*Journal of Industrial Psychology, 1998, 24(2), 25-33*
*Tydskrif vir Bedryfsielkunde, 1998, 24(2), 25-33*

# DIFFERENTIAL ITEM FUNCTIONING IN THE FIGURE CLASSIFICATION TEST

E van Zyl
D Visser

*Department of Human Resource Management*
*Program in Industrial Psychology*
*Rand Afrikaans University*

## ABSTRACT

The elimination of unfair discrimination and cultural bias of any kind, is a contentious workplace issue in contemporary South Africa. To ensure fairness in testing, psychometric instruments are subjected to empirical investigations for the detection of possible bias that could lead to selection decisions constituting unfair discrimination. This study was conducted to explore the possible existence of differential item functioning (DIF), or potential bias, in the Figure Classification Test (A121) by means of the Mantel-Haenszel chi-square technique. The sample consisted of 498 men at a production company in the Western Cape. Although statistical analysis revealed significant differences between the mean test scores of three racial groups on the test, very few items were identified as having statistically significant DIF. The possibility is discussed that, despite the presence of some DIF, the differences between the means may not be due to the measuring instrument itself being biased, but rather to extraneous sources of variation, such as the unequal education and socio-economic backgrounds of the racial groups. It was concluded that there is very little evidence of item bias in the test.

## OPSOMMING

Die uitskakeling van onregverdige diskriminasie en kultuursydigheid van enige aard, is tans 'n omstrede kwessie in die werkplek in Suid-Afrika. Ten einde regverdigheid in toetsing te verseker, word psigometriese toetse onderwerp aan empiriese ondersoeke na die moontlikheid van sydigheid wat kan lei tot keuringsbesluite wat onregverdige diskriminasie meebring. Hierdie ondersoek is onderneem om die moontlikheid van differensiële itemfunksionering (DIF), of potensiële sydigheid, in die Figuurindelingtoets (A121), met behulp van die Mantel-Haenszel chi-kwadraattegniek, te ondersoek. Die steekproef het bestaan uit 498 mans by 'n produksiemaatskappy in die Wes-Kaap. Alhoewel statistiese ontleding beduidende verskille in gemiddelde toetstellings van drie rassegroepe op die toets aangedui het, is baie min items aangedui wat statisties beduidende DIF bevat. Die moontlikheid word bespreek dat, hoewel sommige DIF in die toets teenwoordig is, die verskille tussen die gemiddeldes nie die gevolg is van 'n sydige meetinstrument *per se* nie, maar eerder die gevolg van eksterne bronne van variasie, soos byvoorbeeld die ongelyke opvoedkundige- en sosio-ekonomiese agtergronde van die rassegroepe. Die gevolgtrekking was dat daar baie min getuienis van itemsydigheid in die toets is.

The notion that psychometric tests can promote unfair discrimination and lead to adverse impact in the workplace has lead to widespread debate and research in the field of testing over the past years (Anastasi & Urbina, 1997; Crocker & Algina, 1986; Gregory, 1996; Hesketh, 1993; Holburn, 1992). The popular criticisms that these instruments, which are based largely on middle-class white values and knowledge, are culturally biased and less valid for other population groups, has lead to the contentious issue of bias in testing.

In simple statistical terms *bias in testing* means systematic errors of measurement. It can be described as a slant in the way a test measures what it is intended to measure, or predict what it is intended to predict (Anastasi & Urbina, 1997; Gregory, 1996; Holland & Thayer, 1988; Jensen, 1984). It is a constant or systematic error that disadvantages the test performance of one group relative to another (Shepard, 1982). 'Bias, or systematic error, is an obtained measurement that consistently over-estimates (or underestimates) the true (error-free) value of the measurement for members of one group as compared with members of another group' (Cleary, 1968, p. 115). Finally, Gregory (1996) stated that bias is present when the meanings or implications of a test score obtained by one subgroup of test takers are different from the meanings or implications that it has for other test takers.

Psychometric tests are generally used in organisations to assist in making decisions about the prediction of future work success of individuals. If, however, it can be demonstrated that these psychometric tests place a particular individual or subgroup at an unfair disadvantage (thus discriminating unfairly), this could constitute an unfair labour practice in terms of the proposed new Employment and Occupational Equity Act in South Africa. The fact that bias, or systematic error, in a test can lead to disadvantaged test performance for members of a certain group, has furthermore lead to the general perception that test bias [always] results in unfair discrimination. The basic concern with fairness in test used in South Africa is thus that when psychometric tests deliver invalid results, previously disadvantaged groups get relatively fewer chances of sharing the social benefits of employment opportunities and occupational advancement, leading to adverse impact in the workplace (Holburn, 1991).

The first move from the new South African government to redress the labour imbalances of our society created by many years of discrimination and segregation is formulated in chapter 3, 'Fundamental Rights', of the Interim Constitution of 1993. Section 27 explicitly provides that 'every person shall have the right to fair labour practices' (Juta's Statutes of South Africa, 1994, p. 270).

Section 8 furthermore provides that 'every person shall have the right to equality before the law and to equal protection by the law' (p. 268). It also states that 'no person shall be unfairly discriminated against, directly or indirectly' and then mentions specific grounds of unfair discrimination: 'race, gender, sex, ethic, or social origin ... culture or language' (Nieuwoudt, 1996).

There is a similar trend in the Green Paper on Employment and Occupational Equity (Government Gazette, 1996). The Green Paper gives a good indication of what may be expected in the final Act. In paragraph 1.5 of the summary the drafters state that 'Employment Equity centres on: Eradication of unfair

*Requests for reprints should be addressed to E van Zyl, Department of Human Resource Management, Rand Afrikaans University, P.O. Box 524, Auckland Park 2006.*

discrimination of any kind ... [and] measures to encourage employers to undertake organisational transformation to remove unjustified barriers ... and to accelerate training and promotion' (p. 6). Finally, and specifically pertaining to recruitment and selection, the Green Paper advises that employers should 'avoid psychometric tests unless they can demonstrate that they respect diversity' (4.5.3.5). In its reply, the Society for Industrial Psychology has since suggested that this statement should be replaced by the statement: 'Psychometric testing should satisfy the criteria for fairness in a diverse society' (Roodt, 1996).

It is against this background of imminent changes in the political-legal sphere of employment policy in South Africa that psychometric tests as selection instruments are under scrutiny of many. Soon employers will have a legal obligation to prove the relevance, validity, and the fairness of psychometric instruments in use so that they are not seen as placing barriers to job advancement for certain groups. The Employment and Occupational Equity Act will demand fair discrimination based on sound selection practices that will promote equal opportunities created by among others [culturally] unbiased and valid test instruments.

Differences in test performance between different groups based on race, sex, age, and socio-economic background has been a constant problem in the field of psychometrics. In South Africa in particular, group score differences among various racial groups have been a prominent problem that was historically handled by designing and standardising separate tests for separate groups. However, the new concern for fair discrimination and equal opportunities creates the need to be able to make valid comparisons between people from various race groups based on results from tests that can prove that they respect diversity. Existing studies that provide empirical guidelines on the equivalence of test results when individuals from different race groups are competing for the same position, are few (Taylor & Radford, 1986). Although research on the topics of test and item bias is not an entirely new field of study in South Africa, the process of gathering data pertaining to our own unique situation, has only recently gathered momentum.

The relative lack of research information seems to echo the historic perception of non-importance of this issue. Werbeloff and Taylor (1982), Vorster (1983), Taylor and Radford (1986), Owen (1986, 1989), and Holburn (1990, 1992) have investigated this issue in South Africa over the past 15 years. Some of the abovementioned authors have concentrated their studies specifically on the investigation of bias in various tests (e.g. Senior Aptitude, Mechanical Insight, Mental Alertness, and Junior Aptitude tests) during the construction phase, while others discuss the possible causes of item bias, the need for unbiased, culture fair instruments, and the problems that are faced regarding differential performance on tests.

The validity of tests that are used within organisations and institutions should be investigated by the test users, and test bias studies should be conducted to shed light on issues that pertain to bias and fairness in testing (Holburn, 1990). Charges of bias are expected from those believing that a test is more appropriate for the group heavily represented in the standardisation sample (in South Africa that is mainly whites) and hence discriminates against those who are not, or that test scores are affected by sources of variation other than the construct the test sets out to measure (e.g. educational background). It is therefore propagated that it would be wise to answer the question of bias virtually on a test-by-test basis (Reynolds & Brown, 1984). Gregory (1996) stated in this regard that the validation of tests is a developmental process that begins with test construction and continues indefinitely.

There is a difference between the two concepts *bias in testing* and *fairness in testing* which needs early clarification in order to guard against common misconception and confusion. Jensen (1980) was very specific in his differentiation between these two separate, but interrelated terms: He said that bias is a general term which is not only limited to culture bias. He furthermore stated that bias is a purely statistical term which means that the expected error of estimate of the true value of the measurement is not zero, and that the assessment of bias is an objective, empirical, statistical, and quantitative matter. According to Flaugher (1978) the definition of test bias has many aspects, all forming part of the one real issue, namely bias in testing. Test bias has been discussed as mean score differences between different groups, as sexism, as single-group or differential validity, as content, as the selection model, and as atmosphere. He warned against losing perspective of all of these interrelated aspects and to guard against over-simplifying the issues in research. Questions about the particular content of tests is one aspect of the overall test bias issue which is addressed in this research. According to Cleary and Hilton (1968) this aspect of the test bias issue refers to the study of individual test items. A biased test from this perspective is one that contains questions that in some sense are 'unfair' to some subgroup of the population (Flaugher, 1978). A biased item produce an uncommon discrepancy between the performance of members of the one group and members of another group.

Fairness, on the other hand, refers to the ways in which test scores (whether of biased or unbiased tests) are used in any selection situation to endorse selection fairness. Fairness in this sense has to do with the characteristics of the decision rule or the selection model adopted in choosing one candidate above another (Crocker & Algina, 1986; Gregory, 1996; Huysamen, 1996; Van Wyk, 1993). It reflects subjective social values and philosophies of test use, envelopes the social consequences of using a test, and reflects the perception of what a just society is (Gregory, 1996). According to Flaugher (1978) all of these models endorse the application of double standards, i.e. two candidates for selection who acquire identical scores on the prediction measures, will be treated differently depending on other criteria such as ethnic identity. In terms of certain criteria of fairness, 'unbiased tests can be used unfairly and biased tests can be used fairly' (Jensen, 1980, p. 375).

The criteria of bias are generally classified under three main headings: content validity, construct validity and predictive validity (Crocker & Algina, 1986; Gregory, 1996; Reynolds, 1982; Reynolds & Brown, 1984). The validity of a test sets out to answer the question: Does the test fulfil its primary function? Herein lies the source of bias: A test that is systematically biased to a subgroup because of a source of variance in the test that functions differently for one subgroup than for another, implies that the test is not measuring what it intended to measure, and will therefore not facilitate appropriate inferences and decisions about each subgroup's performance based on the results of this test. It will be [systematically] invalid for a certain subgroup or subgroups. Bias is therefore directly related to validity.

Bias in the construct validity of a test exists when a test is shown to measure different hypothetical traits (psychological constructs) in one group than in another, or when it measures the same trait, but with differing degrees of accuracy (Reynolds, 1982). For a test to be unbiased in construct, all the items comprising the test must measure the same trait or ability for all the subgroups. Comparisons across relevant subgroups should reveal inter alia (1) a high degree of factorial invariance, and (2) the rank order of item difficulties within the test should correspond (Crocker & Algina, 1986; Gregory, 1996). The demonstration of construct validity cannot be simplified and should always rest upon using diverse procedures and a variety of evidence from numerous sources (Gregory, 1996).

Predictive validity refers to the correlation between the test scores and criteria variables external to the test. An unbiased test should show correlation with other variables in two or more populations i.e. it will predict performance equally well for people from various subgroups (Gregory, 1996; Jensen, 1980; Petersen, 1980). Gregory (1996) explains that with an

unbiased test the test scores of all the relevant subgroups will cluster equally well around a single regression line. Bias in predictive validity can be present as either intercept bias (separate parallel regression lines) or slope bias (separate regression lines for the subgroups that are not parallel) or both. Intercept bias (separate, parallel regression lines) is, however, not necessarily an indication of a biased test (Gregory, 1996; Crocker & Algina, 1986; Flaugher, 1978). What it does imply is that, for the test to have predictive validity for the total group, separate regression equations will be needed to make accurate predictions about performances of members from different subgroups. The use of a single regression line for the different subgroups, in this instance, would lead to under-prediction of performance for members of one group and over-prediction of performance for members of the other, thus constituting a clear instance of test bias. The fine line between test bias and fairness in testing is crossed here and the determining factor will be the selection model. The use of separate regression lines will certainly avoid the problem of consistent non-zero errors, and therefore the accuracy of the prediction should be improved in terms of this aspect on non-zero errors.

When slope bias exists, the regression lines for the different groups are separate and non parallel. The use of a single regression line in such an instance will result in both under- and over-prediction of scores for certain subgroups in both groups. In this instance the test possesses a high degree of bias in its predictive validity (Gregory, 1996).

Reynolds (1982, p. 188) and Gregory (1996, p. 264) described bias to exist in the *content* of a test when an item (or items) 'can be demonstrated to be relatively more difficult for members of one group than for another when the general ability level of the groups being compared is held constant and no reasonable theoretical rationale exists to explain group differences on the item in question'. A biased test, if understood to be a test that is systematically disadvantaging some group, also infers that the test must measure different things for different groups. The possibility that a certain subgroup's performance on an item may be influenced by sources of variation other than differences on the particular construct being measured and the fact that these 'other' sources of variation may influence performance in a way that differs systematically for one subgroup than for another, has lead to item bias studies particularly concerned with the relative difficulty of individual test items (Anastasi & Urbina, 1997; Gregory, 1996). Bias in item/test content is very often stated as the criticism against the use of tests across groups, and critics rely on subjective, 'expert' judgements of items to substantiate this opinion. Previous research, however, could not prove conclusively that expert-nominated items are necessarily culturally biased; and found that 'expert' judges did not succeed in identifying culturally biased test items based on the analysis of the item characteristics (Anastasi & Urbina, 1997; Gregory, 1996). The investigation of item difficulties, however, offers an empirical approach to address the question of bias in test items.

In modern psychometric theory this area of scrutiny of items has become known as the investigation of *differential item functioning (DIF)* (Anastasi & Urbina, 1997; Hambleton, Clauser, Mazor & Jones, 1993; Holland & Thayer, 1988; Shepard, 1982). DIF replaces the term 'item bias' because of the semantic conflict that the word *bias* has created and the fact the DIF focuses on the results of an analytical procedure rather than making inferences about the effect, as is the case with the term bias.

Hambleton et al. (1993, p. 5) and Rudas and Zwick (1997) stated a widely accepted definition for DIF as: 'An item is DIF (or potentially biased) if examinees of equal ability, but from different subgroups (for example males and females) do not have an equal probability of correctly responding to that item'. Dorans and Holland (1993) stated that DIF is an unexpected difference between groups of examinees which should be comparable on the attribute measured by the item and test on which it appears.

DIF studies are therefore conducted to identify those items for which equally able persons from different cultural groups have different probabilities of answering a test item correctly (Anastasi & Urbina, 1997; Rudas & Zwick, 1997). If different groups of comparable examinees react differently to the same question, the possibility exists that performance on that item is influenced by other sources of variation in the various subgroups. These other sources of variation affecting the responses could account to possible bias in the test. The aim of these studies is essentially to identify those items that are (potentially) biased against certain subgroup(s). These items should then be evaluated before a decision is made to modify or remove them from the test in order for the test to become less disadvantageous to the relevant subgroups. In DIF studies the primary interest is in identifying DIF items, because of the potentially negative impact that biased items have on test validity (Hambleton et al., 1993).

Therefore, DIF studies focus on the internal behaviour of a test by empirically investigating the relationship between an item and the total test. The statistical methods used to identify DIF are techniques to detect items that are anomalous, that is, they detect the items that behave differently for different subgroups (Shepard, 1982). Shepard mentioned that an inherent caveat of all item bias methods is that they cannot detect pervasive bias because of lack of an external criterion. Pervasive bias refers to the circularity inherent in using the total score of the test or average item to act as criterion to identify individuals of equal ability and therefore representing the standard of unbiasedness. This means that negligible bias in the test will be absorbed into the total score and as such will not be detected by empirical, internal methods. According to Shepard, the major weakness of all bias detection methods is that they lack an independent external criterion to validate findings. Hambelton et al. (1993, p. 8) consequently reported on the research results of their associates who found that internal and external criterion measures produces 'highly similar (MH DIF) results'.

To find DIF in an item(s), however, may, or may not, mean that the test as a whole is biased and thus unfair to some subgroup. A test showing little DIF is not necessarily unbiased, and similarly, an unbiased test may or may not exhibit differential item performance. For DIF to be linked to test bias *per se*, it must be linked by an external criterion and/or expert judgement (Anastasi & Urbina, 1997; Angoff, 1993; Hambleton et al., 1993; Reynolds & Brown, 1984).

Once DIF items have been identified by whatever empirical method, it is suggested that a next natural step should be to inquire about the nature and source of the difference(s). Although a single best method for investigating this is not proposed, various authors suggest the use of a combination of methods that include some statistical and some judgmental procedures (Anastasi & Urbina, 1997; Angoff, 1993; Hambleton et al., 1993; Jensen, 1980).

Item response theory (IRT) and chi-square methods are currently the most prevalent empirical approaches employed for DIF studies and include IRT area methods, and techniques such as the item characteristic curve (ICC), transformed item difficulty (TID), delta plots, and the Mantel-Haenszel chi-square technique (Anastasi & Urbina, 1997; Ackerman, 1992; Crocker & Algina, 1986; Hambleton et al., 1993; Hills, 1989; Holland & Thayer, 1988; Raju, Drasgow & Slinde, 1993).

The Mantel-Haenszel technique (MH) proposed by Holland and Thayer (1988) has emerged as one of the preferred methods for detecting DIF. This technique uses a variation of the chi-square method as an index of DIF and has emerged as an application of chi-square with a very powerful test for the null hypothesis. The procedure tests the null hypothesis that there is no difference in the way a particular item is functioning in two groups of interest (e.g. black-white) after the groups have

been matched in ability levels. To test the hypothesis of no DIF in the two groups of interest for the item (i.e. the white group and the black group perform identically across all score groups), the MH approach uses the common odds ratio across all K 2x2 tables as an estimate of DIF. The parameter $\alpha$ is called the common odds ratio, which is the average factor by which the odds that a member of the reference group is correct on an item, exceeds the corresponding odds for a comparable member of the focal group (Holland & Thayer, 1988; Nandakumar, 1993). Underlying the study of DIF with this technique is the notion that only comparable members of two groups are compared. This means that to detect DIF, the two groups (referred to as the focal and reference groups, respectively) are divided into several score groups based on their total test score prior to comparing their performance on a studied item.

The MH technique is a very powerful, practicable and rather inexpensive procedure to use for DIF studies (Hambleton et al., 1993; Holland & Thayer, 1988). It is a noniterative, contingency table method based on the premise of unidimensionality in a test, i.e. item response can be ascribed to a single latent trait and therefore the total test score reflects a person's standing on a single latent trait (Ackerman, 1992; Anastasi & Urbina, 1997; Crocker & Algina, 1986; Holland & Thayer, 1988). The technique furthermore requires a test that is very reliable which is a prerequisite for validity (Dorans & Holland, 1993; Hambleton et al., 1993). Holland and Thayer concluded that the MH technique provides a one-degree-of-freedom chi-square test that allows detailed matching on the relevant criteria; and yields a single measure of the size of the deviation from the null hypothesis exhibited by the studied item. It may be used with samples as small as 100-200 per group, although increased sample sizes result in increased power (Hambleton et al., 1993; Hills, 1989).

In view of the above, the aim of the present study is examine the Figure Classification Test (A121) (FCT) at item level for the detection of possible DIF by means of the empirical DIF method, the Mantel-Haenszel technique. The FCT is a South African, nonverbal test of abstract reasoning ability based on figural stimuli which attempts to minimise cultural/language bias (Werbeloff & Taylor, 1982). It is a very popular instrument for use across various cultural groups and is widely used in industry for purposes of intellectual screening. It is hypothesised that certain items of the FCT will be identified as DIF when the test is applied to different racial groups as measured by the MH technique. Cross-group comparisons of relative item difficulties will also be made to investigate a possible link between item difficulty and DIF.

The company at which the study was conducted entered the competitive international market in the early 1990's and a subsequent demand for better quality products moved its management to rethink their business strategies. Amongst others they envisaged their personnel to be multi-skilled and decided to focus heavily on training and development of the potential of their workforce in order to realise this. Simultaneously they wanted to assure that new people entering their organisation had the potential to be trained and developed in their specific technologies.

Historically, very few of the semi-skilled jobs in this organisation (mainly machine operators) were filled by black or coloured people. They were mainly occupied by white males. The new policy, however, demanded that this situation be addressed and rectified. Subsequently it was decided to use the FCT (because of its inherent 'culture fair' properties) as a measuring instrument that will be administered to all current, as well as prospective, employees in order to provide information on examinees' abstract reasoning ability.

## METHOD

### Subjects

The subjects were 498 males between the ages of 18 and 64. All the subjects had between seven and nine years of schooling. The sample included all the present employees of a production company in Cape Town, whereas the rest of the subjects included in the sample were tested as part of the company's recruitment process. Of these subjects, 100 (20%) were white, 253 (51%) were coloured, and 145 (29%) were black.

### Measuring instrument

The FCT, after Taylor (1976), was administered to all subjects. The FCT is a test of abstract reasoning ability which was developed in 1974. It was developed with the specific objective of overcoming problems being experienced with 'verbal' tests (i.e. Mental Alertness) that were administered to different racial groups and which could lead to accusations of cultural bias, unfairness and poor predictability (Werbeloff & Taylor, 1982). In this test, abstract reasoning is used as a measure of Spearman's g factor. Gregory (1996, p. 235) reminds that Spearman referred to g as the 'eduction of correlates' and that the term 'eduction' in this context refers to a 'process of figuring out relationships based on perceived fundamental similarities between stimuli'.

The FCT is a 36-item test designed for people with a schooling range of seven to nine years. Each test item consists of six drawings that must be classified into two groups of three. The criterion is the presence of an underlying concept or similarity which is shared by all drawings that are organised in the same class. These concepts or similarities can inter alia be categorised as uniformity, symmetry, inversion, repetition and series. The six diagrams of each item are labelled by the letters A to F and the examinee indicates his/her classification by recording the three identifying letters of each group on the answer sheet (Taylor, 1976). Initial studies showed that the FCT has a reliability coefficient of 0.90 (Werbeloff & Taylor, 1982).

### Procedure

The test was administered to examinees over a period of three years at the company's training centre by a registered A test user under normal, standardised conditions. All existing staff as well as any new applicants were tested. The test was scored by the administrator and stored in the company's training records. Instructions in the Test Administrators' Manual for Figure Classification Test (Taylor, 1976) were meticulously followed.

## RESULTS

Summary statistics (i.e. age, educational level in school standard) for the various groups contained in the sample, were computed to reflect important characteristics of the compared groups. Total test score statistics for the three racial groups were also computed to determine whether there were significant differences between the performance of the various groups on the total test. A computation of item difficulty index (p) was carried out to establish whether difficulty values per studied item differed among the various racial groups and to study the order of the difficulty values. The summary statistics for the black, coloured and white groups are shown in Table 1. The mean age for the black group was somewhat lower than the mean ages of the coloured and white groups. The mean educational level, in school standard, of the black group was similar to that of the coloured group, but slightly lower than

TABLE 1
SUMMARY STATISTICS FOR THE BLACK, COLOURED AND WHITE GROUPS

| GROUP | N | AGE M(SD) | EDUCATION M(SD) |
|---|---|---|---|
| Black | 145 | 33,50 (9,08) | 6,79 (1,49) |
| Coloured | 253 | 34,23 (9,78) | 6,65 (1,52) |
| White | 100 | 34,66 (11,59) | 7,64 (0,68) |

TABLE 2
ITEM DIFFICULTY INDICES PER RACIAL GROUP PER STUDIED ITEM

| Item | p (total) | p (black) | p (coloured) | p (white) |
|---|---|---|---|---|
| 1 | 0,88 | 0,80 | 0,88 | 0,98 |
| 2 | 0,82 | 0,70 | 0,84 | 0,95 |
| 3 | 0,80 | 0,61 | 0,85 | 0,93 |
| 4 | 0,79 | 0,66 | 0,82 | 0,88 |
| 5 | 0,65 | 0,41 | 0,72 | 0,81 |
| 6 | 0,70 | 0,63 | 0,70 | 0,79 |
| 7 | 0,64 | 0,40 | 0,70 | 0,87 |
| 8 | 0,66 | 0,46 | 0,72 | 0,82 |
| 9 | 0,37 | 0,21 | 0,40 | 0,55 |
| 10 | 0,73 | 0,57 | 0,75 | 0,90 |
| 11 | 0,69 | 0,38 | 0,79 | 0,91 |
| 12 | 0,65 | 0,49 | 0,68 | 0,81 |
| 13 | 0,70 | 0,51 | 0,75 | 0,84 |
| 14 | 0,43 | 0,27 | 0,47 | 0,56 |
| 15 | 0,56 | 0,32 | 0,60 | 0,80 |
| 16 | 0,71 | 0,48 | 0,77 | 0,89 |
| 17 | 0,62 | 0,39 | 0,66 | 0,84 |
| 18 | 0,57 | 0,37 | 0,63 | 0,70 |
| 19 | 0,52 | 0,41 | 0,52 | 0,67 |
| 20 | 0,49 | 0,32 | 0,50 | 0,72 |
| 21 | 0,22 | 0,12 | 0,24 | 0,29 |
| 22 | 0,46 | 0,23 | 0,50 | 0,71 |
| 23 | 0,29 | 0,17 | 0,29 | 0,45 |
| 24 | 0,51 | 0,28 | 0,58 | 0,68 |
| 25 | 0,51 | 0,41 | 0,53 | 0,60 |
| 26 | 0,44 | 0,27 | 0,47 | 0,61 |
| 27 | 0,59 | 0,35 | 0,62 | 0,84 |
| 28 | 0,35 | 0,17 | 0,39 | 0,54 |
| 29 | 0,26 | 0,16 | 0,28 | 0,37 |
| 30 | 0,48 | 0,37 | 0,49 | 0,60 |
| 31 | 0,35 | 0,16 | 0,37 | 0,56 |
| 32 | 0,55 | 0,23 | 0,63 | 0,83 |
| 33 | 0,29 | 0,17 | 0,30 | 0,41 |
| 34 | 0,54 | 0,26 | 0,62 | 0,76 |
| 35 | 0,40 | 0,19 | 0,45 | 0,55 |
| 36 | 0,35 | 0,18 | 0,39 | 0,50 |
| Total | 19,57 | 13,11 | 20,90 | 25,52 |

that of the white group. However, these differences are negligible, because the matching criterion for a DIF analysis is total test score, which means that only comparable member per score group were compared.

The mean total score for the white group (M = 25.53; SD = 7.34) was 4.53 points higher than the mean for the coloured group (M = 21.00; SD = 9.09), and 12.34 points higher than the mean for the black group (M = 13.19; SD = 7.99). Statistically significant differences between total test score means were obtained for the three racial groups following a oneway ANOVA, F(2,495) = 69.83; p < 0.0001. Effect sizes were computed using Cohen's estimated f (Cohen, 1988) and the eta statistic (Rosenthal & Rosnow, 1991). The ANOVA yielded a large effect size, because values of 0.74 and 0.47 respectively were obtained. According to Cohen's conventions for f, a value of 0.4 constitutes a large effect size. Post hoc comparisons using Duncan's Multiple Range Test were performed and statistically significant differences were obtained between all three means at the 0.01 level. The white group therefore performed significantly better than both the coloured and black groups on the FCT.

Table 2 provides the item difficulty values per studied item per racial group. Item difficulty per single item (pi) is reflected as the proportion of examinees per racial group that answered the item correctly. In line with the mean total score differences, the p value differences between the racial groups put the coloured groups' performance intermediate to that of the white and black groups. (Note that omits were taken as incorrect and that difficulty values should therefore be interpreted with caution in view of the information in Figure 1.)

A separate analysis was done where omitted items were treated as a separate variable ('omitted', not as correct or incorrect) to provide information on a possible pattern with regard to omits in relation to racial group behaviour on the test. In Figure 1 the number of omitted cases per item per racial group are reflected. Figure 1 indicates that the black group progressively omitted more items toward the end of the test than both the white and the coloured groups. The coloured group omitted progressively more items than the white group toward the end of the test. It seems that omitted items may be related to racial group performance on the FCT, as reflected in this study.

The procedure that was recommended by Nandakumar (1993) when there are not enough examinees representing each possible total test score for a DIF analysis, was followed in this study. Examinees were grouped into five score groups according to their total test scores prior to analysis. These groups represented test scores of 0-7, 8-14, 15-21, 22-28, and 29-36. The data were then analysed using a program for detecting DIF by means of the Mantel-Haenszel statistic, that was originally developed by Nandakumar (1993) and subsequently adapted at the Human Sciences Research Council (HSRC).

The Mantel-Haenszel chi-square (MH-CHISQ) results for the white-black, coloured-black and white-coloured comparisons were extracted. The two groups of interest (e.g. white-black) were referred to as the focal (F) group, which is of primary interest, and the reference (R) group, which is taken as the standard against which the performance of the focal group was compared. The corresponding estimate of the MH alpha ($\alpha$) which is the common odds ratio across the 2x2 tables as an estimate of DIF, was also calculated. The value of MH alpha is

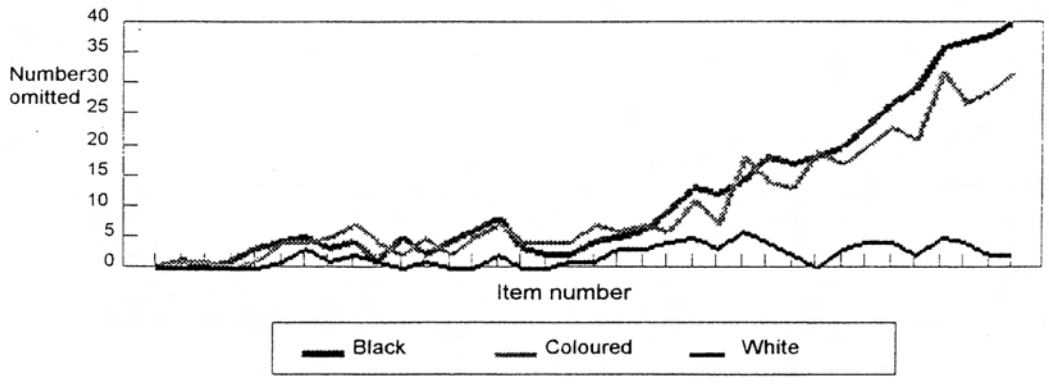


Figure 1: Omitted cases per item per racial group

the average factor by which the odds that a member of the reference group is correct on the studied item, exceeds the corresponding odds for a comparable member of the focal group (Holland & Thayer, 1988). The range of alpha varies between 0 and ∞, with $\alpha = 1$ playing the role of a null value of no DIF. Values exceeding 1 correspond to items on which the focal group performed better on average than comparable members of the reference group. Values smaller than 1 indicate the studied item is favouring the reference group (Nandakumar, 1993).

Holland and Thayer (1988) suggested a conversion for $\alpha$, referred to as delta DIF, which takes the logarithm of $\alpha$ to convert it into a symmetrical scale in which zero is the null value. A zero value of delta denotes no DIF. A significant item with a negative value of delta implies that the item favours the reference group, whereas a significant item with a positive value of delta implies that the item favours the focal group.

The MH chi-square statistics for the white (R) vs. black (F), white (R) vs. coloured (F) and coloured (R) vs. black (F) comparisons are reflected in Table 3. The table shows, for each studied item, the Mantel-Haenszel chi-square statistic (MH-CHISQ), the MH alpha and the delta DIF. Items with significant DIF at statistical significance levels of $p<0.05$ and/or $p<0,01$ were identified by one or two asterisks respectively.

In the MH comparison between the white and black groups, four items were flagged as significantly DIF (Items 11, 22, 32, 36). All four items had significant chi-squares at the 0.05 level, and items 22 and 32 were significant at the 0.01 level as well. Items 11, 22 and 32 favoured the reference group which, in this instance, was the white group. Item 36, however, favoured the black group. The items that were identified as DIF has some of the highest differences in p values e.g. item 32 showed a difference of 0.60 in favour of the white group. It should be kept in mind that Blacks omitted more items towards the end of the test. This figure is therefore relative.

The results of the DIF analysis between the white-coloured comparison showed that no items were flagged as significantly DIF.

The MH comparison between the black and coloured groups yielded more DIF items than the white-black comparison. Seven items were flagged as significantly DIF (i.e. items 6, 11, 19, 22, 30, 32, 34). Items 11, 22, 32 and 34 favoured the reference group, which was the coloured group in this comparison, and items 6, 19 and 30 favoured the focal group (black). Items 11, 19, 30 and 32 were significant at both the 0.05 and 0.01 levels. The items with the greatest MH-CHISQ value (item 11) showed a p value difference of 0.41 between the two groups, in favour of the coloured group. It appeared that the items of the FCT sometimes favoured the performance of the coloured group and sometimes that of the black group.

TABLE 3
MANTEL-HAENSZEL CHI-SQUARE STATISTICS

| | White-black | | | White-coloured | | | Coloured-black | | |
|---|---|---|---|---|---|---|---|---|---|
| Item | CHISQ | Alpha | Delta | CHISQ | Alpha | Delta | CHISQ | Alpha | Delta |
| 1 | 0,3 | 2,0 | -1,6 | 2,4 | 4,1 | -3,3 | 2,2 | 0,6 | 1,4 |
| 2 | 0,1 | 1,5 | -0,9 | 0,7 | 2,0 | -1,6 | 0,8 | 0,7 | 0,8 |
| 3 | 0,0 | 1,3 | -0,6 | 0,0 | 1,1 | -0,2 | 0,7 | 1,4 | -0,7 |
| 4 | 0,3 | 0,7 | 0,8 | 0,3 | 0,7 | 0,8 | 0,4 | 0,8 | 0,6 |
| 5 | 0,2 | 1,3 | -0,6 | 0,1 | 0,8 | 0,4 | 3,6 | 1,7 | -1,3 |
| 6 | 3,8 | 0,4 | 2,2 | 0,1 | 0,9 | 0,3 | 6,2 | 0,5 | 1,8* |
| 7 | 3,5 | 2,6 | -2,2 | 0,3 | 1,4 | -0,8 | 0,3 | 1,2 | -0,4 |
| 8 | 0,0 | 0,9 | 0,3 | 0,0 | 1,0 | 0,1 | 0,1 | 1,1 | -0,3 |
| 9 | 0,1 | 1,2 | -0,4 | 0,0 | 1,1 | -0,2 | 0,0 | 0,9 | 0,2 |
| 10 | 0,0 | 1,0 | 0,0 | 1,2 | 1,7 | -1,2 | 1,1 | 0,7 | 0,9 |
| 11 | 4.6 | 3,7 | -3,1* | 0,0 | 1,2 | -0,4 | 10,3 | 3,0 | -2,6 |
| 12 | 0,0 | 0,8 | 0,4 | 0,0 | 0,9 | 0,1 | 0,2 | 0,9 | 0,4 |
| 13 | 0,8 | 1,5 | -1,0 | 0,1 | 1,2 | -0,4 | 2,1 | 1,5 | -1,0 |
| 14 | 0,8 | 0,6 | 1,1 | 0,6 | 0,8 | 0,6 | 0,3 | 0,8 | 0,5 |
| 15 | 0,3 | 1,4 | -0,8 | 0,6 | 1,4 | -0,8 | 0,0 | 0,9 | 0,2 |
| 16 | 1,4 | 2,0 | -1,6 | 0,2 | 1,3 | -0,6 | 1,1 | 1,4 | -0,8 |
| 17 | 1,2 | 1,7 | -1,3 | 1,1 | 1,6 | -1,0 | 0,0 | 1,1 | -0,2 |
| 18 | 0,6 | 0,7 | 0,9 | 1,4 | 0,7 | 1,0 | 0,0 | 1,1 | -0,2 |
| 19 | 2,8 | 0,4 | 1,9 | 0,0 | 0,9 | 0,1 | 8,8 | 0,4 | 2,0** |
| 20 | 2,3 | 1,9 | -1,5 | 2,4 | 1,6 | -1,1 | 0,1 | 0,9 | 0,3 |
| 21 | 0,0 | 1,1 | -0,3 | 0,2 | 0,8 | 0,4 | 0,1 | 1,2 | -0,4 |
| 22 | 8,7 | 3,1 | -2,7** | 3,0 | 1,7 | -1,3 | 4,1 | 1,8 | -1,3 |
| 23 | 0,4 | 1,4 | -0,8 | 0,8 | 1,3 | -0,6 | 0,2 | 0,8 | 0,4 |
| 24 | 0,1 | 0,8 | 0,5 | 0,6 | 0,7 | 0,7 | 0,6 | 1,3 | -0,7 |
| 25 | 2,0 | 0,6 | 1,4 | 1,3 | 0,7 | 0,9 | 3,7 | 0,6 | 1,2 |
| 26 | 0,3 | 1,3 | -0,6 | 0,0 | 1,0 | 0,1 | 0,0 | 1,0 | 0,0 |
| 27 | 0,1 | 1,3 | -0,6 | 2,1 | 1,9 | -1,5 | 0,1 | 0,9 | 0,4 |
| 28 | 1,7 | 1,7 | -1,3 | 0,1 | 1,1 | -0,3 | 0,1 | 1,2 | -0,4 |
| 29 | 0,1 | 0,8 | 0,5 | 0,0 | 1,0 | 0,1 | 0,5 | 0,8 | 0,6 |
| 30 | 2,5 | 0,5 | 1,6 | 0,4 | 0,8 | 0,5 | 6,9 | 0,5 | 1,8** |
| 31 | 0,0 | 0,9 | 0,4 | 0,0 | 1,1 | -0,2 | 0,8 | 0,7 | 1,0 |
| 32 | 11,4 | 4,7 | -3,6** | 2,2 | 1,8 | -1,4 | 8,2 | 2,5 | -2,1** |
| 33 | 0,9 | 0,6 | 1,2 | 0,2 | 0,9 | 0,4 | 2,8 | 0,5 | 1,6 |
| 34 | 2,6 | 2,1 | -1,8 | 0,0 | 1,0 | 0,1 | 4,1 | 1,9 | -1,4* |
| 35 | 0,0 | 0,9 | 0,2 | 1,4 | 0,7 | 0,9 | 0,2 | 1,2 | -0,5 |
| 36 | 4,1 | 0,3 | 3,2* | 1,2 | 0,7 | 0,9 | 0,5 | 0,7 | 0,8 |

Degrees of freedom for MH-CHISQ = 1
* p <0,05 **p <0,01

Item 11, which was flagged DIF in both the white-black and the coloured-black comparisons, appeared to appeal to the coloured and white groups, but not to the black group. Based on inspection of this items, it is speculated that it relies more on analytical interpretation than, for instance, item 6. Item 6, which favoured the black group in the coloured-black comparison, seems to be much less 'structured', and more abstract. Item 32, which was identified DIF for the black group, in both comparisons, seems to contain a 'Western' concept, namely arrows as indicators of direction. This could be the reason for the under-performance of the black group on this item. Similarly, item 30, which was flagged DIF against the coloured group, seems to contain a concept, namely 'sticks' in various formations, that could be more familiar to the black group than the coloured group. More in-depth, judgmental and statistical analysis of the various DIF items will, however, be needed to explore the causes of DIF.

The greatest number of DIF items existed between the coloured-black comparison, and the white-black comparison was intermediate to this, with no DIF in the coloured-white comparison. Furthermore, the difference in item difficulty indices between the black-white comparison was the greatest. The item difficulty indices between the coloured-white comparison yielded the smallest differences, whereas the differences between the black-coloured comparison were intermediate to these. It appeared that there was a positive relationship between significant DIF and the difficulty value index differences on a studied item. These differences should be viewed against information provided in Figure 1, pointing to the fact that it merely reflects a relative position.

## DISCUSSION

The fact that biased test results can lead to unfair discrimination, racial inequity and adverse impact, is one of the most contentious workplace issues in South Africa today. The Labour Relations Act, 1995, as well as the proposed Employment Equity Act, make specific provisions against this type of discrimination. A psychometric test found to contain DIF can indeed, if not proven otherwise, be regarded as a potentially biased instrument of which the overall validity can be jeopardised. The popular perceptions and misconceptions surrounding psychometric testing in the industry can lead to a test losing its credibility, deservedly or not, if the issues of DIF and bias are not approached from a scientific, test-by-test point of view, and reported in a similar fashion.

This study was undertaken to establish whether the FCT contains DIF as tested by the MH technique. The results presented in the previous section indicate a few items with significant DIF, especially in white-black and coloured-black comparisons. It shows that no significant DIF exists in the white-coloured comparison. The comparisons between mean test scores, however, indicate significant differences in overall performance between the various groups. These differences in racial group performance clearly indicate a skewness in favour of, particularly, the white group. The black group consistently performed inferior to both the white and the coloured groups on the FCT. The results furthermore show that differences in item difficulty indices show a relation to DIF in comparisons between the various racial groups and that there are differences between the various groups with regards to overall item difficulty indices. Gregory (1996) referred to the optimum level of item difficulty as ranging between 0.3 and 0.7, with most of the items hovering around 0.5. When observing the item difficulty indices per item for the groups, there seem to be constant differences, with the lowest difficulty indices for the black group, the highest indices for the white group, and intermediate indices for the coloured group. The item difficulty index mean for the coloured group is 0.58, for the white group it is 0.71, and for the black group it is 0.36. It appears that items comprising the FCT are, in general, difficult for the black group, easy for the white group, and acceptable for the coloured group.

There are, however, no empirical data available to explain the cause(s) of differences in average test scores, and therefore difficulty value indices, between the various racial groups. It is only possible to speculate about the observed differences, of which educational- and socio-economic backgrounds might be the main causes. Statistical artifacts such as number of examinees per score group, as well as further investigation into the type of differences on items, may also shed light on observed differences. Results should therefore be viewed with circumspection. The results of this study therefore appear to support the stated hypothesis, namely that, certain items of the FCT are DIF in comparisons between various racial groups, and corresponding item difficulty indices do relate to identified DIF items.

It was shown in Tables 2 and 3 that p value differences between racial groups on a studied item show a positive relation with corresponding identified DIF items. In all the instances the highest p value differences corresponded with the greatest MH chi-square values, which were between the white and black groups. These results support Anastasi and Urbina's (1997) statement that DIF is a [related] measure of relative item difficulty. Item difficulty difference(s) may therefore be a valid indicator and back-up statistic to indicate and verify DIF. More research on the relation between DIF and p values should yield valuable information in this regard.

It is clear that the rank order of difficulty values per racial group are not in a definite order from easy to difficult as should be the case for mental tests (Gregory, 1996). Items seem to appear in random order of difficulty, with no consistency within a group, or between groups.

The results with regard to number of omitted items in Figure 1 show a strong relation between race and number of omitted items. It indicates that blacks omitted more items than both the white group and the coloured group, and that the coloured group left out more than the white group. The findings concerning omitted items need to be explored by further investigating the relationship between omits, the perceived difficulty of an item, and/or the time allowed for completion of the test. The fact that a larger number of items at the end of the test were omitted by the black group in particular, can point to a variety of possible causes. These may be a lack of understanding of instructions, visual discrimination problems, a not good enough conceptualisation of working under time pressure, or too little time allowed.

Several unanswered questions arise from the present study. Possible causes for differences in performance on the test can probably be ascribed to a variety of other factors and should ideally be further researched. The current instructions for the FCT are available in English and Afrikaans only. The possibility exists that black people experience problems to understand the initial test instructions if it is not presented in their mother tongue. The FCT's instructions are rather complicated and prone to misunderstanding. Adjusting the standardised procedure for testing may be necessary to accommodate all. The possibility exists that more practice items and some coaching are needed to prepare the examinee for the test content to familiarise him/her with the process, in the event that the initial, verbal instructions are not clearly comprehended. Another area for consideration is in the cultural behaviour of black people (men in particular). They do not seem to be comfortable with indicating that they do not understand an explanation or an item (in front of other people). Also, the allowed testing time, namely 60 minutes, might need to be revised in order to allow for slower people to be able to finish. A further related question that arises is whether all cultures perceive time pressure in the same manner and as a reality. Minor changes/adaptations to procedures may be necessary to ensure validity of the test. Whether these changes are minor enough not to invalidate the established norms or whether they are so substantial that the existing norms no longer apply, should however be determined. The connection between non-understanding of instructions, omitted items

and non-completion of the test, should also provide valuable information.

The question that arises from the results is whether it may be claimed that the differences in performance on the FCT reflect real differences in ability between white-black and coloured-black comparisons. Alternatively, does the probability of success for equally able people, who are from different racial groups, differ, thus creating the possibility for bias in the instrument? Or are the observed differences due to extraneous variables, for instance educational background? Another question is whether the differences are negligible as a result of sample size and/or other statistical artifacts?

To ascribe the observed DIF to real differences in ability, will not be accurate seeing that the ability level of the groups being compared, were held constant (matched on total test score). The identified DIF items were relatively more difficult for the members of one group than for the other, probably because of some variable not set out to be measured by the test. The probability for success could have been influenced, in these instances, by other source(s). One broad explanation can be that the FCT is sensitive to extraneous sources of variation such as cultural background or education. The fact that DIF was identified between the white-black and coloured-black comparisons, and not between the white-coloured comparison, suggests that the differences may be ascribed to cultural background more than to educational background, although this relationship cannot be inferred from the results of this study alone.

The differences in average test score performance among the three racial groups can lead to problems with predictive validity. Based on mean score differences it is clear that if a common regression line were to be used for prediction, performances can be over-predicted or under-predicted, depending on the situation. The use of a single regression line can therefore result in biased predictions, conforming to the definition of intercept bias as described by Gregory (1996). The FCT thus has the *potential* to emerge as a biased test. The current use of differential norms for various racial groups is probably the best available approach, compromising the possibility for intercept bias, provided it is used within the parameters of a defined fairness model. This approach is probably the best temporary solution for a situation where different racial groups did not receive equal education or do not come from similar cultural backgrounds.

To declare the FCT a biased instrument, based on the relatively small number of flagged DIF items would, however, also be inaccurate. It must be kept in mind that not all deviant items that were flagged as DIF are necessarily biased. The detection of DIF is purely a recognisance exercise. The further investigation of DIF items by means of other statistical methods and judgmental approaches is necessary to determine whether the DIF is indeed indicative of bias. Item characteristic curves (ICC) could provide valuable information for this examination. The identification of uniform and non-uniform DIF will become possible with this analysis, and as such the characteristics of the item should give valuable clues as to the reason(s) for the identified DIF (Hambleton et al., 1993). Further investigation may find the items totally acceptable, or decisions can then be made to change, omit, or adapt the relevant item.

The FCT itself may not be a 'biased' test, but differences in performance of various racial groups can give the impression of biasedness. One may conclude that the FCT, as it stands, can be used effectively within the parameters of, for instance, a quota system. Any form of differentiation and decision-making outside the parameters of a specific fairness model, which reflects a certain socio-political value decision, will be subject to speculation of bias (Huysamen, 1996). Decisions based purely on total test scores, in isolation, should be considered with caution, as it could lead to biased and unfair selection practices. The test's value appears, therefore, like most other mental tests, to lie in its fair application and use, based on an appropriate fairness model.

An issue which could not be addressed in this study is that little attention was given to empirical investigation of other relevant factors pertaining to a test's validity, i.e. construct and predictive validity. Linking DIF to an external criterion would have increased the value of the results significantly. Similarly, there is no single method that can be guaranteed to identify all the DIF items in a test. A well-known shortcoming of the MH technique, namely that it is not useful for detecting non-uniform DIF, should be kept in mind. A variation of the MH (a three-part procedure) can however be run to detect non-uniform DIF where necessary or required (Hambleton et al., 1993). By using multiple methods this instability problem of the MH technique can be addressed, and flagged DIF items can be cross-validated and verified. Judgmental methods should be employed to add value to this process. Furthermore, increased sample size would have increased both the power of the results, as well as the applicability potential of other techniques which need bigger samples. The sample on which this study was based was drawn from one region of the company's business activity and care should therefore be taken with the generalisation of results. A broader study would be needed to draw final conclusions.

## ACKNOWLEDGEMENTS

The assistance of members of the HSRC with the computations is gratefully acknowledged.

## REFERENCES

Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Erlbaum.

Anastasi, A. & Urbina, S. (1997). *Psychological testing* (7th ed.) Upper Saddle River, NJ: Prentice-Hall.

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.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Crocker, L.M. & Algina, J. (1986). *Introduction to classical and modern test theory.* New York, N.Y.: CBS College Publishing.

Dorans, N.J. & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

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

Gregory, R.J. (1996). *Psychological testing history, principles, and applications* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Government Gazette (1996, July). *The employment and occupational equity act: Policy proposals, 17303*(804).

Hambleton, R.K., Clauser, B.E., Mazor, K.M. & Jones, R.W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9* 1-18.

Hesketh, B. (1993). Measurement issues in industrial and organizational psychology. *International Review of Industrial and Organizational Psychology, 8,* 133-163.

Hills, J.R. (1989). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice, 8,* 5-11.

Holburn, P.T. (1990, May). *Addressing the need for technical staff – apprentice selection.* Paper presented at the congress on

Psychometrics for psychologists and personnel practitioners, Johannesburg.

Holburn, P.T. (1991). *Selection decisions: The quest for fairness* (Pers 424). Johannesburg: Human Sciences Research Council.

Holburn, P.T. (1992). *Differential item functioning in the mental alertness test.* Unpublished master's thesis, University of South Africa, Pretoria.

Holland, P.W. & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds). *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Huysamen, G.K. (1996). The socio-political context of the application of fair selection models in the USA. *Journal of Industrial Psychology, 22*(1), 1-6.

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

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

Nandakumar, R. (1993). A Fortran 77 program for detecting differential item functioning through the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 53,* 679-684.

Nieuwoudt, H. (1996, August). *The green paper, labour relations act and equal opportunity equity act: Proposed measures for employment equity referring to fairness in recruitment and selection.* Unpublished paper delivered at Recruitment and Selection. Fairness in a new South Africa seminar, Bellville.

Owen, K. (1986). *Toets- en itemsydigheid: Toepassing van die senior aanlegtoetse, meganiese insigtoets, en skolastiese bekwaamheidsbatterye op blanke, swart, kleurling- en Indiër technikon studente* (P-55) (Test and item bias: Application of the senior aptitude tests, mechanical insight test, and scholastic proficiency battery on white, Black and Indian technikon students). Pretoria: Human Sciences Research Council.

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

Petersen, N.S (1980). Bias in the selection rule – bias in the test. In L.J.Th. van der Kamp, W.F. Langerak & D.N.M. de Gruijter (Eds), *Psychometrics for educational debates* (pp. 103-122). New York: Wiley.

Raju, N.S., Drasgow, F. & Slinde, J.A. (1993). An empirical comparison of the area methods, Lord's chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning. *Educational and Psychological Measurement, 53,* 301-314.

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

Reynolds, C.R. & Brown, R.T. (1984). Bias in testing: Introduction to the issues. In C.R. Reynolds & R.T. Brown (Eds), *Perspectives on bias in mental testing* (pp. 1-40). New York: Plenum.

Roodt, G. (1996, July). *Commentary on Green Paper from Society for Industrial Psychology.* Unpublished letter, Society for Industrial Psychology, Johannesburg.

Rosenthal, R. & Rosnow, R.L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Rudas, T. & Zwick, R. (1997). Estimating the importance of differential item functioning. *Journal of Educational Behavioural Statistics, 22*(1), 31-45.

Shepard, L.A. (1982). Definitions of bias. In R.A. Berk (Ed), *Handbook of methods for detecting test bias.* (pp. 9-30). Baltimore, M.D: Johns Hopkins University Press.

Taylor, T.R. (1976). *Test administrators' manual for Figure Classification Test.* Pretoria: Human Sciences Research Council.

Taylor, M. & Radford, E.J. (1986). Psychometric testing as an unfair labour practice. *South African Journal of Psychology, 16*(3), 79-86.

The interim constitution of South Africa. (1994). *Juta's Statutes of South Africa,* (5), 266-335.

Van Wyk, M.W. (1993). Employment equity, part 1: Fairness in the selection process. *South African Journal of Labour Relations, 17*(4), 3-39.

Vorster, J.F. (Ed.) (1983). *Simposium oor die problematiek wat ontstaan by die gebruik van dieselfde of afsonderlike toetse vir verskillende bevolkingsgroepe – Oktober 1982* (Symposium on problems related to the use of identical or separate tests for different population groups). Pretoria: Human Sciences Research Council.

Werbeloff, M. & Taylor, T.R. (1982). *Development and validation of the High Level Figure Classification Test* (Pers 338). Johannesburg: National Institute for Personnel Research.