latter writers FI testees performed better than FD testees in mathematics, science, engineering and architecture. Kaufman (1979) reports that FI testees performed considerably better than FD testees in laboratory tests for spatial orientation and in psychometric tests such as hidden figures, block design, etc.

Cognitive styles are not found confined to specific ethnic or socio-economic groups. In the USA it was found that blacks and other low-income groups were prone to use the relational style but that it also occured in higher-income groups (Hale 1982). Although no group used a certain cognitive style exclusively, a definite relationship was found between style and cultural or ethnic

group, particularly when the group could be placed at a certain point on the socio-economic scale. Hilliard (quoted by Hale 1982:42) makes the following claims in this regard:

- Afro-Americans tend to react to objects in terms of the total image and not the various parts, and Euro-Americans tend to divide everything into smaller parts which together form a whole.

- Afro-Americans tend to prefer inferential reasoning to deductive and inductive reasoning.

- Afro-Americans tend to use estimates rather than actual values when working with space, numbers and time.

- Afro-Americans tend to focus on people and their activities rather than on objects.

Several researchers (for example Kaufman 1979; Federico 1980; Hunt 1980; Rohrl 1979; McGee 1979) have identified a relationship between cognitive style and the functioning or activities of the left and right hemispheres of the brain. Among normal right-handed people the functioning of the left hemisphere is associated with a verbal, analytical, sequential, syllogistic and objective way of processing information, while that of the right hemisphere is associated with a spatial, synthetic, nonverbal, creative, holistic and intuitive method of processing information.

According to Kaufman (1979:97) the basic difference between the left and the right hemispheres does not lie so much in the stimuli that are processed as in the way in which these are processed:

> *The left hemisphere specializes in sequential (his underscoring) processing, analytic, linear, and successive in nature; the right hemisphere is lateralized for simultaneous (his underscoring) or multiple processing, holistic in nature.*

This means that each hemisphere of the brain has its own rules for processing information and that the brain therefore has two "intelligences" that function independently (in view of this Witelson (1977) hypothesizes that persons suffering from dyslexia have two right hemispheres). The left hemisphere, which represents the scientific method,organizes data according to the principle of conceptual similarity, while the right hemisphere functions according to the principle of structural

similarity. Kaufman (1979) illustrates this with the aid of the following example taken from the Stanford-Binet Intelligence Test: The left hemisphere regards an apple and a pear as similar because both are types of fruit, but to the right hemisphere they are both round.

During the past decade psychological theories regarding the functions of the left and the right hemispheres of the brain came to the attention of anthropologists who consequently formulated certain interesting hypotheses. One such is that left hemisphere functions increase in importance as communities become dependent on artifactual recording devices such as measuring sticks and scales and calibrated monitoring equipment such as watches and calendars (Rohrl 1979:356).

Ethnic differences in test performance may be related to the functions of the left and right hemispheres of the brain. Although people use both hemispheres, it appears that some cultural and other factors could promote the development of a certain cognitive style or style of thought. In an attempt to identify the relationship between cultural factors and cognitive style, Rohrl (1979) connects certain cultural factors with brain functions.

The cultural correlates of the left hemisphere functions include:

- Specialized social organization
- High degree of dependence on verbal communication
- Large and impersonal community
- Individualism
- Day-to-day pattern of existence not directly dependent upon nature
- A great deal of verbal and formal instruction
- Dependence on watches, calendars, etc.

Cultural correlates of the right hemisphere functions include:

- Shared-function social organization
- Relatively little need for verbal communication
- Small, open face-to-face community
- Tradition and respect for elders and their advice
- Close relationship with nature
- Learning through example rather than verbal instruction
- Relatively little need for watches, calendars, etc.

According to Rohrl (1979) a preponderance of traits from one of these sets indicates a preference for the corresponding brain function. Although the above cultural correlates have not yet been confirmed, Kaufman (1979:103) maintains that...*research suggests a possible right hemisphere leaning among blacks.* His assumption is supported by the fact that American blacks excel in typical right hemisphere functions as measured by figure completion tests and visual spatial tasks such as pattern recognition. American blacks also have considerable talents associated with right hemisphere functions, for example music, creative dancing and visual art.

Some of the cultural correlates identified by Rohrl (1979) with respect to right hemisphere functions seem to apply to a large number of blacks in South Africa (see for example Van der Vliet 1974; Dubb 1974). However this does not justify the conclusion that blacks in South Africa are more prone to right hemisphere functions than whites. If this were the case, it could be a factor in group differences in test performance and possibly also in test bias.

On the basis of the above discussion it emerges that cognitive style is an additional viewpoint from which group differences in test performance can be studied. It has been shown that cognitive style may affect the test performance of individuals and groups in a variety of ways, depending on the

type of test being used, the intellectual tasks that have to be executed and the specific hemispherical brain function involved.

3.0 METHODS FOR DETECTING BIAS IN PREDICTIVE VALIDITY, CONSTRUCT VALIDITY AND TEST ITEMS

Methods for detecting test bias can conveniently be divided into three groups or categories, namely, those used for detecting bias in

- predictive validity,
- construct validity, and
- test items.

This distinction is in many ways artificial since test items and predictive validity are also dimensions of the construct validity of a test. It is nevertheless convenient to make such distinctions as they allow one to focus with greater clarity on certain aspects of the test. It must also be borne in mind that although the methods and techniques pertaining to each category differ from one another conceptually, they all share a common goal: the detection of systematic error in the estimation of some true value for a particular group of individuals.

3.1 Methods for detecting bias in predictive validity

In the context of predictive validity bias refers to *..systematic error in the estimation (i.e. constant over-or-under-prediction) of performance on some criterion measure...* (Reynolds 1982: 215). Since predictive validity involves a correlation between test scores and a criterion measure, methods for assessing bias centre around the correlation coefficient and the regression line (see also Section 4).

3.1.1 SINGLE-GROUP VALIDITY AND DIFFERENTIAL VALIDITY

The use of a tests' validity coefficient as a means of detecting bias dates back to the 1960s in the USA when the issue of test bias in personnel selection for civilian occupations first became a major concern (Jensen 1980: 499). These earliest researchers relied on single-group validity and differential validity findings as the primary indicators of bias. According to Jensen (1980: 499-500)

> *Single- group validity is demonstrated when one group shows a validity coefficient significantly larger than zero and the other group does not. Differential validity is demonstrated when the two groups' validity coefficients differ significantly from one another.*

Results obtained by the two methods are independent, in other words, differential validity cannot be inferred from single-group validity, even if the sample sizes are the same.

It should also be pointed out that, although equal validities imply that the test can be fairly applied to all the groups concerned, this is still no guarantee that the test is not biased for one group as the regression equations could differ between the groups (Jensen 1980:470). It is therefore obvious that validity coefficients are useful but limited as indicators of test bias. (Differential validity and single-group validity are discussed in more detail in Par.4.1.)

3.1.2 SLOPE, INTERCEPT AND STANDARD ERROR OF ESTIMATE

Test bias in terms of predictive validity relates to the homogeneity of the regression line. There are four conditions which can occur when two or more groups are involved and homogeneity of regression is not present. These are variation in the slopes, in the intercepts, in the slopes and intercepts, or in the standard errors of estimate (these conditions are described in greater detail in Par.4.2).

The slope and intercept values across groups can be evaluated by a method described by Potthoff (details are provided by Reynolds 1982: 220-221). With this method the equivalence of regression

coefficients (slopes) and intercepts across a number of independent groups can be simultaneously tested with a single F ratio.

If this test is significant (indicating absence of regression homogeneity), the slopes and intercepts may be tested separately. Significant F values indicate bias.

Conclusions regarding bias can not be drawn on the grounds of slope and intercept findings alone. To justify a claim of no bias, the standard error of estimate (SE_{est}) must be equivalent across groups (see Par.4.2 for an explanation of the SE_{est}). The SE_{est} for a set of predicted scores is given by the following equation (Reynolds 1982: 223):

$$SE_{est} = SD_y\sqrt{(1 - r_{xy}^2)\left(\frac{N-1}{N-2}\right)},$$

where SD_y represents the standard deviation of the scores on Y, r_{xy}^2 represents the squared validity coefficient, and N represents the sample size on which r_{xy} is based. The significance of the difference between independent SE_{est} derived from two samples can be tested by the F ratio formed by the variance errors of the estimate (Reynolds 1982: 223):

$$F = \frac{SE_{est_1}^2}{SE_{est_2}^2}$$

where $SE_{est_1}^2$ represents the square of SE_{est} for group 1 and $SE_{est_2}^2$ represents the square of SE_{est} for group 2.

3.2 Methods for detecting bias in construct validity

Bias in construct validity essentially means that the test measures different constructs in different groups while the assumption is that the same construct is being measured (see Par. 5.1 for a more formal definition). This section deals with bias assessment methods that are based on criteria

internal to the test itself, such as consistency estimates (reliability), item difficulties and the factorial structure of the test.

3.2.1 INTERNAL CONSISTENCY ESTIMATES

According to Jensen (1980: 430) the reliability of a test is a sensitive indicator of bias if the major and minor groups show significantly different reliability coefficients and the coefficient in one of the groups is below an acceptable level, say <0,90. However, Jensen also points out that a significant difference between reliability coefficients is not in itself sufficient to establish bias; other factors, such as group differences in item difficulties, or differences in item correlations should also be taken into account.

Coefficients such as Cronbach's alpha and the Kuder-Richardson 20 (KR20) are examples of internal consistency reliablity estimates. A technique provided by Feldt (Reynolds 1982: 208-209) can by used to determine the significance of the difference between these coefficients for two groups on a particular test:

$$F = \frac{1 - alpha_1}{1 - alpha_2},$$

where $alpha_1$ is the reliability coefficient for group 1 and $alpha_2$ is the reliability coefficient for group 2 on the same test. The quantity 1-alpha represents an error variance term, and the largest variance is always placed over the smallest variance term.

Other reliability coefficient comparisons are (Reynolds 1982: 209-210):

- correlations between alternate forms of a test (may be used when alpha or KR20 are inappropriate), and
- test-retest correlations across groups.

3.2.2 RANK ORDER OF ITEM DIFFICULTIES

The rationale behind comparing the rank order of item difficulties (p values) for different groups is that the items *...should 'behave' in the same way* (Bond 1981: 61) for all groups if the same construct is measured.

To evaluate the consistency of item difficulties across groups, all test items must first be ranked according to difficulty. This is done separately for each group. A rank-order correlation (Spearman's rho) is then calculated between the two sets of ranks. If the sample sizes are sufficiently large (with a subject: item ratio of not less than 10:1), a rho of 0,90 or higher indicates relative consistency of item difficulties across groups (Reynolds 1982: 210).

When the above technique is used, the comments of Osterlind (1983;17) should be borne in mind:

> *It should be clearly understood, however, that comparing the rank order of item difficulty indices between groups is an incomplete strategy for concluding bias in test items. It is, nevertheless, a useful tool as an early indication of whether or not particular items behave differently between groups.*

A related technique is the

- correlation of p decrements between adjacent items, also called delta-decrement analyses (see Jensen 1980: 441-442).

3.2.3 FACTOR ANALYTIC METHODS

Factor analysis is one of the most popular and important methods of assessing bias in construct validity. The concept investigated by means of factor analysis is that the test score variance is composed of the same theoretical constructs, or factors, for different groups (Osterlind 1983: 17). If different factor patterns are found for the different groups, it could be an indication of bias in the construct validity of the test. Factor analysis can also be regarded as a procedure that

...identifies clusters of test items or clusters of subtests of psychological... tests that correlate highly with one another and less so or not at all with other subtests or items. Factor analysis then allows one to determine patterns of interrelationships of performance among groups of individuals (Reynolds 1982: 201).

To determine whether the same constructs are measured in different groups, two factor analytic approaches can be followed, namely, exploratory or confirmatory factor analysis. In exploratory factor analysis a correlation matrix is factor analysed and a number of factors extracted. This can be done separately for each group. A number of methods for determining the degree of similarity between factors for the different groups have been devised. One popular technique, based on the relationship between pairs of loadings for corresponding factors, is the coefficient of congruence (r_c). This coefficient is given by the following equation (Reynolds 1982: 204):

$$r_c = \frac{\sum_1^N a_1 \cdot a_2}{\sqrt{\sum_1^N a_1^2 \sum_1^N a_2^2}},$$

where a_1 represents the factor loading of a variable for one sample and a_2 the corresponding factor loading of the same variable for the second sample. A r_c value of 0,90 or higher is arbitrarily taken to indicate factor equivalence, in other words, factorial invariance across groups.

Confirmatory factor analysis, however, provides a more stringent test of factor similarity than coefficients of congruence. Path analysis can be used to determine whether a specified model fits the data for the different groups involved in the study. The statistical methods used with path analysis models involve estimating all free parameters as well as obtaining measures of model fit (see e.g. Owen 1991b, for an application of path analysis).

3.3 Methods for detecting item bias

The statistical or psychometric problem in detecting item bias is to find methods that differentiate between test performance differences caused by real differences between groups and those caused by bias (e.g. due to the way items have been formulated) (Burril & Wilson 1980). Tech-

niques for detecting item bias are therefore aimed at identifying specific test items that are associated with cultural differences and contribute to inaccurate measurement (Rudner & Getson 1982).

A shortcoming of most of these methods is that they are based on an internal test criterion. This inevitably introduces an element of circularity to the logic of the methods for detecting bias. The total score for the test (to which items that may perhaps be biased have actually contributed) is used to identify persons with equal abilities and the performance of these persons is subsequently used to identify biased items (Shepard, Camilli & Williams 1983). Thus the methods can only succeed in detecting relative bias and not absolute bias.

The following are some of the most important methods or techniques for the detection of biased items (definitions of item bias are given in Section 6):

- analysis of variance
- item discrimination procedure
- transformed item difficulty values
- distractor response analysis
- item-group (partial) correlation
- chi-square techniques
- item characteristic curve (item response theory).

The first four techniques mentioned above are based on the unconditional definition of bias, while the remaining three are based on the conditional definition (see Par. 6.1.).

3.3.1 ANALYSIS OF VARIANCE

In the analysis of variance procedure (ANOVA) the focus is on the interaction of groups-by-items and not the main effects. Significant main effects do not necessarily indicate bias in a test for the simple reason that there may be genuine differences in ability between groups (significant group main effect), or differences among the items in terms of difficulty (significant items main effect) (Osterlind 1983). The rationale behind the assumption that a significant groups-by-items interaction is indicative of bias is as follows:

...items that measure a different trait ...for subgroups ...violate the assumption of unidimensionality of items. Items lacking unidimensionality across subgroups will likely exhibit varying degrees of difficulty regardless of the groups' overall ability level difference. The groups x items interaction of analysis of variance procedures will reveal this effect of different levels of difficulty between or among groups (Osterlind 1983: 21).

When a significant groups-by-items interaction is revealed by ANOVA, other methods are required to identify those items which are biased. Methods recommended by Osterlind in this regard are Scheffe's multiple comparisons and Angoff's transformed item difficulties (see Par. 3.3.3).

Although the ANOVA is an attractive and useful procedure, it has a number of shortcomings. The main criticism is that...*item by group interactions will occur in completely unbiased tests merely as a function of differing levels of ability or mean performance* (Rudner, Getson & Knight 1980: 217). This method is nowadays not frequently used in bias detection studies.

3.3.2 ITEM DISCRIMINATION PROCEDURE

The discrimination value of an item is the correlation of its score with the total score on the test. This correlation (biserial or point-biserial) indicates to what extent a particular item measures whatever is measured by the test as a whole. Items that show a low (non significant) correlation with the total score do not contribute to the true variance in the test scores but only to error variance.

If one assumes that item discrimination values express the degree of relationship between the items and the underlying construct measured by the test, it follows that items with different discrimination values for different groups do not measure the same constructs in those groups and are therefore possibly biased. Conversely, in an unbiased test the item-by-total score correlation for any item should be the same for the groups concerned. According to Jensen (1980: 445), rigorous testing of this hypothesis is difficult for the following reasons: (1) the item-by-total score correlation has a large sampling error, (2) these discrimination values are usually fairly homoge-

neous, and (3) the item discrimination value is affected by the difficulty of the item, which may not be the same for the groups involved.

Although the item discrimination procedure is not a popular method for detecting item bias (Burrill 1981), its value lies in the fact that it is readily available with item analyses and aberrant items can be spotted at first glance. (For a further discussion of the item discrimination procedure, see Angoff 1982: 109.)

3.3.3 TRANSFORMED ITEM DIFFICULTY VALUES (TID)

A description of the transformed item difficulty values (TID) method, also called the 'delta-plot method', can be found in Angoff (1982: 96) and Osterlind (1983: 28).

The TID approach rests on the assumption that an item that is comparatively more difficult for one group than it is for another, is probably biased. "The underlying logic is simply to test if different groups interact with a set of items in the same way" (Osterlind 1983: 29). If this is not the case, in other words, if groups respond differently to certain items, one may conclude that bias exists.

In the TID method, the item difficulty levels (p values) are first transformed to a normal scale (z values). This is done separately for each group. The z scale values are then linearly transformed into a delta scale in order to remove negative values: $\Delta = 4z + 13$. Next, the pairs of deltas (one pair per item) are plotted on a bivariate graph with the delta values for one group plotted on the abscissa and those for the other group on the ordinate. Typically, a plot of these points will appear in the form of an elongated ellipse, extending from the lower left to the upper right of the graph. The closer the resemblance between the item difficulty patterns of the groups, the narrower the ellipse. A straight line (major axis line) is fitted to the data. Items whose points fall at some distance from this line may be regarded as contributing to the item-by-group interaction. Formulas for calculating the major axis line of the ellipse and the distance of points from this line, are given by Angoff (1982: 98). The final step is to identify biased items by evaluating the distance of points from the major axis. This requires that one establish limits for acceptable deviation from the axis. According to Osterlind (1983: 35) a common approach is to place confidence intervals of $\pm 0{,}75$

z-score units about the major axis. Items falling outside these boundaries are regarded as outliers and therefore possibly biased.

The TID method has been widely used, one reason in particular being that it is a visual method for the study of item-by-group interaction. One of its drawbacks, however, is that item discrimination and item difficulty tend to be confounded in cases where the groups differ considerably in mean ability level (Angoff 1982: 104). Under these circumstances, items identified as biased by the TID method may in actual fact only appear to be so because they are highly discriminating.

3.3.4 DISTRACTOR RESPONSE ANALYSIS

A description of this technique can be found in Jensen (1980: 452) and Osterlind (1983: 69).

In objective tests, items usually consist of a stem (the wording/statement of the question) and answer options, one of which is the correct answer while the others are distractors. Test developers usually try to make an item's distractors equally attractive but seldom succeed as can be seen from the percentage of testees who choose the different distractors as their answer.

The choice of distractors can provide valuable clues to why an item functions differently for different groups. If an item is unbiased, its distractors can be expected to display the same measure of attractiveness for all the groups concerned. In 1975 Veale and Forman (see e.g. Rudner, Getson & Knight 1980) proposed a method of investigating item bias that is not dependent upon the test total score like most of the other methods. They suggested that two or more groups' response patterns be compared. If a significance test reveals that the groups are differentially attracted to a test item's distractors, the null hypothesis (no difference) may be rejected and bias may be concluded.

The rationale behind this approach is the assumption that there are certain item characteristics *...that cause a distortion in the item p-value for a specific group* (Rudner, Getson & Knight 1980: 225). This distortion is brought about by distractors which are not equally attractive to all groups. The hypothesis that the frequency with which a distractor is chosen is not the same

for two or more groups, is tested by means of a chi-square statistic (see Osterlind 1983: 70-73 for more details).

A problem with distractor response analysis is that response pattern differences between groups are not so much a function of bias as of ability (Thissen 1976); i.e. testees with higher ability tend to choose other distractors than those favoured by testees with lower ability.

All in all, despite the drawbacks of distractor response analysis, Rudner and Convey (1978: 23) are of the opinion that this technique is one of the more promising and interesting approaches and *...should be considered in future investigations of item bias methodologies.*

3.3.5 ITEM-GROUP (PARTIAL) CORRELATION

Stricker (Angoff 1982: 112) proposes an index of item bias which is based on the correlation between success on an item and group membership, with true scores on the test (with the item under consideration omitted) partialed out. This index is highly attractive because, like the ICC method (see Par. 3.3.8), the procedure controls for group differences in overall ability. According to Angoff, one of the advantages of this index is that it is far less costly in terms of computer software requirements (when compared for instance with the ICC method) and also in terms of the amount of data needed to achieve certain levels of reliability. However, the effectiveness of this method in item bias detection has not yet been sufficiently demonstrated.

3.3.6 SCHEUNEMAN'S CHI-SQUARE

The chi-square approach for the identification of biased items examines the probability that testees from different groups but with the same level of ability will answer a certain item correctly (Osterlind 1983). If an item is unbiased, individuals of the same ability level should have the same probability (chance) of answering the item correctly, regardless of their group membership (Scheuneman 1979). This approach is therefore not dependent on item-by-group interactions for the identification of biased items (as in the case of analysis of variance and some other methods). Two popular methods that use the chi-square approach are Scheuneman's chi-square and the "full" chi-square.

According to Scheuneman (1979, 1980) her proposed chi-square technique should be regarded as a variation of item response theory (latent trait procedures). Both the difficulty and discrimination values of items are allowed to vary, but the continuous curve which is characteristic of the latent trait procedure is abandoned. In its place the test total score is divided into a small number of discrete ability categories (three to five) and the performance of two or more cultural groups within these categories is compared. Thus Scheuneman's chi-square technique involves the use of the total raw score of a homogeneous test to measure or indicate ability. The distribution of the total score is divided into three to five categories or intervals that are each defined by a certain range of the total score (i.e. if the test consists of 30 items counting one point each, the five ability categories may include testees who obtained scores ranging for example from 0-9, 10-15, 16-20, 20-25, 26-30). The distribution of the number of correct responses within each ability category is obtained for each item and for every population group. The number of correct responses within each ability category represents the obtained frequency used in the chi-square test. The proportion of correct responses of all testees (irrespective of ethnic background) whose scores fall within a particular ability category is used to determine the expected frequency. In other words, the expected value for each cell (E_{ij}) is obtained by multiplying the proportion of all testees who answered the item correctly and whose total score fall within interval j by the number of testees within the cell. Thus

$$E_{ij} = \frac{O_{.j}}{N_{.j}} N_{ij}$$

where

$O_{.j}$ = the number of testees in the total score interval (ability category) j who answered the item correctly

$N_{.j}$ = the total number of testees in the score interval j, and

N_{ij} = the total number of testees in group i and in score interval j.

The test statistic (χ^2) is

$$\chi^2 = \Sigma\frac{(B_e - B_o)^2}{B_e} + \Sigma\frac{(W_e - W_o)^2}{W_e}$$

where

B_e = the expected frequencies of Group B

B_o = the obtained frequencies of Group B

W_e = the expected frequencies of Group W, and

W_o = the obtained frequencies of Group W.

The number of degrees of freedom are (K-1) (r-1), where K is the number of population groups and r the number of ability groups.

One of the shortcomings of this method (as well as of the "full" chi-square) is that if a test contains a considerable number of biased items, the test total score will also be biased. Consequently the ability categories, which have been formed on the basis of the total score, will not reflect a group's true potential. When there is a considerable difference between the total score distribution of two groups, for example because one group finds the test much easier than the other, the establishment of comparable ability groups poses a practical problem.

Intasuwan (1979) and Baker (1981), among others, have raised a more serious objection to Scheuneman's technique in particular. In their opinion the fact that she does not incorporate the proportion of incorrect answers in her technique means that her test statistic does not really have chi-square distribution. In other words it is only "half" a chi-square.

3.3.7 FULL CHI-SQUARE

In 1981, Camilli (Shepard, Camilli & Averill 1981) reported that the results of a chi-square statistic based on the proportion of correct responses (p) differ from those in which the proportion of incorrect responses (q) is used. This was an awkward finding, since the proportion of correct and

incorrect answers are supposed to reflect the same information. Camilli, in agreement with Intasuwan (1979), consequently suggested that the full chi-square, which includes both correct and incorrect testee responses, be computed. In other words, a conventional chi-square for a two-by-two contingency table (for example ethnic group x correct-incorrect responses) should be computed for each ability level and summed across all the ability levels. The calculation procedures thus match those of the Scheuneman technique with the exception that incorrect responses are also taken into account. The number of degrees of freedom are r (K-1), where r is the number of ability groups and K the number of population groups. If a particular item score does not form part of the total score, the test statistic calculated for this item can be regarded as a true chi-square (Intasuwan 1979). However this implies that score intervals (ability groups) have to be determined for each item in the test. According to Rudner and Convey (1978), deviating from this approach by keeping the ability groups the same for all the items in a test and including the item under scrutiny in the total score has the following advantage:

> *The inflation of the χ^2 values will be systematic when identical intervals are used for each item ... This systematic inflation allows the χ^2 to be used as a relative index of bias (p. 20).*

A point in favour of Scheuneman's chi-square is that it is far more difficult to create cells of adequate size with the full chi-square in the case of particularly easy items. It can also be argued that there is less of a need for a significance test (and consequently a distribution of known characteristics) than there is for an index of bias. Shepard (1981) also points out that any final conclusion regarding bias should be based on logical grounds and that the deviation of an item - from other items measuring the same construct - should carry more weight than the power of the statistical test.

3.3.8 ITEM CHARACTERISTIC CURVE (ITEM RESPONSE THEORY)

The mathematics involved in item response theory (IRT) is rather complex and this theory will therefore not be discussed in detail. Comprehensive descriptions can be found *inter alia* in Hambleton, Swaminathan, Cook, Eignor and Gifford (1978), Hambleton (1980), and Lord (1980).

The item response theory is based on the idea that latent traits (hypothetical dimensions of cognitive abilities) can explain the coherence among different items of a test (Hambleton & Cook 1977). Latent traits are so called because they are "imperceptible" (i.e. not directly measurable). These traits can be described as psychological dimensions underlying the abilities and attitudes of individuals. Performance in a task related to a certain trait can be predicted by estimating the individual's ability level with regard to that trait. (These estimated scores are used to predict performance.)

The space that is determined by all dimensions required to explain the coherence among items is called the complete latent space (Van der Flier 1980). According to its definition, the conditional distribution of the item scores within this space (given the position of all latent traits) is equal for all relevant populations. When a test consists of homogeneous items, it is assumed that the coherence among the items can be explained by a single underlying trait. The regression of the observed item score on this trait (θ) is called the item characteristic curve (ICC) or item ogive.

The advantages of the IRT over the classical test model can be summarized as follows:

- The testee's ability is estimated on the same ability scale as any subgroup of items fitting the model.
- The item parameters are invariant across subgroups of testees (for which the model fits) making it possible to compare subgroups.
- In some IRT models the scales actually have the properties of an interval scale. Raw scores are usually on an ordinal scale, but for the practical purpose of being able to use certain statistical procedures, it was assumed in the classical test theory that the scales have the properties of an interval scale.

A disadvantage of the IRT is that it seems extremely unlikely that any set of data will ever fully satisfy the requirements of the assumptions made regarding any latent trait model. Apparently the latent trait models may not be robust with regard to deviations concerning the assumption of unidimensionality (Hambleton 1980).

A number of latent trait models have been developed. Most models presuppose a more or less s-shaped item characteristic curve. The best-known models are the three-parameter normal ogive model, the three-parameter logistic model, the two-parameter logistic model and the one-parameter logistic model (also known as the Rasch model). The first two models yield similar ICCs, but the logistic distribution function is mathematically less complicated and is also used more often (Van der Flier 1980). In all of these models it is assumed that an item is unbiased if ... *examinees of the same ability level, but of different group affiliations, have equal probabilities of responding correctly* (Rudner, Getson & Knight 1980: 222).

Since the three-parameter logistic model is used more often than the other two models, the parameters of this model will now be discussed.

The mathematical form of the item characteristic curve of the three-parameter logistic model is as follows (Hambleton *et al* 1978: 473):

$$P_g(\Theta) = c_g + (1 - c_g)\frac{e^{Da_g(\Theta - b_g)}}{1 + e^{Da_g(\Theta - b_g)}}, (g = 1, 2, \cdots, n).$$

where

$P_g(\Theta)$ = probability that a testee with ability Θ will answer item g correctly,

D = a constant scaling factor (usually taken as 1,7),

e = the mathematical constant 2,71828...,

a_g, b_g and c_g = the parameters describing item g.

The a parameter represents the discriminating power of the item. The steeper the slope of the curve the better the item discriminates between low and high ability testees. Discrimination values of 0,5 to approximately 2,5 are quite typical while a value smaller than 0,5 is not adequate for testing purposes (Ree 1979).

The b parameter represents an index of the degree of difficulty of the item and is represented by a point on the ability scale (Θ) that coincides with the inflection point of the curve (i.e. the point

where the curve changes direction). The further to the right the inflection point, the more difficult the item, and the further to the left, the easier the item. Although b usually lies between -3 and +3, the choice of this scale is arbitrary (Ree 1979).

The c parameter (pseudochance parameter) is the lower asymptote of the curve and represents the probability of a testee of extremely limited ability quessing the correct answer to the item. This parameter is usually larger than 0,0 and smaller than 0,30.

Different methods can be employed to estimate the above item parameters. A commonly used method is contained in the computer program LOGIST, compiled by Wood, Wingersky and Lord (1976). This program estimates the item parameters (a, b and c) as well as the ability parameter (θ), and places them on the same scale. When the item parameters are estimated separately for two different population groups, they will not be identical but linearly related.

Therefore, although the parameters of both groups essentially lie on the same scale, there is a linear transformation difference between them (Osterlind 1983). This difference is caused by the fact that θ is arbitrarily defined as having a mean of 0 and a standard deviation of 1 in each separate parameterization. In this regard Dorans (1979: 11) states that

> *... the fact that standard scores from separate derivation samples have identical means and variances does not mean that these scores are expressed in a common metric. Scores are expressed in a common metric only when they share a common reference point and a common unit.*

Despite the fact that the a and b parameters are invariant from group to group, they are not invariant when the origin of θ changes arbitrarily in each parameterization. The scales of two separate groups then have to be equated before the respective parameters can be estimated and compared (Dorans 1979). Various methods may be used for equating the θ scales (see for example Osterlind 1983; Lord & Wingersky 1984).

Once the scales have been equated and the parameters estimated all that remains is to compare the item characteristic curves (ICCs) of the two groups with a view to identifying biased items. If the ICCs of the two groups are identical, the item concerned is unbiased. However if the ICCs differ, certain methods may be used to determine whether the deviation is such that the item should be considered biased. Rudner, Getson and Knight (1980) for example established an index of item bias based on the area between the ICCs of the two groups, while Lord (1980) developed a chi-square statistic for testing the equality of the a and b parameters of the groups (the c parameter is not tested because it is held constant while the a and b parameters are estimated). A useful method for evaluating the differences between the ICCs of two groups, is also suggested by Linn, Levine, Hastings and Wardrop (1980).

Although the ICC method (three-parameter model) is the most theoretically sound and certainly one of the best methods for detecting item bias, it also has a number of drawbacks. One of these is that parameter estimation becomes rather problematical if there are large ability differences between the groups. Another problem is that large sample sizes (N> 1 000) are required in order to obtain stable parameter estimates. The ICC method is also more costly in terms of computer software than most other techniques.

Despite its drawbacks, the ICC procedure (three-parameter model) is the most statistically elegant of all item bias detection techniques discussed and should be used whenever possible. However, in view of the considerable agreement between the ICC and the more practical chi-square procedures (Scheuneman's and full chi-square), the latter ... *may be viewed as a rough approximation to the more complicated three-parameter latent trait procedure* (Ironson 1982: 152). The choice between the two procedures will ultimately depend on sample size and cost.

3.3.9 CONCLUSION

Although not all available item bias detection methods have been discussed in Par. 3.3, those mentioned represent the more important ones. A number of empirical studies have been conducted to evaluate the effectiveness of the various methods (see Taylor 1987 for an overview). A typical finding is that the ICC method (three-parameter model) performs best, followed by the

chi-square procedure, and, in certain circumstances, the TID method (other promising techniques are the log-linear and logit models - see Taylor 1987: 38).

According to Ironson (1982) there are still many unresolved issues in the domain of item bias methodology, e.g. the theoretical differences between methods, their reliability, the validity of the procedures in correctly identifying "truly" biased items, and the extent to which the various indices of item bias are robust to violations of their assumptions. However, methodological issues constitute only one side of the coin. The other crucial question is "... what makes an item biased"? (Ironson 1982: 153). If one is unable to identify and explain those aspects or factors that cause an item to be biased for a particular group, nothing has been learned about the nature of bias in test items. In Section 6 an attempt is made to answer some of the questions in this regard.

4.0 BIAS IN THE PREDICTIVE VALIDITY OF A TEST

Simply put, test validity indicates to what extent a test measures what it is supposed to measure; it also says something about the conclusions one can draw on the basis of test scores. Although several forms of validity are distinguished (Messick 1980), only criterion-related or predictive validity will be discussed here. This again involves two types of validity, namely concurrent validity (test and criterion data were obtained at the same time) and predictive validity (test and criterion data were obtained at different times). These two forms of criterion-related validity are however calculated in exactly the same way.

4.1 Differential and single-group validity

Differential validity exists when there is a significant difference in the correlations between the test and criterion scores of different groups (Jensen 1980).

In terms of Cleary's (1968) regression definition the evaluation of bias in predictive validity takes the form of $\hat{y} = aX+b$, where $\hat{y}$ is the predicted criterion score, a is the regression coefficient, X is the score in the predictor (test) and b is a constant. If the equation is represented graphically as a regression line, a represents the slope and b the Y intercept of the line. Differential predictive validity exists when the slopes of the regression line for two or more groups differ and the same regression line is used for all the groups. According to the definition of bias in predictive validity, the errors in prediction should be independent of group membership if there is no bias. This is not so if the slopes of the regression line for the various groups differ, in other words the expected errors of estimate are not zero.

Bias in differential predictive validity also exists when correlations between the test and criterion scores (the validity coefficients) for various groups differ significantly from each other as a func-

tion of group membership (Shepard, Camilli & Averill 1981). The differential validity hypothesis regarding bias is therefore that certain tests are more valid for majority groups than minority groups, but that the coefficients are not necessarily zero for any of the groups (Schmidt, Pearlman & Hunter 1980).

According to the single-group validity hypothesis regarding bias a test is valid for one group but not for another, in other words the true population validity is zero for the one group but not necessarily for the other (Hunter & Schmidt 1978). Single-group validity is therefore the result of the predictive validity coefficient for only one of the groups being greater than zero. An extreme statement of this hypothesis in terms of whites and blacks is that

> *...black culture is so alien to white culture that a test might be completely meaningless to blacks* (Hunter, Schmidt & Rauschenberger 1984:45).

4.2 The slope, intercept and standard error of estimate of the regression line

Cleary (1968) and Cleary and Hilton (1968) were among the first to attempt a definition of test bias in terms of predictive validity:

> *A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup* (Cleary 1968: 115).

In terms of this definition therefore a test is biased if the criterion score, which was predicted with the aid of the common regression line, is consistently too high or too low for members of the subgroup. In accordance with Cleary's definition many researchers (such as Evans & Reilly 1972; Goldman & Hewitt 1975) agree that a test is unbaised if the regression lines of the groups concerned are identical. Under these circumstances race or sex group membership plays no role.

Jensen's (1980: 379) definition of bias in predictive validity is more comprehensive than Cleary's and all the most important elements that play a role in bias are named in statistical terms:

> *A test with perfect reliability is a biased predictor if there is a statistically significant difference between the major and minor groups in the slopes b_{yx}, or in the intercepts k, or in the standard error of estimates $SE_{\hat{y}}$ of the regression lines of the two groups. Conversely, an unbiased test with perfect reliability is one for which the major and minor groups do not differ significantly in b_{yx}, k or $SE_{\hat{y}}$.*

From this definition it is clear that the emphasis falls on bias in the predictive validity of the test and not on some inherent characteristic of the test itself. If homogeneity of regression (the Cleary model) is not present, the following four conditions can occur:

1. The slopes (regression coefficients) differ.

2. The intercepts differ.

3. The slopes and intercepts differ.

4. The standard errors of estimate differ.

Bias owing to the different slopes of the regression lines of two groups leads to an underprediction of the criterion score of the group with the higher mean criterion score when a common regression line is used.

Bias owing to different intercepts (the intercept of a regression line refers to the point at which the regression line cuts the Y-axis) while the slopes are the same leads to a systematic underprediction or overprediction of the criterion score by the test. A test with the same validity coefficients for two groups can still be subject to intercept bias.

If both the slopes and the intercepts of the groups differ, the amount of bias in the prediction can vary owing to the use of a common regression line. The direction of the bias can even be reversed.

Bias in respect of the standard error of estimate: The standard error of estimate is an index of the number of errors (residuals) made in predicting the criterion, in other words the scatter of the obtained criterion scores around the predicted criterion scores. The standard deviation of the residuals is known as the standard error of estimate (SE_{est}).

To summarize: Bias in predictive validity can be described as

- a type of invalidity that affects one group more than another;
- group differences in test performance which are not also found with respect to the criterion;
- constant and systematic errors in the prediction of a criterion - errors usually associated with group membership;
- unfair discrimination against the group whose criterion score is underpredicted - in reality this group performs better on the criterion than predicted by the test scores.

4.3 Findings regarding bias in predictive validity

In a study with the Junior Aptitude Tests (JAT) it was found that differential and single-group validity were largely absent (Owen 1989a). This finding corresponds with most overseas findings in this regard. Jensen (1980: 515) even refers to ... *a nonexistent phenomenon* ... in this regard. Hunter, Schmidt and Hunter (1979) conducted three validity studies and came to the conclusion that perhaps true differential validity does not exist. According to these writers the fact that they could not find any convincing evidence of the phenomenon of differential validity argues strongly against a conception of test bias that is based on the assumption...*that the meaning of test content differs by race* (p. 733). Contrary to the above writers, Katzell and Dyer (1978) maintain that it is too soon to reject the matter of differential validity completely. The fact that differential and single-group validity are seldom encountered or are difficult to identify suggests to Flaugher (1978) that these sources of bias are not as important as others.

Findings in the USA based on the use of a common regression line indicate that criterion scores for blacks are overpredicted because the intercept of the regression line for whites is usually higher than that for blacks (Jensen 1980). According to Sung and Dawis (1981), when the regressions for whites and blacks differ, the difference can usually be attributed to differences in intercepts and not in slopes, supporting the assertion that differences in the mean test performance of groups are not indicators of bias.

While research in the USA has produced little evidence of predictive bias, these results do not necessarily apply to South Africa because language and SES differences between groups are probably more pronounced here than in the USA.

4.4 Fair selection models

A number of models have been proposed for the fair selection of applicants for employment or to attend an educational institution. These models have been extensively discussed in the literature (see e.g. Petersen 1980) and only some of the more important ones will be mentioned here, namely the

- regression model
- equal risk model
- constant ratio model
- conditional probability model
- modified criterion model
- threshold utility model.

All these models make use of a variable (test) and a criterion score predicted by the variable. They also have the common goal of making the selection of members of different racial, ethnic, sex or other groups as fair as possible. In this regard, Petersen (1980: 119) comments as follows:

> *The issue of fair selection is very closely tied to society's concern for equality of opportunity for all. Originally, equality of opportunity was viewed as selection of those individuals most likely to succeed. It is this conception of fair selection that underlies the regression and equal risk models. In the last few years, however, the issue of fair selection has been viewed*

as a means of minimizing inequality of opportunity by providing preferential treatment for disadvanged groups. But, who will receive preferential treatment and by how much is not easily agreed upon since relative advantage to some necessarily means less advantage to others.

The search for a truly fair selection procedure involves a search for a commonly accepted value position. Hunter and Schmidt (1976) differentiate between three value positions that subsume all the fairness models: unqualified individualism, quotas, and qualified individualism.

Unqualified individualism requires that selection be fair towards the individual, regardless of group membership; use is made of any variable, including race and sex, that will enhance the predictive correlation.

The quota position accepts some loss of predictive validity in favour of social and moral priorities. In extreme cases, predictive validity is completely ignored and applicants from each group are selected in proportion to their group's percentage of the population.

Qualified individualism tries to maximize predictive validity while excluding race as a predictor, even if it might be a valid one. Thus, a common regression equation is also used in instances where separate equations could have been more successful.

From the above it is evident that while fair selection is not itself a characteristic of tests, it is nevertheless of tremendous importance in a multicultural society like South Africa. Any organization which employs members of different groups and uses psychometric tests in the selection of those employees should consider very carefully the fairness of their selection procedures, especially the moral and ethical issues involved.

5.0 BIAS IN THE CONSTRUCT VALIDITY OF A TEST

A test's construct validity refers to the extent to which the test measures a theoretical construct or characteristic such as intelligence, mechanical insight or aptitude for music (Anastasi 1976). This reasonably simple definition conceals an extremely complex theme, as can be seen from the fact that Messick (1980) distinguishes about twelve different types of this validity.

The fact that a test's predictive validity for a minority group is satisfactory is no guarantee that the same can be said of its construct validity. The two types of validity are determined in completely different ways. In contrast to predictive validity, where the correlation between a test score and a criterion score is regarded as an adequate indication of validity, Cleary *et al.* (1975) regard the evaluation of construct validity as a complex combination of logical and empirical actions on the basis of which both the test and its underlying theory (in other words the construct) are evaluated at the same time. This suggests that it is essential when determining construct validity to evaluate several internal aspects of the test. It is also important to note that none of the internal criteria as such provides adequate grounds for making a categorical statement about the construct validity of a test. In this regard Jensen (1980) also points out that a test score is merely an attempt at measuring a specific construct. Nevertheless, when all the internal characteristics (criteria) of a test are considered one can obtain an indication of the extent to which the test succeeds in measuring a specific construct.

5.1 Definitions of bias in the construct validity of a test

Reynolds (1983: 245) defines bias in construct validity as follows:

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

This definition contains a general assumption that can be studied from various points of view. Several writers including Petersen (1980), Bond (1981), and Sundberg and Gonzales (1981) mention that bias in construct validity means that a test measures one thing in one group and something else in another group while the assumption is that the same construct is being measured. Poortinga (Van der Flier & Drenth 1980) refers to the above as functional equivalence and states that it is only one of three possible forms of test score comparability between groups. The other two are score equivalence (the construct or ability should be measured on a comparable scale for the groups) and item equivalence (the relations between the items should be similar for the various groups).

In Jensen's (1980) definition of construct validity another element is added to those mentioned above. According to him bias exists when there are systematic errors that are associated with the group membership of the individuals.

As far as the incidence of construct bias is concerned Scheuneman (1981) maintains that although tests are basically valid (in other words, there is no bias in construct validity), bias is often found in the form of an underestimation of the abilities of minority groups. The latter assumption is supported to a certain extent by the conditional probability definition of Loyd (1983: 3):

When individuals from different groups (cultural, racial, sexual, etc.) with the same level of ability have different probabilities of success on a particular test, the test is said to be baised.

Scheuneman and Loyd's views therefore imply that although tests basically measure the same constructs in different groups, bias means that the total test scores of members of certain minority groups underestimate their "real" abilities (owing to systematic errors related to the group membership of the individuals).

5.2 Findings regarding bias in construct validity

The question of bias in the construct validity of tests was extensively researched in the USA with a great variety of tests, for example the Wechsler Intelligence Scale for Children (WISC), revised WISC, Otis-Lennon School Ability Test, California Achievement Tests, Scholastic Aptitude Test, Peabody Picture Vocabulary Test, Stanford-Binet, Lorge-Thorndike, Raven Progressive Matrices and Iowa Test of Basic Skills. A common finding was that the tests measured essentially the same factors in different groups and were largely unbiased in their construct validity. This conclusion supports an assertion made by both Jensen (1980) and Reynolds (1980), namely that American black-white differences with regard to mental tests are real and cannot be ascribed to cultural bias in the tests. When differences do occur in the factor structure of the groups, they are mainly in the magnitude of the loadings, with smaller loadings for blacks either indicating that the factor is not as well defined within the black group as it is in the white group, or that more error or specific variance occurs in the test scores of blacks than in those of whites (Jensen 1980). One should however take note of Scheuneman's (1981) warning that although tests are basically valid for different groups (indicating the absence of bias in construct validity), bias is sometimes found in the form of an underestimate of the abilities of minority groups.

In general, factor analysis of test scores has shown that although a certain amount of unique variance can be ascribed to cultural factors, the basic factor structure of tests is largely similar for different cultural groups. It has also been found (Sung & Dawis 1981) that race and sex influence the difficulty value of ability tests but not their factor structures.

Findings in South Africa largely agree with those in the USA. Bias in construct validity was investigated by administering the Senior Aptitude Test (SAT) to technikon students (first-year engineering and physical science technicians) from various population groups. It was concluded that, although there were considerable mean test score differences, the tests measured essentially the same traits in the different groups (Owen 1986).

In another study (Owen 1989a, 1991a & b), the Junior Aptitude Tests (JAT) were administered to Standard 7 pupils from the various education departments. The results did not contradict the hy-

pothesis that the JAT measures the same constructs in the different groups. The dimensions underlying the abilities measured by the JAT are to a large extent similar for the different groups. This conclusion has important implications for the construction of common tests for different population groups: Differences between the mean ability levels of groups are large but nonetheless not so fundamental that different psychological principles are required to explain the test behaviour of the groups. This further implies that differences in test performance as such do not necessarily affect the construct validity of a test.

The main finding of the JAT investigation, namely that population group membership affected the difficulty value of ability tests but not necessarily their factor patterns, concurs with the results of several researchers abroad, for example Stodolsky and Lesser (1967). The results also support the viewpoint of Hakstian and Vandenberg (1979) that there is greater similarity in the cognitive structure of different cultural groups than is generally believed.

Based on findings similar to those of the above-mentioned investigation (for example the absence of bias in construct validity), Jensen (1980) and Reynolds (1980) conclude that the differences in mean test performance of whites and blacks in the USA signify actual differences in the ability of these groups. However this conclusion cannot simply be generalized to the population groups in South Africa. For example the results of the above-mentioned investigation showed that language played a prominent role in all the tests containing language items, having a negative effect on the performance of the black testees in particular. The findings also showed that the absence of bias in construct validity did not exclude the possibility of biased items systematically underestimating the ability of a group, as a result of language or other factors. When the purpose is to measure language ability as such, as in the case of the JAT 4 (Synonyms), the discrepancy between the groups probably signifies an actual difference in respect of the specific ability. Such a discrepancy is not necessarily a result of bias. To say that it is, would be the same as saying that all spelling tests are biased against poor spellers or that all arithmetic tests are biased against those who cannot add or subtract.

6.0 ITEM BIAS

The aim of research into bias is not just to provide test compilers with guidelines and procedures for identifying and eliminating apparently biased test items. A more particular aim is to identify variables or factors that may be responsible for bias with respect to specific groups (Schmeiser 1982). So far two main approaches have been followed in studying the problem of item bias, namely the judgmental approach (Tittle 1982) and the statistical approach (Angoff 1982 and others). Many test manufacturers in the USA, such as the Educational Testing Service and the Psychological Corporation, use both methods to identify biased items (Carlton & Marco 1982; Coffman 1982; Lenke 1982). A general lack of agreement between the statistical and subjective evaluations of item bias (Plake 1980) is a particular problem, and experts are often unable to explain what causes bias (Shepard 1981). The assertion made by Burrill (1981: 143) in this regard is a typical conclusion: *The item data can show how bias occurs but not why.* The cardinal question is therefore: What causes an item to be biased for a specific group?

6.1 Definitions of item bias

Investigations of item bias are aimed at determining whether different ethnic, racial or cultural groups display different behaviour patterns with respect to test items (Mellenbergh 1983). A typical statistical indication that a test item may not be appropriate or suitable for a specific cultural group is that the item is obviously too difficult (or too easy) for the group concerned (Thorndike 1982). According to one definition an item is regarded as biased if it favours one group unfairly over another group (Wilcox 1984/85), or if it systematically functions differently for different groups (Burrill & Wilson 1980).

According to Cleary and Hilton (1968: 61) an item is biased for members of a group

...if, on that item, the members of the group obtained an average score which differs from the average score of other groups by more or less than expected from performance on other items of the same test.

According to this definition item bias can be identified only on the basis of the relation between the item concerned and other items in the test. Further to this, Shepard, Camilli and Averill (1981: 317) state that methods for disclosing item bias

...identify deviant or anomalous items in the context of other items.

Cleary and Hilton's definition is based on features of classical test theory and is unconditional with respect to assumptions about the ability level of the testees, but more recent definitions of item bias are increasingly based on item response theory (latent trait theory) (Humphreys 1986). An important feature of these later definitions is a precondition regarding ability level: only testees with the same level are compared with one another (Mellenbergh 1983). The following is a typical definition of item bias based on the item response theory (IRT):

...an item is generally considered biased if equally able members of different groups have unequal chances of success on the item (Subkoviak *et al.* 1984: 49).

Definitions and descriptions of item bias based on IRT methods can be summarized as follows:

- An item is biased against members of a group if the expected performance on the item is lower for persons in that group than for persons of a similar ability level in another group.

- In this context bias may be the result of multidimensionality. In other words, the probability of answering an item correctly depends on more than one latent trait and the item does not therefore measure the same characteristic in different groups (Shepard 1981).

- Item bias is a contextual property; the biased item is an anomaly within the context of other items. This means that bias cannot be identified in an isolated item and test items designed to measure the same construct should consequently be studied together (Shepard 1982).

- Item bias studies search for evidence of an interaction between performance on an item and group membership, while differences in the ability levels of the groups are held constant (Traub & Lam 1985).

In conclusion, it should be emphasised that the mere fact that groups differ on a particular item is not necessarily an indication of bias against the lower scoring group. Even if the different groups are assumed to have the same abilities or potential, they can still not be assumed to have had equal opportunities for acquiring the required knowledge and skills. Jensen (1980) refers to this problem as the so-called egalitarian fallacy. The egalitarian philosophy states that all human populations are identical with regard to all cognitive properties or abilities. In support of Jensen, Reynolds (1982) points out that there is no *a priori* basis for the claim that one group does not differ from another as far as cognitive abilities are concerned. Shepard (1982: 13) summarizes the viewpoint of those who oppose the egalitarian philosophy as follows:

> *An a priori assumption of equivalent group means, however, has been rejected by most scholars who believe that the existence of a difference between groups is not automatically a sign of bias.*

One must however remember that even though differences of mean test scores between groups may exist with regard to an ability test, there may be, and often is, considerable overlap in the different frequency distributions of the test scores for the groups in question.

In order to identify the variables that are potentially responsible for item bias it is necessary to determine whether bias can be associated with a certain type of item - in terms of content as well as format (Linn & Harnisch 1981).

6.2 Item content

6.2.1 KNOWLEDGE

According to Scheuneman (1978: 6) the main source of item bias is content ... *which presupposes experience or knowledge not equally available to members of different groups for reasons of cultural background or economic disadvantage.* However Jensen (1980) maintains that this plays a relatively minor role. The removal of such items is also not always advisable because it could affect the predictive validity of the test, for instance in the case of a language test for Puerto Ricans in the USA.

In the case of Test I (Classification) of the JAT, a common element in most of the items that displayed large indices of bias was that they presupposed some knowledge on the part of the testee, for example knowledge of tools (pliers, screws and nails; garden tools), instruments (electrical appliances; old-fashioned and modern lighting; magnifying instruments, such as a microscope; equipment for catching fish, mice and insects), and general knowledge (for example of different kinds of antelope and Western type of ladies' accessories).

6.2.2 FORMULATION

The way in which an item is formulated can be to the advantage or disadvantage of some groups. In this regard, the following aspects have been identified:

- Information in the item. On the basis of the knowledge that American blacks generally have a less well-developed vocabulary than American whites, Scheuneman (1983) speculates that white testees tend to use the information contained in an item more than black testees do when selecting responses, whether through elimination of some of the distractors or by means of another strategy. Handrick and Loyd (1982) as well as Plake and Huntley (1984) found that some pupils react to subtle grammatical cues in the item stem and distractors when answering an item. This finding is supported by results obtained with testees from different groups in the RSA. In this regard, Owen (1989b) reports that white testees were generally not only better able to understand items due to their superior knowledge of the

language, but were also able to use subtle grammatical cues, unintentionally provided by the item-writer, in the stem and distractors of items.

- Negatively phrased items. Scheuneman (1980) analysed more than 2 000 items of the Metropolitan Readiness Tests and the Otis-Lennon School Ability Test and found that American blacks apparently performed less well when the stem of the item was phrased negatively or when one of the possibilities was FALSE or something similar. However Burrill (1981) maintains that Scheuneman's finding is the exception rather than the rule. Dudycha and Carpenter (1973: 120) also found that negatively phrased items were more difficult than positively phrased items, perhaps because *..the negative stem is a departure from expectation and requires a shift of mental set, which test takers fail to do.* This finding is strongly supported by research involving Standard 7 pupils in the RSA (Owen 1989b).

- False or untrue answers. In the RSA it has also been found that the groups under investigation performed considerably worse on items requiring the identification of the one false or untrue answer than in similar items requiring the one true answer (Owen 1989b). In the same vein, Dudycha and Carpenter (1973) found that items where one of the distractors was comprehensive (for example "none of these") were more difficult than items in which the distractors were specific - perhaps because an item with a comprehensive distractor demanded more knowledge from the testee: not only did he have to know what the correct answer was, but also which answer options were incorrect. However the inclusion or omission of the category "none of these" did not affect the reliability or validity of the test (Williamson & Hopkins 1967).

The above-mentioned results seem to reflect a generalizable principle that should be taken into account by test constructors, particularly in the South African situation.

6.2.3 MATHEMATICAL PROBLEMS

Mathematical problems, particularly when verbally formulated, often appear to be biased with respect to American blacks (Linn & Harnisch 1981), perhaps because more than one ability is required to solve such items correctly (Reckase 1985). Simple mathematical problems, such as

determining the square root of an integer, are extremely difficult for minority groups. (This type of knowledge is generally acquired at school). On the other hand, items involving everyday mathematical knowledge such as how to count money, are experienced as relatively easy (Ironson & Subkoviak 1979). In general it appears that American blacks perform less well than whites with quantitative material (Scheuneman 1979). The mistakes made (by all groups) are often systematic, for instance the consistent use of a multiplication sign to denote summation (Tatsuoka & Linn 1983).

The following are some of the most common errors made by testees in the JAT (Owen 1989a):

- In answer to the question "What is the smallest number that can be subtracted from 35 in order to make it exactly divisible by 8?" three times more black testees than white testees did not indicate the number to be subtracted but the number that would be divisible by 8.

- For the question "The sum of three successive numbers is 27. What is the smallest number of the three?" a relatively large percentage of the black testees simply chose the distractor in which the smallest of the five given numbers appeared. A possible reason for this shortcut is that many testees did not understand what was meant by "...sum of three successive numbers".

- With respect to number series, it was found that those series in which the increase is constant, for example 2 4 6 ?, caused remarkably few problems. When, on the other hand, there was an irregular increase, for example 1 3 6 10 (15), many testees tended to ignore the first numbers in the series and based their answers on the last two numbers only. This also occured in items containing a double series, such as 1 8 2 7 3 (6), where the 1st, 3rd and 5th numbers formed one series and the 2nd, 4th and 6th numbers another.

A final finding worth mentioning is that language - and specifically the story element - was almost certainly responsible for the problems that many testees (irrespective of group membership) had with arithmetical or mathematical "story" problems. This finding indicates that it is inherently more difficult to follow the "story" of the problem than it is to perform the numerical calculations involved

- probably because the solution of this type of item requires more than one ability. It also appears from reports P-96 and P-106 that a considerable number of black pupils did not have the numerical skills and knowledge required to solve more complex mathematical problems (in other words problems demanding more than the most basic skills). These results imply certain problems for the test constructor: the mean test score differences between groups in the case of story problems could perhaps be decreased by manipulating the language, but numerical calculations, where poorer performance is due to a lack of basic skills which the majority of testees should have, leave the test constructor almost no room for manoeuvring.

6.2.4 TIME-RELATED CONCEPTS

One of the hypotheses of the investigation described in Report P-106 was that black testees would perform relatively less well than white testees on test items in which time-related concepts played a role. This supposition was based on an earlier (and very tentative) observation that black technikon students seemed to experience problems with such test items (Owen 1986). The underlying rationale can be found in Rohrl's (1979) view that one of the cultural correlates of right brain functioning in certain communities is that the community members have little need for watches, calendars, etc. Therefore if black testees are more right brain orientated, they may be expected to perform relatively poorly on items containing time-related concepts.

The results required that the above-mentioned hypothesis be rejected. Compared with the white testees, the black testees did not perform any worse with time-related concepts than with others. However this result draws attention to another very important matter, namely the danger of an *ex post facto* approach in identifying biased items. The investigation reported in P-106 compared time-related items with similar items of which the content was not time-related. Without this comparison, the conclusion would have been that items involving time are problematic for black testees. Since on average only 29% of the black testees answered the time-related items correctly and all these items displayed large bias index values, it seems natural to conclude that time-related concepts were responsible for the bias. However the necessary perspective is provided by the fact that similar "other" items were also answered correctly by an average of only 25% of the testees and also displayed large bias indices.

Since bias against black testees was found in items with and without time-related concepts but with the same item format, one must conclude that this bias cannot be ascribed to any one single cause but to a combination of item format and item content effects. A study of the item difficulty values (p values) revealed that (for the black testees)

- contents in the form of series functioned better than contents given in an analogy format, and
- the story type item (format and content) led to very poor performance regardless of whether numerical values were expressed in numbers or in words or whether the problem involved concepts of time or money.

As far as the story type item is concerned the underlying problem (bias factor) was definitely language: if the testee did not understand the context and interrelationships of the words, it made no further difference whether the numerical content was expressed in numbers or in words. The fact that black testees performed equally well on analogy items expressing the numbers in words and in figures supports the conclusion that it was the story element that caused the most problems. Although some of the testees were apparently not very familiar with the new form of expressing time (e.g. 16h30 rather than 4:30 pm), the impact of this problem cannot be compared with that of language difficulties.

6.2.5 DIRECTION AND ORIENTATION

In Report P-106 it is noted that, considering the difference in their knowledge bases, black testees did not perform worse than their white counterparts on items concerning direction and direction orientation. However the crux of the matter is that the knowledge base of the black testees was considerably more limited than that of the white testees, as revealed by the large percentage (58%) of black testees who were unable to indicate correctly the direction in which an arrow pointed. Items based on a knowledge of compass directions are therefore inevitably biased with respect to black testees. Although one can only speculate at this stage about the reasons for this lack of elementary knowledge of compass directions represented on paper, it is reasonable to assume that the school and general reading experience of black testees is more limited in this regard than that of white testees. Difficulties in orienting themselves with regard to a direction

represented on paper could indicate field dependence, but there is as yet insufficient information to successfully pursue this thought.

The most important point made by the results of this study is that the test constructor who wants to narrow the gap in test performance between white and black testees should not include items involving compass directions in tests that primarily measure reasoning ability.

6.2.6 VERBAL ITEMS

- Synonyms, antonyms and analogies: Of the three commonly used verbal item types, antonyms and analogies have the higher incidence of bias with respect to American blacks (Scheuneman 1980), and are also often accompanied by language problems (Valencia & Rankin 1985). According to Scheuneman (1980: 147) American blacks appear to have difficulty with comparative terms such as "fewer", "closer" and "larger". This probably indicates the presence of certain language deficiencies. In the same regard Bond (1981) points out that it is impossible to separate vocabulary from verbal-analogical reasoning despite the conceptual difference between them. As far as sex differences are concerned it was found that antonyms are more often biased with respect to girls (Donlon, Hicks & Wallmark 1980); that antonyms and analogies with practical or scientific content favour men, and that women do better if these items involve human relations (Stricker 1982).

  The findings reported in P-106 do not concur with the above-mentioned results - neither for black nor white Standard 7 pupils. Although both white and black testees found synonyms and antonyms more or less equally difficult, there was a remarkable difference between the performance of the two groups. An average of 64% of the white testees answered the antonym items correctly as opposed to 18% of the black testees. In the case of synonyms the groups' mean percentages were 59% and 14% respectively. In addition to the limited number of items that were used (with more items the findings might have been reversed), a possible reason for the black testee results is that the small proportion of testees who answered either of the two item types correctly left little room for the two percentages to differ significantly from each other. The fact that the black testees found both item types equally

difficult, that only a small percentage of them could answer either of the two item types correctly and that both types were equally biased, emphasizes the problems black testees experience with vocabulary items.

Of all the tests in the JAT, Test 4 (Synonyms) gave black pupils the most problems (Owen 1989a). It seems that in addition to displaying a lack of vocabulary, a large percentage of these pupils tended to form an association with the stimulus word rather than provided a synonym for it. Since it was the objective of the test to measure knowledge of vocabulary, the fact that the test as a whole was not suitable for the black testees cannot be attributed to bias as such, but rather to deficient knowledge of the language.

- Verbal Analogies: As in the case of number series a large percentage of the black testees did not appear to consider the analogy as a whole, but based their answers on the second part only. The more difficult the item, the stronger this tendency was. It is also evident from the results in Report P-96 that a considerable number of testees simply looked for a word associated with the last term in the analogy. Futhermore, it is striking to note that a large percentage of the pupils were attracted to the most improbable of the five answer options.

- Word Classification: Examples of this type of item are "Which word fits in least with the other four?" and "What goes best with a (object) and a (object)?"

  Apart from a lack of specific word knowledge, the biggest problem appeared to be the fact that many pupils did not classify the objects on the basis of conceptual similarity, but on the basis of associative resemblance. Poor performance could also be ascribed to an insufficiently critical or logical attitude. Apparently it did not occur to many testees when they chose a specific answer that other distractors might very well be correct answers too.

- Shuffled letters: Some items of the JAT2 (Reasoning) consist of letters that must be rearranged to form a word. From the responses it seemed that a large percentage of the black pupils in particular answered the questions without reshuffling the letters. Two possible reasons can be suggested for why so many black testees experienced difficulties with these

items: (1) they did not understand the instructions and therefore did not know exactly what to do, or (2) they understood the instructions but were not able to reshuffle the letters to form a word. Language difficulties are a factor in both of these explanations.

- Sayings: Many testees simply did not have the knowledge and experience required to interpret information figuratively rather than literally. This is illustrated very well by the following item.

  "Which conclusion can be drawn from the statement below?"

  Statement:

  "If you do not have wings you cannot fly."

Four times as many black testees as white testees selected distractors that indicated a literal interpretation (Owen 1989b). It is probable that items containing sayings are unsuitable for black testees because the reasoning process by which they are solved requires a level of familiarity with the language that these testees do not possess. To white testees, responding to such items would merely be a question of knowledge.

To summarize: For many testees verbal item bias can be ascribed to the factor of language. However language deficiencies are not the only reason why a large percentage of the black testees did not stand the same chance of answering an item correctly as their white counterparts. The limiting factors do lie not so much in the type of item as they do in the form of test behaviour. Important issues in this regard are *inter alia* a tendency to associative conceptualization, selective attention to the facts in an item, a carelessness concerning discrepancies and an uncritical attitude. The latter three point to an absence of logical attitude.

6.2.7 VERBAL VS. NON-VERBAL ITEM CONTENT

In the study dealt with in Report P-106, it was expected that groups would differ with regard to the way in which their performance was influenced by verbal and non-verbal items. This hypothesis was studied with the aid of two item formats, analogies and series. The verbal items were presented as verbal analogies and letter series and the non-verbal items as figure analogies and number series. The two content types within each format were compared, in other words verbal

analogies were compared with figure analogies and letter series with number series. The hypothesis was confirmed to some extent in the analogy comparisons but not in the series comparisons. As regards analogies, it seemed that the performance of black testees was not improved by the mere replacement of words with figures. The bias indices clearly revealed that an equal number of verbal and figure analogies were biased. Nethertheless the largest indices were found in respect of verbal analogies, supporting the findings of researchers such as Scheuneman (1980). Verbal analogies also displayed a greater incidence of bias than figure analogies in the case of Indian testees.

As regards series, the population groups not only reacted similarly, but the two types of series (letters and numbers) did not differ in their effect on test performance. Irrespective of their content, series items are probably a relatively suitable format for both Indian and black testees and few biased items were found for either of the two population groups. This finding, admittedly based on a limited number of items, therefore does not support Scheuneman's (1981) observation that bias is found more often in letter series than in other types of series.

A general problem concerning bias in non-verbal items is that the causes of the bias, or the reasons for the items' differential functioning, are far less "visible" than in the case of verbal items.

6.2.8 CONCRETE VS. ABSTRACT ITEM CONTENT

The same study (Report P-106) was expected to reveal not only that concrete and abstract items would differ in the effect they had on test performance, but that the various groups would also differ in their reactions to these items. It was also expected that, because of the lower standard of their socio-economic environment, certain testees would perform better on concrete than abstract items. According to reports P-106 and 1991/1 (Owen 1989b, 1991a) neither of these speculations were confirmed. The results suggested that the performance of white testees was not affected by content type. In the case of black testees the matter is more complex because the item format used to test these hypotheses, namely the syllogism, was not at all suitable for the black testees. As a result of the problems they experienced with the item format, it is not possible to say whether the black testees performed better on concrete than on abstract items.

6.3 Item format

The format in which an item is presented, relatively independent of content and context, is another possible cause of bias. According to Scheuneman (1981), figure series are often more biased than figures presented in a different format. Bias is also more often found in letter series than in other types of series. Bias of this nature cannot, however, be adequately traced unless more than one type of item format is used to measure the same underlying ability.

6.3.1 OPEN-ENDED ITEMS

According to Frederiksen (1984b) modifying the format of an item, for example from a multiple-choice question to an open-ended type, has only a minor effect on the construct being measured. However such changes can make an item more difficult. In this regard Dudycha and Carpenter (1973) found that items with an "open" stem (for example "steam is a form of ...") were more difficult than those with a closed stem. No explanation could be found for this phenomenon.

6.3.2 SYLLOGISMS

According to Hunt (1982) reasoning/thinking can be described as a set of intellectual skills linked to specific content areas. For example persons who are able to work out syllogisms whose content is familiar to them are often unable to answer similar syllogisms expressed in an abstract manner. This observation has been confirmed by cross-cultural research. Haggard (1954) also points out that of all item types syllogisms are the most difficult to modify in order to decrease "middle-class bias", as they are largely academic, artificial and completely alien to the everyday world of the lower-class child.

The syllogism (concrete as well as abstract) is probably an unsuitable item format for the black testees described in Report-106 for the same reason as that identified by Haggard (1954) with regard to low SES children. On the basis of the findings of an experiment described by Brislin (1983) it can also be concluded that if syllogisms are formulated in an African language and the black testees are given enough practice, their performance will improve considerably. For the present however syllogisms, except perhaps for those comprising concrete-meaningful material, should

be avoided in a common test for black and white testees. Although all syllogism types were generally suitable for Indian testees, concrete-meaningless, concrete-unreal and abstract contents appeared to be somewhat less suitable than, for example, concrete-meaningful material.

6.3.3 STATEMENTS

The following is an example of a statement-type item format:

> Which true conclusion can be drawn from the statement below
> if only the information in the statement is taken into account?
>
> **Statement:** No hunter is a gamewarden.
> McPhearson is a hunter.

The results in Report P-106 reveal that the black testees performed worse on verbal reasoning items in which a statement was made than on other verbal item types. Furthermore, up to six times as many white testees as black testees answered some of these items correctly. All statement items also indicated quite large bias indices. The fact that on average, statement items were answered correctly by 10% fewer black testees than were other verbal item types can naturally not be attributed to language problems. A more probable explanation is that the same factors that lead to poor performance in syllogisms were involved. In other words, statement type items are too academic, artificial and foreign to the everyday world of black testees. Owing to lack of practice and experience black testees are not used to reasoning within the strict confines of a statement and limiting their thoughts to the facts provided. Seen from another angle, it is not impossible that persons who tend to think holistically - often associated with right-brain functioning (cf. Kaufman 1979) - will experience problems with statement type items (and syllogisms) that demand analytical thinking.

On the other hand, Hale's (1982) assertion that African blacks regard a direct statement as rough and unimaginative cannot be completely ignored. A black testee, for example, will definitely not be motivated to go to any trouble over an item that he regards as an unintelligent verbal construction. However further information in this regard cannot be obtained from a group test situation

and all the possible reasons given for poor performance with regard to statements must therefore be regarded as only tentative.

6.3.4 FIGURE SERIES VS. FIGURE ANALOGIES

Only two item formats, namely figure series and figure analogies were formally compared (Report-106). The same item content was used so as to determine the effect of item format on test performance. The results revealed that all the groups found figure series easier than figure analogies. The extent to which the same item content was made more difficult by the format in which it was presented was clear from the mean proportions (x100=%) in Table 5.13a (Owen 1989b: 53). For example an average of 66,2% of the white testees answered the figure series correctly, while the corresponding percentage for the figure analogies was 47,7% (a difference of almost 19%). In the case of the Indian testees the corresponding percentages were 54,5% and 38,9% (a difference of 15,6%), and for the black testees 36,7% and 27,7% (a difference of 9%). This finding supports the view of Sternberg and Gardner (1983: 90) that

> *...one would expect that for problems in these formats with terms of equal difficulty (i.e. drawn from the same universe of stimuli), analogies would be slightly more difficult than series completion...*

According to these writers series are easier than analogies because fewer information-processing parameters are required to solve them.

Although the series is undoubtedly one of the best item formats for use with black testees, its suitability depends on the complexity or difficulty of the item concerned: the more difficult the item (as measured by the performance of white testees), the less suitable it is for black testees. This finding highlights a fundamental question in many bias investigations, namely whether deviating items are the result of bias (as it is defined) or whether they should be ascribed to the inability of the testees. Although no definite answer can be given to this question, the solution strategies used by testees - as reflected in their choice of item distractors - can provide some clarity in this regard.

6.4 Distractor choice as an indication of test behaviour

Test compilers seldom construct item distractors in an arbitrary way. Because distractors often reflect the typical reasoning mistakes made by testees, the selection of distractors is a valuable source of information about test behaviour. According to Report P-106 black testees displayed certain forms of test behaviour that are detrimental to test performance. In this regard it appears that black testees tended more than white testees to

select illogical answers;

adopt an uncritical attitude;

ignore contradictions, and

pay selective attention to the facts in an item:

- They did not adhere to the information provided in the stem of the item.
- They did not use all the information provided in the stem of the item.
- They based their answers on only one of the two statements in a syllogism.

It was also clear that they lacked the linguistic knowledge and finer understanding of language nuances required to interpret certain material figuratively rather than literally (items involving sayings or adages are therefore not suitable for black testees). Lack of arithmetical knowledge and skills also played a major role in calculation problems.

These forms of test behaviour, which are reasonably typical of test-unsophisticated persons of a low SES, contributed greatly to the identification of deviant and therefore biased items. In a previous report concerning the same testees (Owen 1989b) the above mentioned aspects of test behaviour were identified as causes of apparent item bias - the statistical indices show that items are biased, but the real causes of the bias are located in the test behaviour of testees rather than in the items as such.

One of the most important causes of true item bias is the role of language in verbal constructions such as word classification, linear syllogisms, verbal analogies and mathematical story problems. Language problems prevented the black testees from revealing their true potential with respect to these item types. Any test consisting mainly of such items will necessarily be biased for these testees.

7.0 RECOMMENDATIONS

To make a common test battery for white and black testees in South Africa a reality, the considerable gap that presently exists between the test performance of the two groups has to be narrowed. The item content and item format strategies discussed below could contribute to better performance by black testees on verbal and non-verbal reasoning tests. Owing to the complex nature of the problems they address these "solutions" are by no means complete and other factors such as mastering thinking strategies, cognitive enrichment, practice and training will also have to receive attention.

It should be kept in mind that the differences in mean test scores which are found between white and black testees in South Africa, are probably largely a reflection of the enormous differences in the socio-economic conditions and educational opportunities which have existed, and still exist, between these groups. Tests should not be condemned for highlighting these differences. One should rather direct policy and resources in the country to the eradication of educational backlogs and to social upliftment schemes.

Language is indispensable to the measurement of the full spectrum of intellectual abilities required in a technological society. In a multilingual country the test compiler will probably always come up against test bias caused by language deficiencies. However a start can be made towards the establishment of efficient common tests by following the guidelines given below.

Item content

The following should be avoided as far as possible:

- Negatively formulated item stems and answer options.

- Items requiring one false or untrue answer.
- Item distractors that are overly attractive and therefore confusing to less sophisticated testees.
- Subtle cues in item stems and/or distractors, for example

  Someone who has been influenced in making a decision
  based on preconceived opinions, is said to be ...

  A. influential. B. hypocritical. C. prejudiced. D. accusatory. E. impartial.

In this item, the pre in preconceived could prompt smarter testees to select option C (prejudiced).

- Difficult words in items such as word classification and analogies.
- Complicated syntax in mathematical story problems.
- The use of items involving direction and directional orientation (geography tests are a different matter). Items such as the following should therefore be avoided:

  Mac lives 5 km south-west of Tom.
  Sally lives 7 km east of Mac. In which direction
  must Sally walk to visit Tom?
  A. North-west B. North C. South-east
  D. West E. South

- Synonyms and/or antonyms in tests that do not primarily measure language skills.
- Involving basic arithmetic skills in items that are not primarily concerned with measuring numerical skills.
- Test items containing sayings or adages, for example

  "Let sleeping dogs lie."

- Using verbal analogies when figure analogies could be used.

Item format

The following should be avoided as far as possible:

- Syllogisms, for example

  All A are B.

  Some A are C.

  Therefore :

- Test items containing statements, for example

  Which one true conclusion can be drawn
  from the statement below if only the information
  in the statement is taken into account?
  Statement:

  The world's population is increasing at a
  tremendous rate but food production
  remains constant.

Series of any content type (letters, numbers or figures) is possibly the best format. Analogies are less suitable but cannot always be replaced by series without affecting the construct validity of the test.

Distractor attractiveness is another aspect to consider in reducing the role of irrelevant factors in differential test performance. Certain distractors regarded by most white testees as completely wrong are apparently highly persuasive in the case of black testees. Test constructors usually strive to include distractors that are as tempting as possible in order to attract the attention of those testees who do not really know the correct answer. These efforts to mislead pupils usually take the form of imitating typical thinking errors. Despite the fact that this is standard practice in constructing items, it does not only confuse the less sophisticated testees who cannot identify the correct answer, but also reinforces certain undesirable thinking patterns.

It is therefore recommended that test constructors reconsider the procedures and strategies for the construction of test items and that special attention be given to item distractors. Distractors cannot of course be a means of preventing pupils from reasoning illogically but by not providing a distractor that represents for example an associative solution, testees can at least be forced to consider other options as well. In the case of unsophisticated testees attempts to mislead by means of distractors should generally be avoided as far as possible.

Research has shown that a large percentage of black pupils do not have the knowledge or skills to exploit the properties and format of various item types in such a way as to obtain high test scores. Illustrating the principles underlying the answers to the different item types by means of a limited number of practice examples is apparently not sufficient to provide these testees with the necessary insight and comprehension. Based on the views concerning practice and training of, for example, Alley and Foster (1978), Borkowski and Krause (1983) and Frederiksen (1984a), it is recommended that urgent attention be given to the training of language and cognitive skills in general and problem-solving strategies in particular. In order to render the test performance of white and black testees comparable, black pupils must be equipped with a basic "supply" of reasoning strategies on which they can draw to solve problems encountered in the form of, for example, analogies, series completion, classification and statements. It is therefore recommended that training tests be developed which illustrate the principles required for answering the various item types often encountered in ability tests. No testee should take an actual test before proving that he has mastered certain basic problem-solving skills.

As far as a long term solution is concerned, it is recommended that experimental work be conducted on the item formats and item contents responsible for the differential performance of groups.

The macro-approach to studying item bias, for example, by means of group testing, limits the amount of information that can be obtained from testee responses. A micro-approach in the tradition of cognitive psychologists such as Sternberg, Pellegrino, Glaser, Hunt and others makes it possible to study the cognitive processes underlying the answering of items and offers an explanation for the apparently illogical test behaviour of some testees.

To conclude: Although the test developer cannot raise the economic level of the underprivileged group, he can at least try to narrow the gap between the test performances of low and high SES groups by paying special attention to the way in which items are formulated. For example, items should be formulated positively, the "untrue" type of answer should be eliminated by asking which of the answer possibilities is "true", cues should not be given and the invention of confusing, misleading or deliberately attractive item distractors should be avoided.

8.0 METHODS USED BY TEST PUBLISHERS IN THE USA TO "DEBIAS" STANDARDIZED TESTS

In order to make test developers, users and researchers more aware of the issues involved in the differential test performance of groups, some of the strategies used by four leading test publishers in the USA to "debias" tests are presented here.

(1) CTB/McGraw-Hill (Green 1982)

According to Green (1982: 230-1) the position of CTB/McGraw-Hill concerning test bias (most of their tests are achievement tests) is based on some general propositions.

> *First, it is held that when students come to school they may differ in their background knowledge, cognitive and academic skills, language, and attitudes and values. To the degree these differences are great, no one curriculum and no one set of instructional materials will be equally suitable for all, and therefore no one test will be appropriate. It is difficult to specify what amount of difference can be called great, and reasonable differences of opinion on this matter exist.*
>
> *Secondly, it is the task of schools to increase the amount of knowledge that is common to all, to develop certain basic cognitive skills in all students, to generate proficiency in at least the English language among all students, and to foster certain common attitudes and values in our society. Therefore, there is a need for general tests that measure the knowledge and skills taught in school. Thus, the test publisher's task is to develop tests measuring these common skills and bodies of knowledge, without introducing any extraneous elements into the performances on which the measurement is based. If these tests do require that students have knowledge and skills not taught in school, differences in performance among some*

students will occur because of differences in school learning, whereas for other students differences in out-of-school learning will matter. Thus the test is measuring different things for different groups and can be called biased .... .

The third general proposition is that for some groups, notably those from families whose language and culture differ sharply from that of 'Middle America', no test designed to be used nationally can be completely unbiased. The best one can do is to minimize the role of the extraneous elements, thereby increasing the number of persons for whom the test is appropriate. If care is taken in test construction, however, the influence of these elements can be minimal for most people; but for some groups they will continue to play a substantial role. For example, one simply cannot conclude that students are illiterate if they fail an English test when they speak Spanish, since perhaps they are literate in Spanish. Only when such students have acquired a substantial degree of both facility with the English language and acculturation can one expect our standard tests of basic academic knowledge and skills to elicit performances adequately representative of their academic status.

CTB currently uses four procedures to minimise test bias (Green 1982: 233):

- *careful attention to content validity,*
- *the inclusion of bias considerations and the application of the various McGraw-Hill guidelines in the test specifications used by the writers and editors,*
- *bias reviews by both CTB editors and by external experts, and*
- *analyses of item tryout data separately by ethnic group in order to find and delete items that appear to be undesirable for one or more groups.*

(2) Riverside Publishing Company/Houghton Mifflin (Coffman 1982)

A major aim of research on test bias at the Riverside Publishing Company...*is to provide information not only to test developers about how to minimize bias at the test construction stage, but*

also to test users about factors that need to be weighed when interpreting test scores for particular subgroups of the population" (Coffman 1982: 241).

According to Coffman the approach to insuring test fairness consists of five stages, namely,

- *applications of careful professional judgments at the item writing stage,*
- *systematic reviews by representative panels of test users at the test assembly stage,*
- *statistical analyses based on a variety of bias models,*
- *comparisons among the results of stages 2 and 3, and*
- *special follow-up studies designed to seek answers raised by the comparisons at stage 4* (p. 242).

> *Since any one author's knowledge is necessarily limited, it is common practice to have each test question reviewed by several individuals other than the original item writer before the item is placed in the pool of items to be presented. Such reviews, and the discussions that ensue, contribute to increasing the sensitivity of writers to the multiplicity of factors that might contribute to unfairness* (p.242).

(3) The Psychological Corporation (Lenke 1982)

The Psychological Corporation publishes *inter alia* the Metropolitan Readiness Test, Otis-Lennon School Ability Test and the Wechsler Scales. Like other test publishers the Corporation has given considerable attention to the question of bias and fairness in testing. A problem in this regard according to Lenke (1982: 255-6), is that

> *The fairness issue is most often discussed in relation to 'bias' of one kind or another. Unfortunately, no one has yet been able to come up with a definition of bias that is either universally acceptable or universally applicable to all types of test or in all types of situation ...*

Until we can all agree on a definition of bias that can be applied in a meaningful way to the examination of test content and to the interpretation of test performance, there will be no such thing as a 'bias free' test.

Apart from ensuring that their norming samples are representative of the nation as a whole by including all ethnic groups, socio-economic levels, and geographic regions in approximately the same proportions as the national population, The Psychological Corporation also pays special attention to facial bias and item bias. The importance of these two concepts is evident from the following description by Lenke (p. 256):

Facial bias refers to situations in which a test item, or a group of test items, appears to reflect some prejudice, stereotype, derogatory or offensive association, over- or underemphasis of the worth of particular ethnic or sex groups, and over- or underrepresentation of particular types of environmental settings ... . Item bias refers to situations in which a test item functions differently in some systematic way for different groups. While several methods exist for the detection of item bias, it is not yet clear which method, or combination of methods, is most appropriate for the identification of 'unfair' items. For this reason The Psychological Corporation views statistical item bias detection procedures as helpful in evaluating the performance of individual items; however, such procedures are not used in the absence of sound, professional judgment.

The Psychological Corporation, like other test publishers, also makes use of a panel of minority-group educators to review items. As far as ethnic bias is concerned both subjective and objective rating procedures are used. In this regard Lenke (p. 258) states that

The subjective raters were members of an advisory panel of minority educators, selected on the basis of their sensitivity to minority-group concerns and their active involvement in education. The role of the panel members was to provide needed input to each of the five major stages in the development of a major test series: blueprinting, content development, item analysis, standardization, and publication. At each of these stages, panel members were

asked to review the materials for potential problems in content, such as irrelevance or insensitivity to minority students, and unclear or potentially biased artwork.

As objective rating procedures a number of statistical methods are used, e.g. Angoff's delta-plot method.

(4) Educational Testing Service (Carlton & Marco 1982)

According to Carlton and Marco (1982: 278) Educational Testing Service (ETS)

> *... has a longstanding commitment to producing tests that acknowledge the multicultural nature of our society, to avoiding language or content that is offensive to or demeaning of any group within the population, and to attempting to understand and use information about the performance of different groups on test items.*

Among the guidelines for ETS products and services are the following pertinent points which all test developers in multi-cultural countries should bear in mind (Carlton & Marco pp. 278-9):

- *Specifications for tests should require material reflecting the cultural background and contributions of women, minorities, and other subgroups.*
- *Individual test items, and the test as a whole, should be reviewed to eliminate language, symbols or content which are generally considered potentially offensive, inappropriate for major subgroups of the test-taking population or serving to perpetuate any negative attitude which may be conveyed towards these subgroups. No item in any test should include words, phrases or description that is generally regarded as biased, sexist or racist... .*
- *...studies* (should be undertaken) *to determine the sources of significant differential performance of sex, ethnic, handicapped, and other relevant subgroups on ETS tests.*

- *Studies relating item performance to subgroups should be carried out for new or substantially revised tests when there are adequate data concerning sufficient samples of large subgroups whose education and experience may be different from the majority of examinees.*

Currently, the system of sensitivity reviews at ETS include the following (p. 281):

1. *a group of reviewers who undergo training;*

2. *a set of training materials;*

3. *a list of specific criteria for sensitivity reviews;*

4. *a list of words, phrases, and concepts that may signal offensiveness;*

5. *adjudicators to make decisions in instances of intractable disagreement between reviewers and test developers; and*

6. *a steering committee to oversee the entire process.*

According to the authors (p. 282) test items are evaluated from two perspectives: *the cognitive dimension, which deals with the factual accuracy of an item, and the affective dimension, which deals with the feelings an item may evoke in group members ... . Examples of the latter would include items that focus on the high birth rate in Third World nations or on the high suicide rate among Native Americans.* Highly controversial issues (e.g. issues relating to genetic inferiority) are usually avoided in a test unless these issues are

> *... both relevant and essential to effective measurement ... . When both the test developer and the sensitivity reviewer agree that such issues are both relevant and essential, items must be worded in such a way as to make it clear that ETS does not subscribe to the position stated* (p. 282).

Another important point is that

... no item may contain material that reinforces offensive stereotypes, nor may it include underlying assumptions about a group that reflect an individual's ethnocentric beliefs. Examples of the latter would include items that imply an inferiority or deficiency on the part of one or more groups ... .Examples of ... stereotypes would include statements or implications that a particular group is deserving of a particular fate or that a group is by nature dependent on the majority culture (p. 283).

ETS also has a set of criteria, labeled 'context considerations', which contains judgmental guidelines for reviewing material that may be sensitive to some groups but is necessary in testing. The four areas or domains covered are the historical, literary, legal and psychological. Test developers in South Africa should take cognisance of these domains, especially the historical, because they represent material likely to be described as 'offensive' in the 'new' South Africa.

Regarding the historical domain, the authors (p. 283) declare that:

When testing one's knowledge of history, it is sometimes desirable to draw from material written during earlier periods when social values were markedly different from present values. Thus, material that was not considered offensive at the time has become potentially offensive when judged by present standards. For example, a passage describing the condition of Southern Blacks during the reconstruction period may include the term "colored people" or "Negro". While it may be desirable to avoid the use of such material where possible, the sensitiveness of the item must be judged in the overall context in which it is presented.

In conclusion, before test developers or test users label a deviant item as biased they should consider Carlton and Marco's (p. 286) distinction between these two terms:

Statistical procedures can identify items that are deviant, that is, items that seem to behave differently for different groups. However, the content of an item must be taken into consideration in determining whether an item is biased. Differential item performance may be an

artifact of the analysis method; it might be due to unequal exposure to content; it might be due to cultural differences. Bias cannot be inferred without considering the reasons why an item is deviant.

To close this section on the methods used by test publishers in the USA to "debias" standardized tests, reference is made to Scheuneman's (1982: 196-197), very useful and practical recommendations which merit careful attention, especially by test developers. She suggests the following five guidelines to facilitate the item review process:

1. *Prepare item cards for use in the review process. Prepare for your review by having items with high bias indices put onto cards - one item per card. Record conventional item statistics, including the number selecting each option, separately for each group of interest on the backs of the cards. Use computer-generated labels if available; these save clerical labor and are apt to be more accurate. Code items to indicate apparent direction of bias (if direction can be determined). The items that were not identified as biased should be readily available for reference, but need not be placed on cards, although it may be more convenient to do so.*

2. *Sort all items into relatively broad categories. Usually, content categories outlined in the test specifications or blueprint will be appropriate, although item formats or some other characteristic believed to be important may be preferred. Tabulate the number of biased and unbiased items in each category (if this was not done previously). If the distribution of biased items into the categories is very similar to what would be expected if the items had been selected at random, the division may be not meaningful and another classification scheme might be tried. If it does not appear to be random, the concentration into specific categories may be the first clue to possible sources of bias or unexpected performance differences.*

3. *Group biased item cards by content category. Carefully review those items identified as biased, working with one classification category at a time. Examine the items, singly or in groups, looking for item flaws and clues suggesting plausible explanations for the differences found and using the conventional item statistics where they may be helpful. Try to find patterns of differences that may support or disprove some of the possible explanations or that may suggest new*

hypotheses concerning the differences. Do not expect to find an explanation or hypothesis to account for all items. Remember that it is almost certain that some items have been incorrectly classified as biased, and the proportion of such items can be quite high depending on sample size and the decision rule(s) used for selecting biased items.

4. *Verify hypotheses by checking against the set of unbiased items. Sometimes the similarities seen among the biased items are simply characteristics of the test and will be found among unbiased items as well. If hypotheses suggest differences that might generalize across categories or into new item pools, determine whether such differences have occurred. For example, in the certification exam, items worded negatively ("which of the following is not a true statement"), occurred in the biased item pool more often than would be expected regardless of the content of the area.*

5. *Consider what actions might be taken to correct problems revealed by this analysis. If the hypothesis is incorrect and if an action can be taken that should have no adverse impact on scores, such as adding sample items, this can be done without further verification. Otherwise, it may be necessary to involve others in a decision-making process to consider (1) the consequences of making changes if you are wrong versus (2) not making changes if you are right, and (3) the approximate likelihoods of these outcomes. Such discussions may cause a reevaluation of the purposes of the exam, which should be of benefit even if no changes in the test are made as a result.*

9.0 SUMMARY AND CONCLUSION

The purpose of this report is to promote in the test developer and test user a greater understanding of the issues involved, and particularly the issue of bias, when a psychometric test developed for one group is used for another. With this purpose in mind, attention was given to methods for detecting bias, research findings regarding bias, and steps that can be taken to reduce bias in testing.

In Section 2 of the report a brief overview was given of some of the background factors that could play a role in test bias, namely, culture, socio-economic status, language and cognitive style.

In Section 3 methods for detecting bias in predictive validity, construct validity and test items were discussed. As far as item bias is concerned, it was recommended that the item characteristic curve of the item response theory be used whenever possible, although good results can also be obtained by means of the chi-square methods; the TID method and the item discrimination procedure are particularly useful as a quick indication of aberrant - and possibly biased - items.

Bias in the predictive validity of a test was discussed in Section 4. Attention was given to differential and single-group validity as well as to the slope, intercept and standard error of estimate of the regression line. Research findings in the USA indicate that predictive bias is a rather rare phenomenon; however, the criterion scores for minorities are sometimes overpredicted by a common regression line (due to differences in the intercepts for the groups - which are not necessarily an indication of bias).

Bias in the construct validity of a test was discussed in Section 5. Overseas as well as local findings indicate that tests measure essentially the same underlying psychological constructs in

different groups. Construct bias is therefore not really an issue as far as the differential test performance of groups is concerned.

In Section 6 attention was given to item type (in terms of content as well as format) as a possible source of item bias. As far as item content is concerned, it has been found that apart from language factors (and items containing sayings) differential performance is not so much the result of a specific item type but largely of a particular form of test behaviour (e.g. a tendency towards associative conceptualization or an illogical approach). With regard to format, it has been found that syllogisms and statement-type items in particular are unsuitable for some groups.

In Section 7 certain recommendations were made to test developers and some guidelines given regarding item contents and item formats that should preferably be avoided in the construction of common tests for blacks and whites. Section 8 supplemented the guidelines given in the previous Section. In order to make test developers, users and researchers more aware of sensitive aspects involved in the construction of common tests for different groups, some of the strategies used by leading test publishers in the USA to *debias* standardized tests were presented.

The development of unbiased common tests for the various population groups in South Africa is certainly one of the major challenges in the field of psychometrics that has to be met before the turn of the century. This is undoubtedly a formidable task in view of the numerous factors, cultural, political, educational, sociological and economical, which divide South African society and segment its people into different groups.

Groups and group differences are, however, facts of life and the best the test developer can do is to measure those differences as accurately as possible without interference from bias. Ideally, a test should only register individual differences and not group differences as well. Unfortunately, this is not always achieved in practice - one reason being that the items which best discriminate between individuals are sometimes the items that best discriminate between groups. Although the test developer has no control whatsoever over all the factors in the formation of an ability, in the measurement of that ability he can at least ensure that his measuring device is as fair and unbiased as possible.

What action can be taken by the test developer and test user when bias is detected in a test? The following are some of the options.

1. If the test measures different things in different groups (indicating bias in construct validity), the test as a whole is unsuitable as a common test and its use should be restricted to the group for whom it was standardized.

2. On the other hand, if bias in construct validity is absent (which is usually the case) but there are considerable differences in the mean test performances of the groups, then the test user could consider the use of separate norms. However, this line of action is not always advisable and can lead to other problems. Furthermore, it must be remembered that mean test score differences per se are not an indication of bias; they could signify real differences in ability that must be acknowledged as such.

3. Bias in predictive validity, i.e. the over- or underprediction of a criterion score by a test score, can be dealt with by test users using the test for selection purposes by means of separate regression lines for the different groups. It must be remembered, however, that bias in this sense usually works to the advantage of the underprivileged group, i.e. the criterion scores for the latter group are usually overpredicted by a common regression line.

4. Biased test items can be dealt with in various ways. When a new test is being developed for different groups, the 'no bias' responsibility rests with the test developer (see Sections 7 and 8). Biased items in existing tests are more difficult to handle. The following are some of the options:

   a. The test developer can withdraw the old test and replace it with a new one (but the costs involved could be enormous).

   b. The test developer can revise the test, culling all the biased items. If a large number of items have to be replaced, this amounts to the development of a new test.

c. If bias is confined to verbal items, test users are advised to use those scores with caution. In other words, some allowances should be made for groups whose language skills are limited. However, if the primary objective of the particular test is to measure verbal ability, the test scores are probably a true reflection of the groups' verbal competence.

d. The test user can score only the unbiased items. In this case the norms given in the test manual are not applicable and the test user will have to calculate his own norms. There is another unfortunate aspect to this procedure. It has been the experience of quite a number of researchers (including the author) that the elimination of a few biased items from a test is usually insufficient to reduce the mean test score differences between groups in any significant way. This is so because the biased items are not necessarily the most difficult ones. Thus, although the removal of biased items is psychometrically and morally defensible and a 'must' for test developers, the effect of such removal on the mean test score differences between groups tends to be rather small. Furthermore, reducing the number of items in a test can result in lower test reliability and also an attenuation of the construct measured.

From the above it is evident that there are no easy solutions or one best way of handling bias in existing tests. The investigation of bias is a long-term undertaking with many facets, only one of which is the identification of biased items. Of more importance are the reasons why certain item types are more likely to be biased than others. In other words, the emphasis is (or should be) on promoting insight into and understanding of the real nature of bias rather than merely identifying and eliminating aberrant items. It is hoped that this document will make some contribution towards this end.

10.0 REFERENCES

Abrams, N.E. 1979. *A comparison of performance of black and white students on tests of racially orientated information.* Unpublished PH.D Dissertation, Columbia University, (Univ. Microfilms International, Ann Arbor, Michigan, 1982).

Alley, G. & Foster, C. 1978. Nondiscriminatory testing of minority and exceptional children. *Focus on Exceptional Children*, 9, 1-14.

Anastasi, A.. 1970. On the formation of psychological traits. *American Psychologist*, 25, 899-910.

Anastasi, A. 1976. *Psychological testing* (4th ed.). New York: Macmillan.

Anastasi, A. 1985. Some emerging trends in psychological measurement: A fifty-year perspective. *Applied Psychological Measurement*, 9, 121-138.

Angoff, W.H. 1982. Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias.* Baltimore: The Johns Hopkins University Press.

Baker, F.B. 1981. A criticism of Scheuneman's item bias technique. *Journal of Educational Measurement*, 18, 59-62.

Bond, L. 1981. Bias in mental tests. *New Directions for Testing and Measurement*, 11, 55-77.

Borkowski, J.G. & Krause, A. 1983. Racial difference in intelligence: The importance of the executive system. *Intelligence*, 7, 379-395.

Brislin, R.W. 1983. Cross-cultural research in psychology. *Annual Review of Psychology*, 34, 363-400.

Burrill, L.E. 1981. *A comparative investigation into the identification of ethnic bias in items assessing current educational status.* Unpublished PH.D Dissertation, Fordham University, New York. (Univ. Microfilms International, Ann Arbor, Michigan, 1984).

Burrill, L.E. & Wilson, R. 1980. *Fairness and the matter of bias.* Test Service Notebook 36. New York: The Psychological Corporation.

Buss, A.R. 1977. On the relationship between the psychological environment and the development of differences in abilities. *Intelligence*, 1, 192-207.

Carlton, S.T. & Marco, G.L. 1982. Methods used by test publishers to "debias" standardized tests: Educational Testing Service. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias.* Baltimore: The Johns Hopkins University Press.

Cleary, T.A. 1968. Test bias: prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, 5, 115-124.

Cleary, T.A. & Hilton, T.L. 1968. An investigation of item bias. *Educational and Psychological Measurement*, 28, 61-75.

Cleary, T.A., Humphreys, L.G., Kendrick, S.A. & Wesman, A. 1975. Educational uses of tests with disadvantaged students. *American Psychologist*, 30, 15-41.

Coffman, W.E. 1982. Methods used by test publishers to "debias" standardized tests: Riverside Publishing Company/Houghton Mifflin. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias.* Baltimore: The Johns Hopkins University Press.

Cole, M. & Scribner, S. 1974. *Culture and thought: A psychological introduction.* New York: John Wiley.

Coop, R.H. &. Sigel, I.E. 1971. Cognitive style: Implications for learning and instruction. *Psychology in the Schools,* 8, 152-161.

Donlon, T.F., Hicks, M.M. & Wallmark, M.M. 1980. Sex differences in item responses on the Graduate Record Examination. *Applied Psychological Measurement,* 4, 9-20.

Dorans, N.J. 1979. *The need for a common metric in item bias studies.* TM 79-20. Washington, D.C.: U.S. Office of Personnel Management.

Dudycha, A.L. & Carpenter, J.B. 1973. Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology,* 58, 116-121.

Ebel, R.L. 1979. Intelligence: a skeptical view. *Journal of Research and Development in Education,* 12, 14-21.

Eells, K.W., Davis, A., Havighurst, R.J., Herrick, V.E. & Tyler, R.W. 1951. *Intelligence and cultural differences.* Chicago: University of Chicago Press.

Evans, F.R. & Reilly, R.R. 1972. *A study of test speededness as a potential source of bias in the admission test for Graduate Study in Business quantitative score.* Report RB-72-9. Princeton, N.J.: Educational Testing Service.

Federico, P.-A. 1980. Adaptive instruction: Trends and issues. In R.E. Snow, P.-A. Federico & W.E. Montague (Eds.), *Aptitude, learning, and instruction. Vol.1: Cognitive process analyses of aptitude.* Hillsdale, N.J.: Lawrence Erlbaum.

Federico, P.-A. 1983. Changes in the cognitive components of achievement as students proceed through computer-managed instruction. *Journal of Computer-Based Instruction,* 9, 156-168.

Ferguson. G.A. 1954. On learning and human ability. *Canadian Journal of Psychology,* 8, 95-112.

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

Frederiksen, N. 1984a. Implications of cognitive theory for instruction in problem solving. *Review of Educational Research,* 54, 363-407.

Frederiksen, N. 1984b. The real test bias: Influences of testing on teaching and learning. *American Psychologist,* 39, 193-202.

Ghuman, P.A.S. 1980. A study of the concept of equivalence and divergent thinking among four sub-cultural groups of Punjabi children. *International Review of Applied Psychology,* 29, 89-103.

Globerson, T., Weinstein, E. & Sharabany, R. 1985. Teasing out cognitive development from cognitive style: A training study. *Developmental Psychology,* 21, 682-691.

Goldman, R.D. & Hewitt, B.N. 1975. An investigation of test bias for Mexican-American college students. *Journal of Educational Measurement,* 12,187-196.

Green, D.R. 1982. Methods used by test publishers to "debias" standardized tests: CTB/McGraw-Hill. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias.* Baltimore: The Johns Hopkins University Press.

Guilford, J.P. 1980. Cognitive styles: What are they? *Educational and Psychological Measurement,* 40, 715-735.

Guthrie, G. 1963. Structure of abilities in a non-western culture. *Journal of Educational Psychology*, 54, 94-103.

Haggard, E.A. 1954. Social-status and intelligence: An experimental study of certain cultural determinants of measured intelligence, *Genetic Psychology Monographs*, 49, 141-186.

Hakstian, A.R. & Vandenberg, S.G. 1979. The cross-cultural generalizability of a higher-order cognitive structure model. *Intelligence*, 3, 73-103.

Hale, J.E. 1982. *Black children: Their roots, culture and learning styles.* Provo, Utah: Brigham Young University Press.

Hambleton, R.K. 1980. Latent ability scales: Interpretations and uses. *New Directions for Testing and Measurement*, 6, 73-97.

Hambleton, R.K. & Cook, L.L. 1977. Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement*, 14, 75-96.

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R. & Gifford, J.A. 1978. Developments in latent trait theory: Models, technical issues, and applications. *Review of Educational Research*, 48, 467-510.

Handrick, F.A. & Loyd, B.H. 1982. Methods used by test publishers to "debias"standardized tests: The American College Testing Program. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias.* Baltimore: The Johns Hopkins University Press.

Havighurst, R.J. & Breese, F.H. 1947. Relation between ability and social status in a Midwestern community. III Primary mental abilities. *Journal of Educational Psychology*, 38, 241-247.

Horn, J.L. 1976. Human abilities: A review of research and theory in the early 1970s. *Annual Review of Psychology*, 27, 437-485.

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

Humphreys, L.G. & Taber, T. 1973. Ability factors as a function of advantaged and disadvantaged groups. *Journal of Educational Measurement*, 10, 107-115.

Hunt, E. 1980. The foundations of verbal comprehension. In R.E. Snow, P-A. Federico & W.E. Montague (Eds.), *Aptitude, learning and instruction. Vol.1: Cognitive process analyses of aptitude.* Hillsdale, N.J.: Lawrence Erlbaum.

Hunt, E. 1982. Towards new ways of assessing intelligence. *Intelligence*, 6, 231-240.

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.

Hunter, J.E. & Schmidt, F.L. 1978. Differential and single-group validity of employment tests by race: a critical analysis of three recent studies. *Journal of Applied Psychology*, 63, 1-11.

Hunter, J.E. Schmidt, F.L. & Hunter, R. 1979. Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin*, 86, 721-735.

Hunter, J.E. Schmidt, F.L. & Rauschenberger, J. 1984. Methodological, statistical, and ethical issues in the study of bias in psychological tests. In C.R. Reynolds & R.T. Brown (Eds.), *Perspectives on bias in mental testing.* New York: Plenum.

Intasuwan, P. 1979. *A comparison of three approaches for determining item bias in cross- national testing.* Unpublished D.Phil. Dissertation, University of Pittsburgh. (Univ. Microfilms International, Ann Arbor, Michigan, 1982).

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

Ironson, G.H. & Subkoviak, M.J. 1979. A comparison of several methods of assessing item bias. *Journal of Educational Measurement*, 16, 209-225.

Irvine, S.H. 1970. Affect and construct - a cross-cultural check on theories of intelligence. *Journal of Social Psychology*, 80, 23-30.

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

Jensen, A.R. 1974. How biased are culture-loaded tests? *Genetic Psychology Monographs*, 90, 185-244.

Jensen, A.R. 1980. *Bias in mental testing.* London: Methuen.

Katzell, R.A. & Dyer, F.J. 1978. On differential validity and bias. *Journal of Applied Psychology*, 63, 19-21.

Kaufman, A.S. 1979. Cerebral specialization and intelligence testing. *Journal of Research and Development in Education*, 12, 96-107.

Lenke, J.M. 1982. Methods used by test publishers to "debias" standardized tests: The Psychological Corporation. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias.* Baltimore: The Johns Hopkins University Press.

Lesser, G.S., Fifer, G. & Clark, D.H. 1965. Mental abilities of children from different social-class and cultural groups. *Monographs of the Society for Research in Child Development*, 30, 1-115.

Linn, R.L., Levine, M.V., Hastings, C.N. & Wardrop, J.L. 1980. *An investigation of item bias in a test of reading comprehension.* Technical Report No 163. Center for the study of reading. Champaign, Illinois: University of Illinois at Urbana-Champaign. (ERIC Document Reproduction Service No. ED 184 091).

Linn, R.L. & Harnisch, D.L. 1981. Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement*, 18, 109-118.

Lord, F.M. 1980. *Application of item response theory to practical testing problems.* Hillsdale, N.J.: Lawrence Erlbaum.

Lord, F.M. & Wingersky, M.S. 1984. Comparison of IRT true-score equipercentile observed-score "equatings". *Applied Psychological Measurement*, 8, 453-461.

Lorge, I. 1952/53. Intelligence and cultural differences as seen by the test maker. *Teachers College Record*, 54, 190-193.

Loyd, B.: 1983. *The effect of number of ability intervals on the stability of item bias detection.* Paper presented at the Annual Meeting of the Eastern Educational Research Association. Baltimore, February 24-27. (ERIC Document Reproduction Service No ED 247 297).

MacKenzie, A.J. 1981. Level I and Level II abilities in primary school children. *British Journal of Educational Psychology*, 51, 312-320.

Marmorale, A.M. & Brown, F. 1979. Psychological testing of the nonwhite, nonmiddle class child. *International Journal of Group Tensions*, 1-4, 195-200.

McGee, M.G. 1979. Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological Bulletin*, 86, 889-918.

Medley, D.M. & Quirk, T.J. 1972. *Race and subject-matter influences on performance on general educational items of the National Teacher Examinations.* Report RB-72-43. Princeton, N.J.: Educational Testing Service.

Medley, D.M. & Quirk, T.J. 1974. The application of a factorial design to the study of cultural bias in general culture items on the National Teacher Examinations. *Journal of Educational Measurement,* 11, 235-245.

Mellenbergh, G.J. 1983. Conditional item bias methods. In S.H. Irvine & J.W. Berry (Eds.), *Human assessment and cultural factors.* New York: Plenum Press.

Merryfield, M.M. 1985. The challenge of cross-cultural evaluation: Some views from the field. *New Directions for Program Evaluation,* 25, 3-17.

Messick, S. 1980. Test validity and the ethics of assessment. *American Psychologist,* 35, 1012-1027.

Novick, M.R. 1980. Policy issues of fairness in testing. In L.J.T. van der Kamp, W.F. Langerak & D.N.M. de Gruijter (Eds.), *Psychometrics for educational debates.* Chichester: John Wiley.

Oakland, T. 1977. *Psychological and educational assessment of minority children.* New York: Brunner/Mazel.

Oakland, T. & Feigenbaum, D. 1979. Multiple sources of test bias on the WISC-R and Bender-Gestalt Test. *Journal of Consulting and Clinical Psychology,* 47, 919-927.

Olmedo, E.L. 1981. Testing linguistic minorities. *American Psychologist,* 36, 1078-1085.

Osterlind, S.J. 1983. *Test item bias.* Beverley Hills: Sage Publications.

Owen, K. 1986. *Toets- en itemsydigheid: Toepassing van die Senior Aanlegtoetse, Meganiese Insigtoets en Skolastiese Bekwaamheidsbattery op blanke, swart, kleurling- en Indiërtechnikonstudente.* HSRC Report P-86. Pretoria: Human Sciences Research Council.

Owen, K. 1989a. *Test and item bias: The suitability of the Junior Aptitude Tests as a common test battery for white, Indian and black pupils in Standard 7.* HSRC Report P-96. Pretoria: Human Sciences Research Council.

Owen, K. 1989b. *Bias in test items: An exploration of item content and item format.* HSRC Report P-106. Pretoria: Human Sciences Research Council.

Owen, K. 1991a. *Toets- en itemsydigheid: Die geskiktheid van die Junior Aanlegtoetse (JAT) as 'n gemeenskaplike toetsbattery vir standerd 7-leerlinge van twee onderwysdepartemente met spesiale verwysing na iteminhoud en itemformaat as oorsaak van itemsydigheid.* HSRC Office Report 1991/1. Pretoria: Human Sciences Research Council.

Owen, K. 1991b. Test bias: The validity of the Junior Aptitude Tests (JAT) for various population groups in South Africa regarding constructs measured. *South African Journal of Psychology,* 21, 112-118.

Peterson, N.S. 1980. Bias in the selection rule - bias in the test. In L.J.T. van der Kamp, W.F. Langerak & D.N.M. de Gruijter (Eds.), *Psychometrics for educational debates.* Chichester: John Wiley.

Plake, B.S. 1980. A comparison of a statistical and subjective procedure to ascertain item validity: One step in test validation process. *Educational and Psychological Measurement,* 40, 397-404.

Plake, B.S. & Huntley, R.M. 1984. Can relevant grammatical cues result in invalid test items? *Educational and Psychological Measurement,* 44, 687-696.

Radford, J. & Burton, A. 1972. Changing intelligence. In K. Richardson, D. Spears & M. Richards (Eds.), *Race, culture and intelligence.* Middlesex, England: Penguin Books.

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

Reckase, M.D. 1985. The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9, 401-412.

Ree, M.J. 1979. Estimating item characteristic curves. *Applied Psychological Measurement*, 3, 371-385.

Reynolds, C.R. 1980. Differential construct validity of intelligence as popularly measured: Correlations of age with raw scores on the WISC-R for blacks, whites, males and females. *Intelligence*, 4, 371-379.

Reynolds, C.R. 1982. Methods for detecting construct and predictive bias. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias.* Baltimore: The Johns Hopkins University Press.

Reynolds, C.R. 1983. Test bias: In God we trust: all others must have data. *The Journal of Special Education*, 17, 241-260.

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.* New York: Plenum Press.

Rohrl, V.J. 1979. Culture, cognition and intellect: Towards a cross-cultural view of intelligence. *The Journal of Psychological Anthropology*, 2, 337-364.

Rudner, L.M. & Convey, J.J. 1978. *An evaluation of select approaches for biased item identification.* Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, March 27-31. (ERIC Document Reproduction Service No. ED 157 942)

Rudner, L.M., Getson, P.R. & Knight, D.L. 1980. Biased item detection techniques. *Journal of Educational Statistics*, 5, 213-233.

Rudner, L.M. & Getson, P.R. 1982. *Item bias research has its limitations.* Paper presented at the Annual Meeting of the National Council for Measurement in Education, New York, March.

Scarr, S. 1981. Testing for children: Assessment and the many determinants of intellectual competence. *American Psychologist*, 36, 1159-1166.

Scheuneman, J.D. 1978. *Further considerations in the assessment of bias in test items.* Paper presented at the Meeting of the American Psychological Association, Toronto.

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

Scheuneman, J.D. 1980. Latent-trait theory and item bias. In L.J.T. van der Kamp, W.F. Langerak & D.N.M. de Gruijter (Eds.), *Psychometrics for educational debates.* Chichester: John Wiley.

Scheuneman, J.D. 1981. A new look at bias in aptitude tests. *New Directions for Testing and Measurement*, 12, 3-35.

Scheuneman, J.D. 1983. *Bias in test items: Causes and effects.* Paper presented at the Annual Meeting of the American Psychological Association, Anaheim, California, August.

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.* Baltimore: The Johns Hopkins University Press.

Schmidt, F.L., Pearlman, K. & Hunter, J.E. 1980. The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology*, 33, 705-724.

Schultz, L.J. & Fortune, J.C. 1981. The three I's: Sources of test bias. *Education*, 102, 117-123.

Shepard, L.A. 1981. Identifying bias in test items. *New Directions for Testing and Measurement*, 11, 79-104.

Shepard, L.A. 1982. Definitions of bias. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias*. Baltimore: The Johns Hopkins University Press.

Shepard, L.A., Camilli, G. & Averill, M. 1981. Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, 6, 317-375.

Shepard, L.A., Camilli, G. & Williams, D.M. 1983. *Accounting for statistical artifacts in item bias research*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, April.

Sternberg, R.J. & Gardner, M.K. 1983. Unities in inductive reasoning. *Journal of Experimental Psychology: General*, 112, 80-116.

Stodolsky, S.S. & Lesser, G. 1967. Learning patterns in the disadvantaged. *Harvard Educational Review*, 37, 546-593.

Stricker, L.J. 1982. Identifying test items that perform differentially in population subgroups: A partial correlation index. *Applied Psychological Measurement*, 6, 261-273.

Subkoviak, M.J., Mack, J.S., Ironson, G.H. & Craig, R.D. 1984. Empirical comparison of selected item bias detection procedures with bias manipulation. *Journal of Educational Measurement*, 21, 49-58.

Sundberg, N.D. & Gonzales, L.R. 1981. Cross-cultural and cross-ethnic assessment: Overview and issues. In P. McReynolds (Ed.), *Advances in psychological assessment*. Vol. 5. San Francisco: Jossey-Bass.

Sung, Y.H. & Dawis, R.V. 1981. Level and factor structure differences in selected abilities across race and sex goups. *Journal of Applied Psychology*, 66, 613-624.

Tatsuoka, K.K. & Linn, R.L. 1983. Indices for determining unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement*, 7, 81-96.

Taylor, T.R. 1987. *Test bias: The roles and responsibilities of test user and test publisher*. HSRC Report Pers-424. Pretoria: Human Sciences Research Council.

Templer, A.J. 1972. The relationship between field dependence-independence and concept attainment. *Psychologia Africana*, 14, 121-129.

Thissen, D.M. 1976. Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement*, 13, 201-214.

Thorndike, R.L. 1982. *Applied psychometrics*. Boston: Houghton Mifflin.

Thorndike, R.L. & Hagen, E. 1969. *Measurement and evaluation in psychology and education*. New York: John Wiley.

Tittle, C.K. 1982. Use of judgmental methods in item bias studies: In R.A. Berk (Ed.), *Handbook of methods for detecting test bias*. Baltimore: The Johns Hopkins University Press.

Traub, R.E. & Lam, Y.R. 1985. Latent structure and item sampling models for testing. *Annual Review of Psychology*, 36, 19-48.

Triandis, H.C. & Brislin, R.W. 1984. Cross-cultural psychology. *American Psychologist*, 39, 1006-1016.

Valencia, R.R. & Rankin, R.C. 1985. Evidence of content bias on the McCarthy Scales with Mexican American children: Implications for test translation and nonbiased assessment. *Journal of Educational Psychology*, 77, 197-207.

Van der Flier, H. 1980. *Vergelijkbaarheid van individuele testprestaties.* Lisse: Swets & Zeitlinger.

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 Gruijter (Eds.), *Psychometrics for educational debates.* Chichester: John Wiley.

Vernon, P.E. 1968. What is potential ability? *Bulletin of the British Psychological Society*, 21, 211-219.

Vernon, P.E. 1969. *Intelligence and cultural environment.* London: Methuen.

Walberg, H.J. & Haertel, G. 1984. Community influences on learning. *Evaluation in Education: An International Review Series*, 8, 1-82.

Wigdor, A.K: & Garner, W.R. 1982. *Ability testing: Uses, consequences and controversies.* Part I and II. Documentation Section. Washington D.C.: National Academy Press.

Wilcox, R.R. 1984/85. A note on measuring item bias. *The Journal of Experimental Education*, 53, 114-116.

Williamson, M.L. & Hopkins, K.D. 1967. The use of "none of these" versus homogeneous alternatives on multiple-choice tests: Experimental reliability and validity comparisons. *Journal of Educational Measurement*, 4, 53-58.

Witelson, S.F. 1977. Developmental dyslexia: Two right hemispheres and none left. *Science*, 195, 309-311.

Witkin, H.A., Goodenough, D.R. & Karp, S.A. 1967. Stability of cognitive style from childhood to young adulthood. *Journal of Personality and Social Psychology*, 7, 291-300.

Witkin, H.A., Moore, C.A., Goodenough, D.R. & Cox, P.W. 1977. Field-dependent and field-independent cognitive styles and their educational implications. *Review of Educational Research*, 47, 1-64.

Wolf, B. 1978. *Bias in testing.* Paper presented at the 19th International Congress of Applied Psychology.

Wood, R.L., Wingersky, M.S. & Lord, F.M. 1976. *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters.* Research Memorandum RM-76-6 (modified 1/78). Princeton, N.J.: Educational Testing Service.