Special Report Pers-424 1987

Test Bias: The Roles and Responsibilities of Test User and Test Publisher

Terence R Taylor

Nationale Instituut vir Personeelnavorsing — NIPN
National Institute for Personnel Research — NIPR

RGN·HSRC

Test Bias: The Roles and Responsibilities of Test User and Test Publisher

Special Report Pers-424

Test Bias: The Roles and Responsibilities of Test User and Test Publisher

Terence R Taylor

Pretoria
Human Sciences Research Council
1987

Terence R Taylor, D. Litt. et Phil., Chief Research Specialist

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

ISBN 0 7969 0551 7

Printed and published by HSRC
134 Pretorius Street
Pretoria

ACKNOWLEDGMENTS

Dr G K Nelson, Executive Director of NIPR

Members of NIPR Library staff

A special word of thanks to Mr Charles Chemel. I had a number of very useful discussions with him. These discussions helped me to shape my thinking on the topics covered in this report.

TABLE OF CONTENTS

TABLES AND FIGURES

EKSERP

Baie Suid-Afrikaanse sielkundige toetse is vir 'n enkele populasie of groep ontwikkel en gestandaardiseer. Tot redelik onlangs het segregasie in die werksplek tot gevolg gehad dat baie tipes werk deur slegs een groep gedoen is, en toetsing vir keuring of plasing kon gevolglik gedoen word met behulp van 'n instrument wat vir die betrokke groep ontwerp is.

Hierdie situasie verander egter vinnig. Die werkmag, veral op die middel- en laermiddel-posvlak, word toenemend uit a wye spektrum van groepe getrek. (Met "groep" word nie net slegs ras bedoel nie; ander biografiese verandelikes soos geslag of taal kan ook gebruik word om groepe te definieer.)

As gevolg van hierdie ontwikkelings moet die vergelykbaarheid van toetstellings ten opsigte van verskillende groepe ondersoek word. Vergelykende inligting oor verskillende groepe se resultate ten opsigte van toetse is nodig indien billike besluite by keurings geneem moet word.

In hierdie verslag word die rol en verandwoordelikhede van die toetsgebruiker en toetsuitgewer bespreek met betrekking tot sydigheid in billikheid. Die uitskakeling van itemsydigheid is identifiseer as een van die hoofverantwoordelikhede van die toetsuitgewer. Hierdie onderwerp word breedvoerig in die verslag bespreek in 'n strategie vir die uitskakeling van hierdie tipe sydigheid word voorgestel.

Sydigheidsnavorsing work gewoonlik op a priori-groepe gedoen (met ander woorde groepe wat voor die versameling van inligting bekend is en wat op biografiese verandelikes gebaseer is). In die verslag word voorgestel dat sodanige navorsing ook op sogenaamde a posteriori-groepe gedoen moet word (groepe wat gevorm word op grond van antwoorde op die toets wat ondersoek word). Die voordele van hierdie tipe navorsing word bespreek en 'n navorsingstrategie word voorgestel.

ABSTRACT

Many South African psychological tests were developed and standardized on a single population or group. Until fairly recently, segregation in the work-place meant that many jobs were done by only one group, and testing for selection or placement could be done effectively using an instrument designed for the group in question.

This situation is rapidly changing. The workforce, especially at the middle and lower-middle job levels, is increasingly drawn from a wide spectrum of groups. (By "group" we do not mean only race; other biographical variables such as sex or language can also be used to define groups.)

As a result of these developments, the comparability of test scores across different groups has to be investigated. Comparative information on the scores of different groups on the tests is necessary to make fair selection decisions possible.

In this report the roles and responsibilities of the test user and test publisher are discussed with regard to bias and fairness issues. The removal of item bias is identified as one of the main responsibilities of the test publisher; this topic is covered in some detail in the report and a strategy for removing this kind of bias is proposed.

Bias research is usually done on a priori groups (i.e. groups which are known in advance of the data collection, and which are based on biographical variables). A case is made in the report for also doing bias research on what we call a posteriori groups (groups formed on the basis of responses to the test under investigation). The advantages of doing this kind of research are discussed and a research strategy is proposed.

1.0 INTRODUCTION

Until fairly recently job reservation and other social and political factors in South Africa had the effect of limiting the candidates for most types of job to a single population group. This situation is changing rapidly. Although some jobs at the highest and lowest levels remain the preserve of a single group, many types of work in the middle ranges are now done by members of all population groups. For instance, only a few years ago only whites did artisan jobs; now artisan work is done by members of all population groups.

These changes in employment patterns have important ramifications for selection. Tests often form a vital part of the selection process, but most of the tests used in South Africa for this purpose were developed for only one population group. There was no need to develop tests for more than one population group for reasons mentioned above. If all candidates competing for a given job are from the same group, an appropriate test developed for that group can be safely used to compare candidates. However, if candidates are from different population groups and all available tests have been developed and standardized on only one group, comparison becomes problematic, as it cannot be assumed that scores of members from different groups have the same meaning. The issues associated with this problem have been discussed in the literature under a number of headings, including "bias", "fairness", "comparability", and "culture-loadedness".

Up to this point we have been distinguishing groups in terms of ethnicity. This is not the only type of grouping which is made in bias research. Sex is another discrete variable which has been used to form groups; and continuous variables, such as socio-economic status (SES) can be arbitrarily "cut" at various values to form any number of groups. (One can, for instance, create groups made up of people of high, medium, and low SES.) The bulk of overseas research in this domain has, however, concentrated on ethnically-based comparisons. One of the reasons for this is that race groups have political "clout". Also, court cases can be based on claimed discrimination against a race group. The US constitution protects individuals against discrimination on the basis of race, creed, or sex. Several cases have come to court in the USA on the issue of test bias against a given race group, or unfairness to a certain race group with the regard to the use of a given test.

In South Africa the social structure has been based on race for a very long time and to a large extent remains so. Certain other important variables such as education and SES might correlate with race to a greater extent than they do in the USA. In South Africa race membership therefore probably summarizes a whole collection of factors to a greater extent than it does in most other countries. Thus there is in this country an even greater impetus to do research on test bias based on groups composed of individuals from different race groups.

The first thorough examination of bias to be undertaken in this country (Owen, 1986) was based on groups composed of individuals from different race groups (whites, Asians, blacks, and coloureds). Race or ethnicity should not, however, form the basis of all bias research. Simple biographical characteristics such as race, sex, SES, etc. can be used to form groups of people who have much in common, but these characteristics might not get to the very nub of what makes people perform differently on tests, or what makes tests perform differently on people. In the cognitive domain (the domain in which most tests measure) it is differences in knowledge and styles of information processing rather than differences in skin colour which underlie differences in performance on intellectual tasks (Taylor, 1987). It is for this reason that a chapter has been included in this report (Chapter 4, entitled "Going Beyond a Priori Groups") in which the possibility is explored of forming groups based on patterns of item responses rather than on some simple biographical variables like race or sex.

The prominence of test bias and fairness issues in the USA has drawn into the debate many individuals who may have a valid stake in this domain but who do not always have the background to grasp some of the concepts. It has become popular to brand tests as being "biased against" or "unfair to" a certain group on the basis of arguments which will not hold water. A number of fallacies have emerged and gained some credence. I shall discuss two of the more prominent ones here. Both were identified and articulated by Jensen (1980).

The first fallacy is what Jensen (1980) calls the egalitarian fallacy. Ability is held to be equally distributed in all groups; consequently, group differences in mean and standard deviation are prima facie evidence of bias. No serious students of bias accept this. The rejection of the egalitarian argument does not mean that one accepts a hypothesis of the fundamental inferiority of one group in comparison with another. A number of factors which have nothing to do with genetic endowment may cause one group to perform more poorly than another on a given test. These factors include education, culture, and exposure to certain types of experience. Differences in performance can also be caused by "self-selection" factors. For instance, vacancies for a particular kind of work might attract the "brighter" members of one group and the "duller" members of another group.

Hence group differences in test mean and standard deviation do not immediately indicate that bias is present. Of course one must also be careful not to commit the opposite mistake: To impute a genuine difference between groups on the basis of a difference in test scores. Factors related to bias may account for all or some of such a difference.

The second fallacy mentioned by Jensen (1980) is the "culture-bound" fallacy. The argument here is that if a test was developed for a certain group (usually the white middle-class or so-called "majority" group) then the items are automatically biased against other groups. The reason given for this is that the items incorporate features which reflect the culture of the group on which the test was developed. Jensen takes the view that the determination of bias must be based on objective psychometric criteria. Many instances have occurred of items which have been judged to be biased because they apparently reflect white middle-class culture, but which on statistical analysis have turned out not to be biased. Conversely, items which appear not to be culture-loaded have emerged as biased in psychometric criteria.

Jensen (1980) takes an extreme view, in that he regards the inspection of the content of items as irrelevant and unimportant. Many researchers advocate the inspection and removal of items which seem to be culture-loaded, but see this as an adjunct to the more important task of statistically examining the items for bias. The aim of a dual approach like this is to produce test material which both looks unbiased and is unbiased on statistical criteria.

What is the purpose of this report? It is not mainly intended as an academic review of the literature on bias and related concepts. There already are sufficient competent publications on this matter to render another one superfluous. In this report the roles and responsibilities of test constructors and test users are examined and guidelines are given for the development of procedures which the test constructor should apply in order to eliminate or reduce item bias in tests.

Tests being developed now and ones which will be developed in the future should be subjected to these procedures, unless there is very little chance that they will be used on more than just one group. Existing tests must also be "put through the hoop", especially those which are being used regularly for selection and placement of individuals of different races. As a result of this process, certain existing tests may have to be modified.

This report therefore has a very practical purpose; consequently I will not spend too much time discussing abstruse theoretical issues. The aim is to make concrete suggestions which can be implemented in practice in the form of procedures which can efficiently find sources of bias.

This report is geared mainly to the needs of the test constructor. The test constructor has certain responsibilities which he must fulfil before making a test available to the test-using community. On the other hand, there are certain checks on the performance of the test which are not the responsibility of the test constructor. The test constructor cannot apply his or her instrument in all possible contexts before releasing it. Certain responsibilities therefore devolve upon the test user, who has to examine the performance of the test in the applications for which he or she is using it, especially if the test is intended to play a key role in the selection or placement of individuals from a variety of "natural" groups.

The domain which we are dealing with in this report is something of a terminological minefield. There are several closely-related concepts and no standardized terminology. One man's culture-loadedness is another man's bias. We must be clear about what we are discussing. In the next chapter I will investigate the relevant concepts and define them as I shall use them in this report. I shall also apportion responsibilities to test maker and test user.

As the detection and elimination of item bias turns out to be the most important responsibility of the test constructor, we devote the bulk of the report (Chapters 3 an 4) to an evaluation of different techniques of item bias detection and propose an overall strategy of bias detection in which specific techniques are "embedded". A summary and evaluation are given in the final chapter (Chapter 5).

2.0 DEFINITION OF TERMS AND ORIENTING COMMENTS

As I mentioned in the previous chapter, there is a lack of standardization of meaning of the terms which are commonly used in this area of research -- terms such as bias, comparability, culture-loadedness, fairness, etc. It is therefore necessary to discuss and define the terms which will be used in this report. We shall begin by discussing the term which can best be used as a conceptual framework: Comparability. Then I shall relate comparability to bias and other concepts. Finally I shall assign responsibilities and roles to test users and test constructors.

2.1 COMPARABILITY

Van der Vijver and Poortinga (1982) point out that there are different types of comparison which can be made between cultures or groups. These authors distinguish four types of universals: Conceptual, functionally equivalent, metrically equivalent, and scalar equivalent universals. These correspond to the four types of measurement which are commonly distinguished: Categorical, ordinal, interval, and ratio.

Conceptual universals are theoretical concepts at a high level of abstraction, like "the psychic unity of mankind", "intelligence" or "adaptability". Van der Vijver and Poortinga (1982) state that the universality of concepts such as these cannot be refuted in experimental studies, as they are at too high a level of abstraction to permit empirical research to be done.

Functionally equivalent or weak universals contain concepts for which empirical referents have been specified, although these may differ across cultures. Hence, different instruments may be used in different groups to establish the equivalence of a given concept; the decision on which instruments to use rests on the appropriateness of these instruments for a given group. It may occur that the same test does not measure the same construct or constructs in different cultures (Frederiksen, 1977). For instance, a test designed to measure crystallized intelligence in one culture may measure fluid intelligence in another because of the unfamiliarity of the concepts in the latter culture. Different tests may therefore be needed to measure the same construct (crystallized intelligence) in the two cultures.

Correlational and factor analytic techniques are generally used when investigating functional equivalence. An example of this kind of research can be found in Owen (1986). Owen administered a battery of cognitive and scholastic proficiency tests to whites, Asians, blacks, and coloureds. The factor structures for each group were simultaneously rotated to a target structure. Owen concluded that the same structure could account for the data of all four groups.

Metrically equivalent or strong universals are measured in the same metric across cultures, although the scales may have a different origin in each culture. Hence, cross-cultural comparisons of absolute scores are meaningless; score differences, however, may be compared.

Scalar equivalent or strict universals contain concepts which are measured in the same metric across cultures and which have the same zero point in each culture. Hence, comparisons of absolute scores across cultures are possible. This is the kind of measurement that is the norm in the physical sciences. The distance to the moon and the diameter of an atomic nucleus can be measured in the same metric.

These four universals should be compared with three types of equivalence or comparability proposed by Poortinga (1971). The three types of equivalence distinguished are: Functional equivalence, score equivalence, and item equivalence.

None of these three types of equivalence corresponds to conceptual universality, presumably because inter-group comparison is not a very meaningful notion at this level of abstraction. Functional equivalence is to be found in both sets of definitions and has the same meaning in both. Score and scalar equivalence also refer to the same concept; this type of equivalence is the basic requirement for scores from different groups to be meaningfully and validly compared. Item equivalence imposes a more stringent requirement: Each item in the measuring instrument must be score equivalent. Hence, each item is treated as a scale in its own right and must be demonstrated to possess the requirements for score equivalence before the test can be called item equivalent. If a test is composed entirely of items which are comparable, it follows that the test score will be comparable. Stated rigorously: A test (X) is score equivalent in two populations if the predicted true score on a common reference or other score equivalent measure (Y) is the same for equal true scores on X obtained in the different populations. A test is item equivalent if each item (taken as a measure in its own right) satisfies the requirement for score equivalence (Poortinga, 1971).

An "eternal regress" problem, which Poortinga (1971) does not address, exists in implementing this procedure. A score equivalent reference measure is required in order to determine whether other tests or items are score equivalent, but where does one get a first or "ur" (reference) measure to establish score equivalence of the reference measure? The only satisfactory way out of this problem is to start with simple psychophysical constructs which can be measured in a direct way. Measures of this kind could be accepted as reference measures without the need for referencing against some other measure. Each measure which was established against an original reference measure as score equivalent (with regard to certain groups) could then be used itself as a reference measure, until ultimately reference tests were available for more abstract psychological constructs.

Clearly this approach is not practicable, as it would entail "re-growing" cognitive psychology and psychometrics. In addition, psychological data are usually not orderly enough to make such an undertaking possible, and the theory is not sufficiently sophisticated to guide the process.

We are therefore in the situation of knowing the theoretical requirements for score and item equivalence, but being unable to perform the required checks in practice. It is therefore necessary to move to a "fallback" position. Perhaps it is possible at least to establish whether some of the <u>necessary</u> conditions for score equivalence are present, even if these conditions are not <u>sufficient</u> to establish score equivalence unequivocally. These issues will be dealt with in the course of the discussion in the next section.

2.2 BIAS AND OTHER CONCEPTS

Comparability is a useful theoretical concept, but psychometricians concerned with the practicalities of detecting and removing incomparability from tests tend to speak of bias rather than comparability. Adherents of the anti-test movement also refer to incomparability as bias.

Unfortunately, the term "bias" is not used to refer to only one concept. At least three usages of the term can be distinguished.

1. If an item or test is <u>judged</u> to be a more demanding or difficult task for Group B to do than for Group A to do, and this difference in difficulty is seen to result from causes which are unrelated to the construct being measured, then the item or test may be called biased against Group B. The main feature distinguishing this usage of the term is that it refers to a judgment; no psychometric analysis is involved.

2. Systematic errors in the prediction of scores on a criterion (such as job or training course performance) which are associated with group membership may be referred to as bias.

3. Test items which are statistically demonstrated to be anomalously "hard" or "easy" for a given group may be regarded as biased.

These three types of bias will be referred to as *judged* bias, *predictive* bias, and *item* bias. Judged bias is sometimes called culture-loadedness (e.g., Chemel, 1985).

Bias is not always *against* a particular group. Bias can also be *in favour of* a particular group. In a study where only two groups are included, bias in favour of one group also means bias against the other group.

We are now in a position to relate these concepts to comparability. As we saw in the previous section there are several types of comparability, but we shall be concerned here only with those types which are necessary for inter-group comparison of absolute scores, viz., score equivalence and item equivalence. Judged bias is clearly not closely related to these types of equivalence or comparability because it is not based on statistical criteria. Predictive bias is also not closely related to comparability because it is detected through the use of an external criterion (score and item comparability on the other hand are examined through the use of a reference instrument which taps the underlying domain measured by the test under investigation).

Item bias and item equivalence are related concepts. In order to investigate the relationship more closely, it is necessary to define item bias in rigorous terms, as was done for item and score equivalence. Two definitions are commonly given:

1. An item of a test is biased against (or for) members of "minority" Group B if, on that item, the members of this group obtain an average score which differs from the average score of "majority" Group A by more (or less) than expected from Group B's performance on other items on the test. In terms of analysis of variance, bias is defined as item x group interaction (Cleary & Hilton, 1968). This definition will be called the unconditional definition of item bias because bias is not defined conditional on ability level.

2. If one holds constant the score on the underlying trait being measured, an item is biased if the probability of a right answer differs among the groups under study (Humphreys, 1986). This definition will be called the conditional definition of item bias because bias is defined conditional on ability level.

If item equivalence or comparability is (somehow) established for a given test in certain populations, then it will follow that the test is unbiased according to both the definitions of bias given above. Hence, both of the above definitions impose necessary requirements for item comparability; the conditions are not, however sufficient. If a pervasive factor (irrelevant to the ability being measured) is present which depresses or enhances the performance of Group B on all items, and which has no impact on the performance of Group A, then the test scores will not be comparable in the two groups. Despite this, the test might not be found to be biased according to the criteria given in the above definitions. Hence it cannot be said with certainty that test scores are comparable in different groups just because techniques based on the above definitions failed to detect bias.

The advantage of the definitions of item bias stated above is that they lead to practical methods of detecting test characteristics which indicate incomparability. It will be remembered from the discussion in the previous section that item and score equivalence/comparability cannot be examined easily in practice.

What is the relationship of fairness to bias, particularly predictive bias? We shall take the view in this report that fairness is at issue only when tests are used for selection and placement. Predictive bias is defined in terms of errors of prediction of performance on criteria, and is consequently closely associated with fairness issues. Predictive bias and fairness are not synonymous, however. A test which is predictively biased against a certain group is not inherently unfair; it is only when decisions are made on test results that one may speak of the fairness of these decisions. A biased test can be used fairly and an unbiased test can be used unfairly.

According to Jensen (1980):

"A test is a biased predictor if there is a statistically significant difference between the major and minor groups in the slopes, or in the intercepts, or in the standard error of estimates of the regression lines of the two groups, when these regression parameters are derived from the estimated true scores of persons within each group" (pp. 381, 382).

If a test is unbiased as a predictor of a particular criterion, then the use of a single regression line to select individuals on the basis of test scores is fair, at least according to some models of fairness. On the other hand, if there is a statistically significant difference on any of the three parameters stated in the definition, the test is biased, and the use of a single regression line to select individuals from both groups would be unfair according to most models of fairness. If, however, separate regression lines are used for each group, the (biased) test may be used fairly to make selection decisions.

Sophisticated models have been developed to search for evidence of predictive bias. Lautenschlager and Mendoza (1986), for instance, present a hierarchical model which enables bias due to differences in slope or intercept to be pinpointed. More primitive approaches, such as Hunter, Schmidt and Hunter's (1981) approach which aims only at detecting differential validity, are not recommended. A differential validity approach is not sensitive to differences in intercept. Also, Drasgow (1982), has shown that item bias does not always show up in differential validity. Using synthetic data, Drasgow created a test in which 25% of the items were biased against a hypothetical minority group. There were only very small differences in the predictive validities of the test in the majority and minority groups, despite the large bias against the minority group.

Chemel (1985) points out that, in the definition of predictive bias, no reference is made to the construct that the test is intended to measure. Differences in slopes, intercepts, and standard errors of estimate are the only concerns. Hence a test could, in theory at least, measure different constructs or combinations of constructs in different groups and still not be biased according to Jensen's (1980) definition. Functional equivalence is not a requirement for predictive validity.
It is unlikely, but not impossible, that a test which measures different things in two groups will have a relationship between test scores and a given criterion which does not differ significantly between groups on any of the three parameters mentioned above.

It should also be noted that predictive bias research always involves a particular criterion. If in Groups A and B predictive bias is not found for a given combination of test (T) and criterion (C1), this does not mean that if another criterion (C2) is used in combination with T predictive bias will not be found. Linn (1984) discusses the implications of this for fairness in selection decisions:

"....the notion that fairness can be achieved by selecting those people with the highest qualifications and treating equally those with equal qualifications enjoys rather a broad support. But this broad support tends to become fragmented once idealized qualifications start being translated into operational measures. It is one thing to agree that the most qualified should be selected. It is quite another to agree that a particular measure of performance on the job or in the educational institution, that is, a criterion measure, is an adequate measure of those idealized qualifications. ....The Achilles' heel of criterion-related validity is, of course, the criterion. The real concern for a test to be used in selection is with its validity as an indicator of the idealized qualifications. The degree to which it predicts a criterion measure is of concern only to the degree that the criterion measure is itself a valid indicator of those idealized qualifications." (pp. 37, 38).

The implication is that there is no "definitive" predictive bias study which will give the "final word" on a test. Apart from the fact that criteria are fallible indicators of "qualifications", the nature of these qualifications will vary from setting to setting.

Predictive bias has been researched fairly extensively, especially in the USA. Not much evidence of predictive bias against "minority" groups has been found (Drasgow, 1987; Reynolds, 1983). Most of the research has used groups composed of American blacks and whites (e.g., Cleary, 1968). Zeidner (1987), however, examined predictive bias in groups of Israeli students of different ethnic membership; again the tests were in general found to be predictively unbiased.

Where evidence of predictive bias has emerged, the bias has generally been against the "majority" or "advantaged" group. This situation can occur when a single regression line is used to make selection decisions for both the majority and minority groups. As the regression line for minority groups is found in most studies to be below that of the majority group, the use of a regression line based on a combined population results in the over-prediction of criterion performance for the minority group and the under-prediction of performance for the majority group (Cole 1981; Jensen, 1980).

It would be unwise to assume that these findings have validity for South Africa, especially in the light of Linn's (1984) comments (see above). J M Taylor (1986) examined the relationship between test results and academic performance in samples of black and white technikon students. The Technical Reading Comprehension Test and the Blox tests were the predictors. For both predictors the regression line for blacks was above that for whites. It is possible, however, that the criterion was not totally comparable for blacks and whites. One of the difficulties of doing this kind of research in South Africa is that comparable criteria are generally not available. In the Taylor study, for example, the blacks and whites attended different technikons which might have had different standards.

Although I have referred extensively to regression lines, it should not be concluded from this that all fairness models are based on regression. Many are based on equalizing certain proportions in different groups. (An impressive review of the main models is given in Petersen & Novick, 1976.) Inherent in each fairness model is a value judgment about what constitutes fair selection. As there is no consensus on what values are best, there is not likely to be consensus on what fairness model is best. Decisions on what constitutes fairness in selection should be in the hands of policy makers, not psychometricians. The issues in question are ultimately "political" (in a broad interpretation of the word). The contributions on test bias made by psychometricians can be used only as a technical input to help policy makers to come to more informed decisions (Cole, 1981).

2.3 ROLES AND RESPONSIBILITIES

In Chapter 1 we stated that test constructors and test users have different responsibilities with regard to the investigation of bias. In the first two sections of this chapter we looked at the comparability concept and related it to the various types of bias. We also examined the relationship of predictive bias to fairness in selection. The task which remains is to delineate the roles and responsibilities of the test maker and test user in the light of the analysis of the comparability, bias, and fairness concepts given in the first two sections of this chapter.

Two statements can be made with certainty:

1. The detection and elimination of item bias is solely the responsibility of the test constructor.

2. The provision of a definitive fairness model is <u>not</u> the responsibility of the test constructor.

The application of item bias detection methods should be as much part of the test construction and refinement process as the application of item analysis procedures. Both of these techniques should be used to decide on which items to include in a test and which to exclude. By "item bias detection methods" we mean primarily statistical methods, although judgmental methods are not excluded. Tests which were constructed before item bias detection became a salient issue in psychometrics should also be subjected to item bias detection procedures if they are regularly used in multi-cultural selection contexts. Clearly, these actions can only be undertaken by the publisher of the test material.

As I mentioned earlier in this chapter, fairness models are based on values -- on a conception of what constitutes the greatest good. Fairness models must be decided upon by the institution which is making the employment or enrolment decisions (or possibly by government). The test publisher cannot prescribe for the user in this regard, for this would amount to moral preachment. It should be borne in mind that issues of fairness in employment arise even when tests are not part of the selection process. The furthest the test publisher can go is to publish a document or documents which dispassionately describe the issues which should be considered when framing a policy of "fair" selection.

I shall now discuss the roles of the test constructor and test user with regard to the determination of functional equivalence and predictive bias.

Functional equivalence has to do with the construct validity of a test in different groups. Does the test measure the same construct or constructs in different groups? The test constructor should perform construct validation studies when developing a test. In fact, research should be done on a virtually on-going basis to expand knowledge on what the test measures in different contexts. Research organizations which construct tests generally do this, although not necessarily in a very formal fashion, and not necessarily always using an approach specifically or primarily designed to investigate functional equivalence. The onus should be on publishers of tests to perform more research of this kind and to make the results available to test users.

For test users, predictive bias is often the most pressing concern. Most users would like to be given guidelines by the publisher on how to make selection decisions in a multi-cultural or multi-group context. They would like to be told something fairly simple, like: "Use the same norms and cut-offs for all applicants," or: "Use normtable x for Group A and normtable y for Group B, and cut-offs corresponding to stanine S." Unfortunately, a test publisher cannot make such blanket statements.

For almost all models of fairness, some form of empirical research has to be done in order to collect the data necessary to implement the fairness model in question. (An exception is the quota model.) As we pointed out in the previous section, an "answer to all questions" predictive bias study cannot be performed. Hence, the onus is on the organization using the test to perform its own research, using predictor and criterion data from its own applicants. If the organization lacks the skills to do this, it should hire these skills.

This does not mean that the test publisher should have no part in predictive bias research. Some studies should be done, especially in areas where the test in question is used extensively. From such studies, some regularities concerning the test's performance might emerge. But if a test user simply applies the findings from these studies in his own organization, he does so at his own risk.

This situation is not radically different from that found in single-group validation exercises. The test publisher performs certain illustrative predictive validation exercises before making the test available. If a user decides that the findings of these studies provide sufficient justification for him to use the test without further research on a similar group in his own organization, he does so at his own risk. The user should bear in mind however that the risk of severe consequences such as litigation is greater when selection is done in a multi-group context.

The first and most pressing responsibility of the test constructor is, then, the investigation of item bias in all groups in which the test is commonly used. The rest of this report will be devoted to this topic.

3.0 BIAS DETECTION APPROACHES

In this chapter we will review the main within-test bias detection techniques and discuss their virtues and drawbacks. Certain refinements will be suggested. It should be borne in mind that bias detection methods are not comprehensive specifications of how to identify biased items; they describe the statistical technicalities of bias detection but not the overall strategy of identifying biased items with maximum accuracy. Consequently, there will be a section in this chapter dealing with issues relevant to strategy.

All within-test bias detection methods are incapable of finding pervasive bias. Pervasive bias is present if performance on all items of a test is underestimated (or overestimated) for a given group. Within-test bias detection methods can do no more than identify "relative bias", i.e. identify items which appear to be giving an anomalous reflection of ability in a given group of testees relative to the estimate of ability obtained from the rest of the test. Anomalies are detected by making inter-group comparisons.

In all, I shall examine six bias detection techniques in this chapter. The first two are based on the unconditional definition of bias; the remaining four are based on the conditional definition.

3.1 TECHNIQUES

3.1.1 Analysis of Variance

According to Cleary and Hilton (1968, p. 61):

"An item of a test is said to be biased for members of a particular group if, on that item, the members of the group obtain an average score which differs from the average score of the other groups by more or less than expected from performance on other items of the same test. That is, the biased item produces an uncommon discrepancy between the performance of members of the group and members of other groups."

In analysis of variance terms, discrepancies of this kind reveal themselves in group x item interaction. Main effects are not the primary focus of attention: Significant main effects do not signal bias in a test. Genuine differences in ability between groups will be reflected in a group main effect; similarly, differences among items in terms of difficulty will be reflected in an items main effect (Osterlind, 1983). A significant interaction between groups and items, however, is indicative of bias in the measurement instrument.

If a significant groups x items interaction is found, the application of a secondary procedure is required to identify those items which are biased (e.g., the Scheffe method of multiple comparisons and the Transformed Item Difficulties method). Plake (1981) discusses a sophisticated procedure for identifying those items contributing to the items-by-group interaction.

Osterlind (1983) has a number of criticisms of the ANOVA procedure for bias detection. Bias may be present in a small proportion of items, and this may not be sufficient to result in the rejection of the null hypothesis. One would nevertheless want to identify the offending items so that they could be modified or removed from the test. A further shortcoming of the method is that it is based entirely on group averages, a fundamental characteristic of unconditional methods. If in a two-group study an item is biased against Group A at lower ability levels and against Group B at higher levels, these two effects may cancel each other out and the item will appear to be unbiased. The reader should refer to the section on the Item Characteristic Curve method for a further treatment of the topic of bias as a function of ability level.

The ANOVA method is not frequently used in current research on bias detection. Its popularity in the earlier days of bias detection was largely due to the familiarity of its methodology.

3.1.2 Transformed Item Difficulties

The Transformed Item Difficulty (TID) method was developed by Angoff and Ford (1973). Suppose that a test has been developed and standardized on a given group and that information is available on the difficulty levels (*p* values) of the items. If the test is then administered to another group, one would expect that the pattern of item difficulties would remain the same as those obtained in the original group. By pattern is meant the difficulty levels of items relative to other items in the test. As the groups may differ with regard to the distribution of ability under consideration, one would not require that the *p* values of each item be of the same order in both groups, but one would expect that the *pattern* of item difficulties be comparable. Departures from this situation are indications of bias.

The comparison is done by transforming *p* values to standardized scores based on the *z* statistic. Separate transformations are performed for each group, based on the scores of that group. Not only does this remove any effects due to differences in ability between the two groups, but it also linearizes the difficulty indices; hence meaningful comparisons between pairs of groups can be made on a bivariate graph.

The transformation usually performed on *p* values is the Delta transformation:

$$\Delta = 4\underline{z} + 13,$$

where $\underline{z}$ is defined so that 100 x $\underline{p}$% of the area of a normal (0,1) distribution is greater than $\underline{z}$. This transformation effectively removes negative values. A Delta range of 0 to 26 corresponds to a $\underline{p}$ from 0,999 to 0,001.

For an n-item test and a two-group study, n pairs of Deltas are computed and plotted on a bivariate graph. (In studies which include more than two groups, a number of bivariate graphs can be plotted, with the data from the majority group represented in each case.) Typically, an elongated ellipse results; the narrower the ellipse, the more similar the pattern of item difficulties in the two groups under consideration. A straight line representing the major axis or principal component of the ellipse is drawn through the scatterplot. The vertical distance of a point from this line reflects the degree of bias inherent in the item represented by the point. Points falling above and below the major axis indicate bias in different directions (in favour of different groups).

Various methods exist to evaluate the distance of points from the major axis for the purposes of identifying bias as opposed to random effects (see Jensen, 1980; Rudner, Getson, & Knight, 1980). According to Osterlind (1983), a limit of 0,75 z score units away from the line is often used for establishing acceptable boundaries of bias magnitude.

Oosterhof, Atash, and Lassiter (1984) recommend the use of what they call Delta-charts to evaluate bias trends in the test as a whole. A Delta-chart consists of a graph with items sequentially numbered on the abscissa and the Delta values on the ordinate. The points associated with a given group are joined by a line. All groups under consideration are represented on the same chart. This procedure facilitates the detection of bias effects in adjacent items. If, for instance, the distance between the lines of Group A and Group B widens near the end of a test, this could indicate that time constraints are affecting performance differentially in the two groups.

The TID method is conceptually appealing and easy to apply. It is not surprising, therefore, that it has been widely used. The method has its drawbacks, however, when used to evaluate bias in groups which differ markedly in mean ability (Shepard, Camilli, & Williams, 1985). An item which discriminates effectively between low and high ability individuals (a positive feature for an item) also tends to discriminate between high and low ability groups. This will emerge in the TID method as evidence of bias. This tendency of the method spuriously to identify bias in highly discriminating items is a function of the unconditional nature of the technique. Angoff (1982) suggests matching the samples on ability in order to overcome this difficulty. Usually this means trimming the higher-scoring sample in order to bring its average performance in line with that of the lower-scoring group.

3.1.3 Item Characteristic Curve

The Item Characteristic Curve (ICC) method is universally regarded as the "Rolls Royce" of bias detection methods. This approach is based on Lord's (1980) latent trait theory. Lord (1977) discusses the application of the theory to bias detection. The ICC method is the most satisfactory way of testing for bias as defined in the conditional definition: An item is unbiased if all individuals having the same underlying ability have an equal probability of getting it correct, irrespective of group membership.

The term "underlying" in the above definition indicates that an error-free index of ability is required in order to investigate bias rigorously. Test scores are estimates of underlying ability but contain error; this error might also be more prevalent in one group than another or at one ability level than another.

An item characteristic curve expresses the probability of obtaining an item right as a function of underlying ability. These curves are usually S-shaped. In the complete ICC model, three parameters are used to describe the curve: Item difficulty, item discriminating power, and the probability of answering the item correctly given very low ability (this is a guessing parameter). Different curves will result depending on the values of these parameters.

Techniques exist for plotting item curves for more than one group on the same graph. Consequently, the probability of answering the item correctly for different groups can be compared at all ability levels. If the curve for Group A is above that for Group B at a given point along the ability continuum, this indicates that Group A individuals at this ability level find the item easier than Group B individuals at the same level. The situation might be different at other ability levels. If the curves cross, this indicates that an interaction occurs between group and ability level. Such interactions cannot be detected in unconditional methods as performance is averaged across the ability range. A case where group x ability interaction occurs is illustrated in Figure 1.

ICCs can be used in a wide variety of ways to assess bias. If an item is unbiased, the ICCs will be very similar for the groups under consideration. Some of the ICC-based methods of assessing bias involve measuring the area between the curves. Both "unsigned" and "signed" methods are used for measuring area. When the unsigned method is used, the total area between the curves is calculated, and no account is taken of whether Group A's curve is above Group B's, or vice versa. Hence in Figure 1,

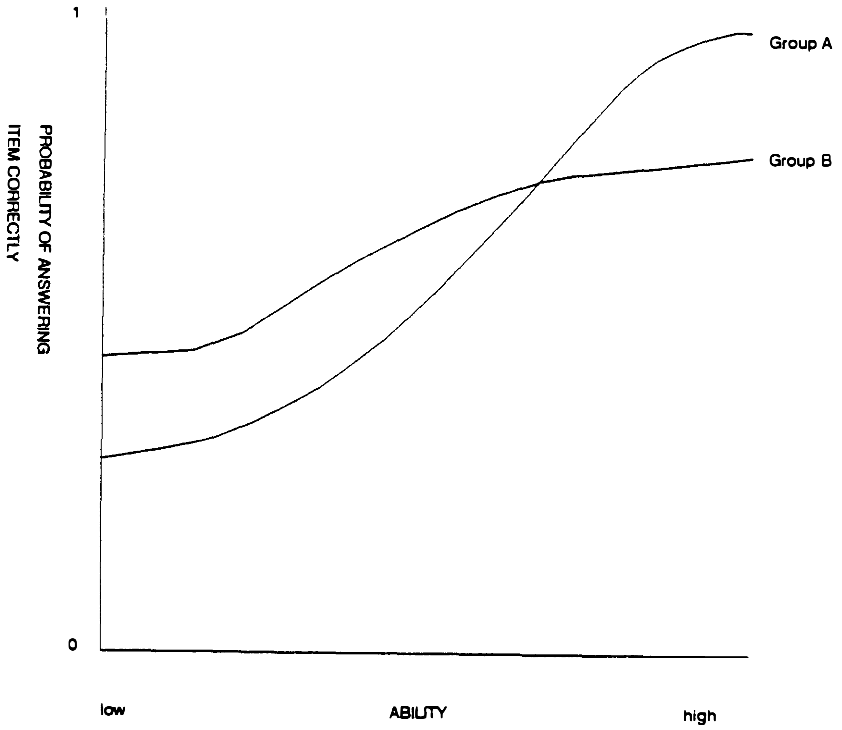

Figure 1: ICCs for two groups of testees.

the area between the curves to the left of the intersection point would be added to that to the right of the intersection point. In the signed method, account is taken of the relative positions of the curves, and areas can cancel each other out.

A description of some of the other ways of assessing bias in the ICC approach are given in McCauley and Mendoza (1985) and Shepard, Camilli, and Williams (1985).

Although the ICC method has a number of admirable features, it also has several drawbacks. Large sample sizes (in excess of 1000) are required to estimate the parameters accurately. Some authors, such as Bock and Aitkin (1981), Linn and Harnisch (1981), and Rigdon and Tsutakawa (1983) have developed latent trait techniques which might function effectively with smaller sample sizes. The Linn and Harnisch approach, for instance, starts by estimating ICC parameters on a combined sample, and then becomes more like a chi-squared procedure, as interval comparisons are done. It is still necessary, however, to have fairly substantial numbers, as the combined sample should still exceed 1000 for reasonable parameter estimates to be achieved. Another problem is that parameter estimation becomes problematical if the groups differ widely in ability. Another drawback of the ICC method is that the computer programs required to use the method are costly to run and require highly skilled personnel.

3.1.4 Chi-Square

We saw from the previous section that one of the most serious drawbacks of the ICC method is that large sample sizes are required in order to estimate the parameters adequately. Clearly there is a need for a method which is based on the same (conditional) rationale as the ICC method, but which will work on smaller sample sizes and will be easier to apply. Both the chi-squared method and the log-linear method (discussed in the following section) attempt to answer this need.

As one might expect, sacrifices have to be made in exchange for these desirable characteristics. The ability continuum is divided into ability ranges or categories; hence ability is no longer treated as a continuous variable as it is in the ICC method. Also, manifest ability scores rather than latent (error-free) scores are used. In order to take account of these limitations, we have to modify our conditional definition of bias: An item is unbiased when individuals in the same ability range (where ability is determined from manifest scores) have the same probability of answering the item correctly, regardless of group membership.

The chi-squared method was developed by Scheuneman (1979). The total range of the test is divided into a number of score intervals or categories (usually five to seven). Care should be exercised in selecting the categories, as frequencies falling into each category should not be too low or too high. The intervals must be selected to suit the data obtained from each item; this can easily be done, as each item is tested separately for bias.

For an m group study in which the total test range is divided into r categories, an m x r contingency table is created, with the entry in each cell representing the frequency of individuals of a given group and ability category who answered the item correctly. Marginal totals are employed to calculate expected frequencies under the hypothesis that group membership is unrelated to the probability of answering the item correctly within each ability category. A chi-squared procedure can then be applied to test this hypothesis. The number of degrees of freedom are (m - 1)(r - 1).

Criticisms have been levelled against the Scheuneman approach because it is based on only "half" a chi-squared model. The frequencies of individuals answering the item incorrectly in each ability category are not represented. The Scheuneman statistic does not have the proposed chi-squared distribution. Rudner et al. (1980) state that chi-squared values are inflated when the observed score distributions are different in the groups under consideration; Chemel (personal communication) is of the opinion that the magnitude of item difficulties plays a major role in the distortion of the statistic. Full chi-squared procedures have been proposed by Intasuwan (1979) and Ironson (1982). Owen (1986) used both the Scheuneman and Intasuwan methods to detect item bias in several HSRC tests. In this study, the Intasuwan full chi-squared method identified more items as biased than the Scheuneman method.

3.1.5 Log-Linear and Logit

Log-linear models provide a means of analyzing qualitative data within a framework which is similar to that of analysis of variance. There is also a strong connection, however, to psychometrics in the form of latent trait theory which employs the logistic function. The goal of the technique is to provide a linear approach to the analysis of frequency data. The major accomplishment of the log-linear model is that it provides a unified approach to the analysis of multidimensional contingency tables (Baker, 1981).

The log-linear model is based on the same conception of bias as the chi-squared model and can be applied to data in the same format as that required for the chi-squared technique. The approach is much more sophisticated, however, in that it permits the rigorous testing of a series of models, and the comparison of models with one another.

Log-linear models acquired their name from the fact that products become sums after logarithms have been taken. Sums are statistically more tractable than products. Each term in a log-linear model represents a contribution to the frequency of the cell; less restrictive models can be created simply by adding the appropriate term or terms to the model, and more restrictive models can be achieved by doing the opposite. The basic goal of the log-linear approach is to find the most parsimonious model that fits the data. The researcher usually (but not always) starts with the most complex model and proceeds towards simpler models. The goodness of fit of each model is ascertained by a chi-squared test. The change in goodness of fit between two <u>nested</u> models can be tested through the residual chi-squared static: The difference between two chi-squared statistics is itself distributed as chi-square. (Model A is nested in Model B if Model A can be created by placing certain restrictions on Model B.)

Logit-linear models differ from log-linear models in two major respects. Firstly, the statistics of the logit-linear approach are based on the multivariate logistic function. Secondly, a distinction is made in logit-linear models between explanatory and response variables, whereas in log-linear models, all variables are regarded as response variables. In practical terms this means that in log-linear models the procedure is to draw a single sample from a population and then categorize the subjects on the basis of one or more response variables, whereas in logit-linear models, groups of individuals are drawn from populations according to some sampling plan based on the explanatory variables; individuals' responses are then categorized on the response variable(s). Despite these differences, it is possible to obtain identical results with the two approaches if the analyses are properly formulated.

An introductory treatment of the statistics of log-linear and logit-linear models is given in Baker (1981). More advanced treatments are to be found in Fienberg (1977) and Bock (1975).

Let us now direct our attention to bias-relevant issues and consider a three-dimensional contingency table which could reflect the responses (k) of different groups (j) in different score categories (i) of a test. As we will regard responses as either right or wrong, k will have only two values, 1 (wrong) and 2 (right). If only two groups are involved in the study (e.g., whites and blacks) then j will also only have two values, 1 (whites) and 2 (blacks). There may be several score categories on the test; if five, i will assume the values from 1 to 5. (Note: The assignment of numbers to categories is quite arbitrary and could be changed without affecting the results.) The frequency in any cell of this matrix will be represented as F(ijk). F(312) for instance is the frequency of whites in score category 3 who correctly answered the item under investigation.

The saturated (least restrictive) log-linear model in the three-dimensional case is:

$$\begin{aligned}\ln F(ijk) = u &+ u(1(i)) + u(2(j)) + u(3(k)) + u(12(ij)) \\ &+ u(13(ik)) + u(23(jk)) + u(123(ijk)),\end{aligned}$$

where u is an overall effect parameter,

u(1(i)), u(2(j)), and u(3(k)) are parameters specific to score category, group, and response (right-wrong) respectively,

$u(12(ij))$, $u(13(ik))$, and $u(23(jk))$ are interactions between two variables, and

$u(123(ijk))$ is the interaction between all three variables.

The similarity to analysis of variance can be seen from the above.

In logit-linear terms, score category and race are explanatory variables, whereas whether the individual answered the item correctly or not is a response variable. The natural logarithm of the "odds" of answering correctly against incorrectly can be expressed:

$$\begin{aligned}
\ln\,(F(ij1)/F(ij2)) &= \ln F(ij1) - \ln F(ij2) \\
&= \{u(3(1)) - u(3(2)\} + \{u(13(i1)) - u(13(i2))\} \\
&\quad +\{u(23(j1)) - u(23(j2))\} \\
&\quad + \{u(123(ij1)) - u(123(ij2))\} \\
&= 2u(3(1)) + 2u(13(i1)) + 2u(23(j1)) \\
&\quad + 2u(123(ij1)) \\
&= C + S(i) + G(j) + (SG)(ij)\ldots\ldots\ldots\ldots\ldots(1)
\end{aligned}$$

where C is the overall item difficulty,

$S(i)$ is the main score category effect,

$G(j)$ is the main group effect, and

$(SG)(ij)$ is the score category x group interaction.

Bias in an item is revealed in the (SG)(ij) and G(j) terms. If the model:

$$\ln (F(ij1)/F(ij2)) = C + S(i) + G(j) \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

does not fit the data adequately, then the item is biased. The nature of the bias is complicated because the interaction between score category and group is needed to explain the data. Mellenbergh (1982) calls an item for which (SG)(ij) has to be included nonuniformly biased. Even if an item is found not to be nonuniformly biased, it might still be uniformly biased. Consider the model:

$$\ln (F(ij1)/F(ij2)) = C + S(i). \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

If model (2) fits the data and the difference in fit between models (2) and (3) is significant, then the item is uniformly biased: The logit differences between the two groups are independent of score categories. If model (3) fits the data and the difference of fit between models (2) and (3) is not significant, then the item is not biased (Mellenbergh, 1982).

Some probability level of the associated chi-square statistic has to be set in order to decide in each case whether the model fits adequately. This could be 5% or 1%, depending on how strict the investigator wishes to be and how willing he is to erroneously identify unbiased items as biased.

The sophistication of the log-linear and logit approach can be seen from the above discussion. The model-fitting procedures which can be done in this approach are impossible in the standard chi-squared approach described in the previous section. The application of log-linear modelling described above is only one of a number of useful applications which have relevance to the bias issue. Baker and Subkoviak (1981) discuss other uses of the technique, including investigations into group differences in response patterns.

Van der Flier and his associates have made extensive use of log-linear and logit models in bias detection (Kok, Mellenbergh, & van der Flier, 1985; van der Flier, Mellenbergh, & Ader, 1984; van der Flier, Mellenbergh, Ader, & Wijn, 1984). The refinements which these authors have introduced into the approach are discussed in section 3.3.

3.1.6 Regression and Partial Correlation

Regression is a technique widely used in studies on predictive or test bias (as opposed to item bias). Slopes, intercepts and errors of prediction are tested for differences in the groups under study. Raju and Norman (1985) attempt to make a case for regarding test bias and item bias as conceptually similar. These authors use a regression approach to investigate what is commonly called item bias.

Raju and Norman (1985) propose that the individual item score be treated as the criterion and the total score (excluding the item under consideration) as the predictor. As in predictive bias research, regression lines are calculated for each group and these are tested for equality. Since most of the predicted item scores are likely to fall between 0 and 1, the predicted item score may be viewed as the probability of success on the item. Non-equality of regression lines means that the probability of success given the total test score depends to a certain extent on group membership. The authors point out that their method (which they call the Regression Bias method) takes into account both the item p value (the proportion of testees answering the item correctly) and the item discrimination index (the correlation between the item score and total test score). The item discrimination index has been used widely by test constructors for many years as a gross indicator of the performance of an item in different groups, but is of limited utility when groups differ substantially on ability, due to the limitations on item-total correlations imposed by extreme p values.

It should be noted that the Regression Bias method is conceptually similar to the ICC method, except that: (1) manifest scores rather than latent scores are used as indices of ability, and (2) straight lines rather than logistic curves are fitted to the data. It will be remembered from the discussion on the ICC method (section 3.1.3) that three parameters are employed in the full ICC model: Difficulty, discrimination and guessing. The Regression Bias method constitutes a crude method of determining group differences in the discrimination parameter (crude because a straight line rather than logistic curve is used). One would expect little similarity between the results obtained by the Regression Bias method and the area-between-the curves bias indices obtained from the ICC method, because the former uses straight lines.

Stricker (1982) proposes a bias detection index based on the partial correlation. The index is the item's partial correlation with group membership (coded numerically), with total score (excluding the item under consideration) held constant. The author points out that like the ICC index, this procedure controls for group differences in overall ability. However, unlike the ICC index, this index compares the general difficulty of the item for groups, essentially averaging differences between them at various levels of ability; hence items which are identified as biased are those in which one group is consistently favoured. The procedure cannot identify items which perform differently at different levels of ability, favouring one group at one point and penalizing it at another.

The assumption that the nature of bias is the same throughout the ability range might not be justified in many instances. In an empirical study, Stricker (1982) found that the ICC and partial correlation methods agreed quite well when the groups under study were males and females, but not when the groups were blacks and whites. This might indicate that the assumptions of the partial correlation method were severely violated in the latter case.

3.2 EVALUATION

Several empirical studies have been conducted to investigate the effectiveness of the major bias detection methods.

Ironson and Subkoviak (1979) compared the following methods: TID, item discrimination (item-total correlations), chi-squared, and ICC (area between curves). These authors tested 1691 blacks and 1794 whites on six tests. White vs. white comparisons were used to give a baseline for measuring inherent error in the bias indices. The magnitude of the bias indices in the black vs. white comparisons were roughly two to three times as large as the magnitude of the white vs. white comparisons. The values of the TID, chi-squared, and ICC indices were highest for the items of a vocabulary plus reading subtest The second-highest bias values were for a mathematics test, and the lowest bias values for a picture-number association test.

Correlations between bias indices varied from essentially zero to 0,82. The largest correlations were between the ICC and chi-squared approaches. As the ICC approach is often regarded as the standard against which other methods are judged, the substantial correlations found between ICC and chi-squared methods suggests that the chi-squared method was the second-best of those studied.

The authors examined the items identified most consistently as biased. The source of bias was not immediately apparent, although there were many verbal items. Several other authors have also commented on the lack of apparentness of sources of bias (e.g., Shepard et al., 1984; Flaugher, 1978).

Subkoviak, Mack, Ironson, and Craig (1984) compared the TID approach, two chi-squared techniques, and the ICC method. Bias was manipulated in the study by constructing certain items designed to favour blacks. These items were then imbedded in a set of more neutral items, and the test was then administered to black and white populations. Unlike the Ironson and Subkoviak (1979) study, therefore, items known in advance to be biased were included in the test; consequently a more satisfactory method of evaluating the performance of the bias detection techniques was available.

The intercorrelations between the bias measures were high (0,85 to 0.95). Correlations were computed between a priori bias (zero-one coding) and the bias indices. Both signed and unsigned indices were used. The ICC method proved to be the most effective at identifying a priori bias. A correlation of 0,87 was obtained between the unsigned measure and a priori bias; the corresponding correlation using the signed measure was 0,88. Chi-squared and TID measures performed about equally well, with correlations in the low 0,70's (unsigned) and 0,80's (signed). Hence, all methods performed well (especially when one bears in mind that some unintended biased items might have been present) with the ICC method being marginally better than the others.

Ironson, Homan, Willis, and Signer (1984) also constructed a test in which bias was manipulated. Tests may be biased against a Group, B, if test scores in this group are influenced by a secondary ability which does not influence scores in Group A. For instance, if Group A are first-language English speakers whereas Group B are first-language Zulu speakers, and the test in question is a verbally presented test of mathematical ability, then the test may be biased against the Zulu group because not all of Group B have the English language skills to understand every item in the test, whereas all of Group A can understand the verbal component of the problems. Ironson et al. "planted" items in a verbally presented mathematics test which were above the reading skills of one of the groups. Three bias detection methods were used: TID, a full chi-squared method, and an ICC procedure adapted for small sample sizes, based on the work of Linn and Harnisch (1981).

Somewhat surprisingly, the biased items did not prove to be relatively more difficult for the group with low reading skills as compared with the other items. Only the ICC method proved effective in identifying the planted biased items. The authors suggested that the ICC method may have worked well because it is more sensitive to test multidimensionality than the other methods.

McCauley and Mendoza (1985) also investigated the effectiveness of various ICC-based detection indices in identifying test items which require a secondary ability on which two groups differ in mean level. In addition, they examined the ability of bias detection techniques to identify as biased items which have different factor structures in different groups. Artificially generated data were used. The results indicated that most of the bias detection indices employed were effective at indentifying items which require a secondary ability on which the groups differ in mean performance. Signed and unsigned between-curves areas proved to be good indices. The bias detection methods were less effective at identifying items which had different factor structures in different groups. Hence, the methods were not effective at detecting lack of functional equivalence in items.

Shepard, Camilli, and Williams (1985) examined the effectiveness of the TID technique, a full chi-squared method, and the Linn and Harnisch (1981) small N ICC method. Both real and simulated data were used. In the case of real data, the standard three-parameter ICC method was used in a cross-validated procedure to identify biased items. In the simulated data, 54 item characteristic curves were generated using a combination of parameter values. On 18 items, bias was introduced by increasing the difficulty of the item for one of the groups.

The small N ICC procedure proved to be the best at identifying bias, followed closely by the chi-squared method and the TID method. The correlations of the TID Delta index with the criterion were the least stable across studies. The performance of the TID method increased substantially when samples matched on ability were used. (It will be remembered from the discussion in section 3.1.2 that the TID method tends to identify items which discriminate between high and low ability groups as biased. By matching samples on ability, this shortcoming can be overcome.)

Theoretically the ICC bias detection method is the most satisfactory. The results of the studies quoted above offer empirical confirmation of the superiority of the method.

However, the ICC method is simply not usable in many applications for a number of reasons, the most important being that sample sizes required by the method cannot be obtained. Even the small *N* variations of the ICC method (such as the Linn-Harnisch technique) require combined samples in excess of 1000.

The chi-squared approach usually turns out to be the second-best method. This method has the advantage that it is easy to apply and does not require very large samples. The log-linear method, which also uses categorical frequency data, is theoretically superior to the chi-squared approach. Unfortunately, little comparative research has been done using the log-linear method, as this method has only recently been developed. Mellenbergh (1982), however, reports correlations in the vicinity of 0,8 between indices based on the chi-squared and log-linear methods.

The log-linear method has been used on its own in several studies and has produced good results (Kok, Mellenbergh, & van der Flier, 1985; van der Flier, Mellenbergh, & Ader, 1984; van der Flier, Mellenbergh, Ader, & Wijn, 1984). The "acid test" when evaluating bias methods is to compare the performance of the methods against a genuine criterion (i.e. known bias). This is equivalent to validating a test against a relevant external criterion. The log-linear method has apparently not yet been put through such an acid test, but it is to be expected that it would out-perform the chi-squared method, but not perform as well as the ICC method.

The TID method usually performs rather indifferently in cases where the groups under investigation differ substantially in mean ability. However, when groups are matched on ability, the method appears to perform quite well. It is an easy method to apply and has the advantage that it is based on different principles from the ICC, chi-squared, and log-linear methods. It can therefore be used to give an independent assessment of item bias.

3.3 REFINEMENTS

Conditional bias detection techniques which use score categories are based on the same basic bias definition as the ICC technique, but they are cruder that the ICC method for two main reasons. One, ability is divided into a number of categories or ranges, rather than being treated as a continuum. Two, test total is used as an index of ability.

Although the division of ability into categories results in the loss of information and consequently in the loss of accuracy in the description of the performance of items, this approach does have the advantage that useful information can be obtained with smaller sample sizes. From the studies reported in the previous section, it appears that the categorical methods are only marginally less powerful than the ICC method; many researchers are willing to "trade off" this reduction in performance in exchange for smaller _N_'s.

Test total is a "second rate" index of ability because it contains error and is not perfectly correlated with the underlying ability. If there are biased items in the test, these contribute to the test total; but test total is used to categorize people on ability so that items can be tested for bias.

Clearly there is a circularity here: A total which may be biased -- and therefore result in the misclassification of individuals of certain groups with regard to ability -- is used as part of the process of finding biased items. If there are several items in a test which are biased against Group B in comparison with Group A, then members of Group B will be placed in lower ability categories than they "deserve" to be; this in turn will result in fewer items being identified as biased against Group B.

One way of minimizing this problem is to "clean up" the test total by using an iterative procedure to remove biased items, so that ability can be estimated using only items which appear from the bias detection procedure to be unbiased.

European researchers have been particularly active in this field. Van der Flier and his colleagues have not only developed the log-linear and logit models to detect bias; they have also developed iterative procedures to minimize the problem of bias in the total score. Van der Flier, Mellenbergh, Ader, and Wijn (1984) describe the basic steps of the procedure as follows.

1. For each item, denoted m, the following steps are taken:

   a. A test score is computed for each subject as the sum of scores for all items except (i) item m and (ii) those items classified as biased in the preceding iteration.

   b. The overall frequency distribution of the test scores is computed. Subjects are then divided into groups of equal size. If necessary, the subjects who obtained a particular total score are randomly assigned to the adjacent categories in order to satisfy this requirement.

   c. A Score x Group x Response table is constructed. If, in a table, a cell with a frequency of zero is found, the frequencies of all cells are raised by 0,5 to prevent undefined values in the formulas used.

   d. The likelihood ratio chi-square for item m is computed.

2. The t items with the highest chi-square values are classified

as biased, and the other items are included in a set considered to be unbiased.

3. Steps 1 and 2 are repeated, with the algorithm being terminated if, at the end of the iteration, one of the following conditions arises: The prescribed number of iterations has been performed; the maximum chi-square of the set of unbiased items is below the critical value.

In each iteration chi-squares are computed for all items of the test, including the ones classified as biased in the preceding iteration. In each iteration, the number of items eliminated is one greater than the number eliminated in the previous iteration. The items excluded from the computation of the total score in one iteration need not be excluded from the subsequent iteration. If an item that was identified as biased in iteration t has an acceptable chi-square in iteration t' (where t'>t), the item is included in the calculation of the total score for iteration t'+1. In this way the score is iteratively freed from biased items, and items are tested for bias using an apparently unbiased test total as ability indicator.

A computer program to perform the above procedures has been developed by Ader (1982). This program, known as BIASIT, permits a more sophisticated investigation of bias to be undertaken than any other approach, short of ICC-based approaches.

Log-linear or logit bias detection techniques are not the only ones into which iterative procedures can be integrated. All techniques which use test total to estimate ability can be improved through the use of iterative refinement of the total score to reduce or eliminate bias.

3.4 RECOMMENDED STRATEGY FOR BIAS DETECTION

No item bias detection method is infallible; even the ICC-based method cannot be relied upon to identify all items in a test which are biased against a particular group. Two types of error are possible: In a false positive error, an unbiased item is identified as being biased, and in a false negative error, a biased item escapes detection. The relative frequency of these errors is a function of the significance level set for the classification of items as biased or unbiased. For instance, if in the chi-squared method, a chi-square value corresponding to the 5% significance level is set, fewer genuinely biased items may be found than would be the case if a 1% level had been set; on the other hand, fewer false positives will be identified at the 5% level than at the 1% level. Apart from the relative frequency of errors, we must consider the absolute frequency of errors. Some bias detection methods make more errors overall. The absolute number of errors made in comparison with other methods reflects the power of the method. More powerful methods make fewer errors overall.

Just as the probability of one's trousers falling down can be minimized if one uses both braces and a belt and keeps one's hands in one's pockets, so the probability of making errors in identifying item bias can be minimized if one engages in a number of bias detection exercises. This means using more than one sample of data and more than one bias detection method. If a single sample is drawn from each group under study and only one bias detection method is used, then there is a strong likelihood that several false positive and false negative errors will be made. Different bias detection methods seldom correlate in excess of 0,8 with one another; hence, the list of items identified as biased is likely to differ fairly substantially from one bias detection method to another.

If more than one bias detection method is used, then the items can be divided into two classes: Those which all bias detection methods identify as biased or unbiased and those which are identified as biased by some methods and unbiased by other methods. One can have some confidence (but not perfect confidence) that the items falling in the former class have been correctly assessed with regard to bias. On the other hand, if only one bias detection method is used, one cannot have any real confidence about any of one's classifications.

The same reasoning can be used to justify the use of more than one sample of data from each group under study. For a given bias detection method, some items will be identified as biased or unbiased in all samples while the information on others will be mixed. One will have some confidence that an accurate assessment has been made of the items in the former category.

If a multimethod and multisample approach is combined, then a fairly large amount of information is available on each item. For instance, if in a two-group study there are three pairs of samples and two bias detection methods, then each item is evaluated six times. (If each sample is not intended a priori to be paired with only one other sample, then up to 18 bias evaluations can be made on each item.) It is inadvisable to make hard and fast rules about how the results of these analyses should be used, but a reasonable strategy seems to be to treat an item as unbiased unless it has been identified as biased on at least half of the analyses done.

When designing a strategy such as that sketched above, one should try to be as varied as possible in one's selection of bias detection methods and samples. One should avoid using methods which are conceptually similar, because such methods will have similar strengths and weaknesses, and might misclassify items in a similar way. A combination of the iterative logit method and the TID method would satisfy the requirement of variety in the selection of bias detection methods. The logit method is based on a conditional definition of bias whereas the TID method is based on an unconditional definition.

No sample is likely to represent a given group comprehensively; hence one should draw varied samples from the group. If the group is "blacks" one should draw samples from different geographical areas, different occupations and different ethnic backgrounds.

Many bias detection indices have associated tests of significance. These are used to identify the probability that a given value would occur under the hypothesis that the item is not biased. Once the probability falls below a certain value (usually 5% or 1%) the hypothesis of no bias is rejected and the item is identified as biased (in the sample of data under investigation).

Although significance tests are certainly useful in the process of determining which items are biased, they are not the only indicators which should be used. If a 5% significance level is used, then 5% of items will be identified as biased merely due to chance effects. In a 40-item tests, two items would be expected to be erroneously identified as biased due to these effects. Also, if any of the assumptions on which the significance test is based are violated, then the significance values obtained are no longer meaningful.

One of the most important of these assumption is the randomness assumption. Suppose that we draw a large number of samples from a given group (say white apprentices). If every individual in this population has the same probability of being assigned to each of the samples drawn, then the randomness assumption is met. If bias detection procedures are applied to pairs of these samples and a significance level of r% is set, then one would expect r% of the tests for bias to exceed the set level. If the randomness assumption is not met, then there are liable to be more than r% of the bias tests exceeding the r% level, as the samples will be composed of slightly different "kinds" of people. Very easy and very difficult items are liable to be affected more than items in the middle of the difficulty range.

It is for this reason that calibration exercises are often done. The aim is to discount the effects on bias indices of differences in samples due to the violations of the randomness assumption which commonly occur in practice. The usual practice is to draw more than one sample from the "majority" group, or group on which the test was initially developed. One of the samples is regarded as the sample representing the majority group, while the other sample or samples represent a minority group. In some cases, the "pseudo-minority" is made the same size as the typical minority sample. Minority samples may be smaller than majority samples because fewer individuals are available for inclusion in the study. The composition of the pseudo-minority sample may also be "tailored" to represent the ability distribution of the minority group. (In the case of the TID method, however, this step is not recommended for reasons which have been discussed.)

Bias detection methods are then applied to the majority and pseudo-minority sample or samples. A study of this kind gives an indication of what values the indices may assume under the condition where there is no bias but where there are variations in the composition of samples drawn from the same population. These values can be used to set cut-offs on the bias indices.

The sophisticated bias detection methods mentioned above are not the only techniques which should be applied when doing bias research. Several techniques routinely used in the development and refinement of tests also give useful information. In fact, it is important that the test or tests under investigation be subjected to these simpler analyses <u>before</u> the more sophisticated ones are applied. Some of the routine analyses which give useful information on bias are listed below. These should be performed on all samples under investigation.

1. Means, standard deviations, skewnesses, etc.
2. Reliabilities.
3. Proportions of individuals attempting each item.
4. Item difficulties.
5. Item-total correlations.

The results of the analyses mentioned above will indicate whether it is worth continuing with more sophisticated bias investigations. If, for instance, the reliability of a test is very low in a given group, the test is then clearly unusable on that group and it is unnecessary to subject the instrument to an item-by-item bias detection procedure. If there are gross differences between groups in the number of later items attempted, this could indicate that the time limit of the test should be revised before the bias investigation is undertaken.

The basic analyses listed above, therefore, give very useful general information on the test which should be carefully looked at before time and effort is expended on executing the elaborate procedures specifically designed to detect bias.

4.0 GOING BEYOND A PRIORI GROUPS

Up to this point we have been discussing bias against (or in favour of) what might be called a priori groups. Groups of this kind are known in advance of the analysis; the criterion for membership is usually some biographical variable, such as race, SES, area of residence, language, sex, etc. The attention which a priori groups have enjoyed in bias research is largely due to what might loosely be called political factors. An individual has very little influence when representing only himself, but the situation may change radically when he uses his group membership to press his suit. In American law, no-one may be discriminated against on the basis of race, sex, or religious persuasion. These terms define large groupings of people; although the American "system" places great emphasis on the protection of individual rights, in practice this is often achieved through group membership.

In South Africa, groups characterized by biographical variables are perhaps even more starkly distinct than in America, as a result of historical factors and the political dispensation in this country. There is therefore a natural tendency to consider only bias against or in favour of a priori groups when designing research projects.

Although there are good reasons for performing bias research on a priori groups, this should not be the only kind of bias research undertaken. There are certain disadvantages in concentrating entirely on groups composed of members with certain biographical characteristics. One is that it perpetuates a rather simplistic way of thinking of individuals in group terms. A negative consequence of this is that gifted members of low-scoring groups might not be offered the opportunities which they would enjoy if they were members of a higher-scoring group. This situation can occur when separate regression lines are used for different groups, and the regression line of the low-scoring group lies below that of the high-scoring group (see Chemel's, 1985, argument on this point). The use of a priori groups such as race or sex may blind the researcher to the fact that other more important characteristics unite or separate the people under study. An individual classified into a given a priori group may, in fact, have certain features which make him more similar to people in another a priori group.

A case can be made for forming a posteriori groups composed of individuals sharing characteristics which reflect their way of doing the test.

I spoke in the previous section of the need to calibrate bias detection indices in order to take account of fluctuations in the values of these indices which result from slight differences in the composition of samples drawn from the same population. The fact that these fluctuations occur indicates that people from the same a priori group may differ in the way that they do a test. These fluctuations might be due in part to variations in the prevalence of different methods and strategies used by individuals who are all "lumped" into the same group because they share the same value on some gross classifier variable.

In this chapter I shall investigate two approaches which may be regarded as successive steps in the direction of a bias research model based on a posteriori, rather than a priori, groups. The first retains the a priori group concept, but addresses itself to identifying those individuals who possibly do not belong in the group. The second aims at identifying types of response patterns which could be used to define a posteriori groups.

One of the advantages of these approaches is that they take the totality of each individual's responses into account, and therefore can be used to form a picture of his or her overall performance on the test. In conventional bias detection methods on the other hand, each item is usually investigated on its own.

4.1 DEVIANT RESPONSE PATTERNS

Item difficulties are expressed in p values which indicate the proportion of individuals in a given sample who answered the item correctly. We expect an individual, irrespective of his ability level, to have a greater chance of correctly answering an item with a high p value than one with a low p value.

Suppose N people completed a T-item test and we order the data in the following way. The items of the test are ordered across the top of the page from easy (on the left) to difficult (on the right). Subjects are ordered down the page, from high scorers at the top of the page to low scorers at the bottom of the page. The body of the (NxT) matrix is filled with 1's and 0's: A value of 1 in location (i,j) indicates that individual i answered item j correctly and a 0 indicates that he answered it incorrectly.

In such a matrix one would expect 1's to predominate in the upper left area of the matrix (where the responses of "bright" individuals to easy items are to be found) and 0's to predominate in the lower right area (where the responses of "dull" individuals to difficult items are to be found).

In a perfectly ideal and regular situation, one would be able to say which items an individual answered correctly merely by knowing his score on the test in question. If the individual obtained a score of m on the test, it could be concluded that he answered the m items with the highest p values correctly. Similarly, if an item has a p value of p' and N individuals did the item, one would expect the p' x N brightest individuals to have answered the item correctly and the remainder to have answered incorrectly.

Unfortunately, this type of situation almost never happens in practice. There are invariably instances of individuals unexpectedly answering correctly items which appear to be too "difficult" for them, and of individuals failing to answer correctly items which appear to be within their ability. These situations are represented in the matrix as "zeroes amongst the ones" and "ones amongst the zeroes". If items are ordered in difficulty as described above, one expects each row (representing an individual) to start with a series of 1's and conclude with a series of 0's (unless the individual is of very high ability, in which case there may be no 0's). The occurrence of 0's in the main series of 1's indicates items failed which were supposedly within the individual's ability; conversely, the occurrence of 1's within the main series of 0's indicates items passed which were supposedly beyond the individual's ability. Anomalies in the performance of items can be detected in a similar way by investigating the patterns of 1's and 0's in the columns of the matrix.

These concepts are illustrated in the example shown in Table 1. In each row, a vertical line is drawn after the number of elements in that row (counting from the left) which represents the individual's score. (If an individual's score is k, a line is drawn after the kth element in the row.) In the ideal case, all 1's in the rows would be to the left of these lines. Similarly, in each column a horizontal line is drawn after the number of elements which represents the number of individuals who answered the item correctly. (If n individuals answered an item correctly, the line is drawn after the nth element in the column.) In an ideal situation, all 1's in the rows would fall above these vertical lines.

When these sets of vertical and horizontal lines are linked together (as shown in Table 1) they form two "curves" which are known as the S (subject) and P (problem) curves (Tatsuoka & Linn, 1983). The matrix of responses is often known as the S-P table (We have not used this notation in the discussion above because of the possible confusion of P with p and p'.) If the S-curve is left unchanged, but all the 0's to the left of the curve are changed to 1's and all the 1's to the right of the curve are changed to 0's, the resulting S-curve is called a perfect S-curve. In a similar way, a perfect P curve can be generated by swopping 1's and 0's in columns.

Table 1.

A Hypothetical Score Matrix with S-Curve (solid line) and P-Curve (broken line)

	ITEM j											
Subject i	1	2	3	4	5	6	7	8	9	10	Total	p
1	1	1	1	1	1	1	1	1	1	1	10	1,0
2	1	1	1	1	1	1	1	1	1	0	9	0,9
3	1	1	1	1	1	0	1	1	0	1	8	0,8
4	1	0	1	1	1	1	0	1	0	0	6	0,6
5	1	1	1	1	0	1	0	0	1	0	6	0,6
6	1	1	1	0	1	0	1	0	1	0	6	0,6
7	1	1	1	1	0	0	1	0	0	0	5	0,5
8	1	1	1	0	1	1	0	0	0	0	5	0,5
9	1	0	0	1	0	1	0	1	1	0	5	0,5
10	1	1	0	1	0	0	1	0	0	1	5	0,5
11	0	1	1	1	1	0	0	0	0	0	4	0,4
12	1	0	0	0	1	1	0	0	0	1	4	0,4
13	1	1	0	0	0	1	0	0	0	0	3	0,3
14	1	0	1	0	0	0	0	0	0	0	2	0,2
15	0	1	0	0	0	0	0	0	0	0	1	0,1
Total	13	11	10	9	8	8	6	5	5	4		
p*	87	73	67	60	53	53	40	33	33	27		

(*decimals removed)

Indices have been developed to reflect the degree of "anomalousness" of an individual's responses over items or of an item's performance over individuals. The Sato caution index for an individual, i, is the ratio of two covariances. The first term of the ratio is the covariance of the observed row vector i (representing the individual's performance on the items) and the vector of sums of columns (representing the whole sample's performance on the items). The second term of the ratio is the covariance of the same two scores, but where row 1's and 0's have been swopped to form a perfect S-curve (Tatsuoka & Linn, 1983; Tatsuoka, 1984). The value of the second term of the ratio (the denominator) is used as a norm value to standardize the first term (the numerator).

The Sato caution index for an item can be generated in a analogous way to the caution index for an individual.

A modified version of this index has been devised which results in all possible values of the index falling in the interval between 0 and 1 (see Harnisch & Linn, 1981). According to Harnisch and Linn, the modified caution index eliminates extreme values of the index that are sometimes obtained in cases where a very high scoring examinee misses a single very easy item.

The caution index was initially developed to identify individuals or groups who were performing anomalously on school tests. A low value in the caution index indicates that caution should be exercised in the amount of faith placed in the test score. A deviant response pattern is reflected in a low value on the caution index; a low value indicates the possibility that some irregularity occurred during testing (e.g., "cribbing" of certain answers). Miller (1986) used the caution index to identify those school classes where subject matter was taught in an anomalous way. A low caution index can also indicate that the individual in question approached the items in the test in a very different way from the "average" testee. Harnisch (1983) points out that the investigation of patterns of deviant responses can be of diagnostic value. He classifies students both according to test score and score on a deviance index.

The Sato caution index is only one of a number of caution indices which have been proposed. Some indices are based on item response theory (see Drasgow & Guertler, 1987, and Kane & Brennan, 1980), while others like Sato's index are based on manifest scores (see van der Flier, 1977, 1982; Tatsuoka &, 1982). Tatsuoka (1984) and Harnish and Linn (1981) compare some of the indices. Harnisch and Linn intercorrelated deviance scores obtained from eight indices. Most of these are highly intercorrelated with one another. For instance, two of the most commonly used, van der Flier's and Sato's, were correlated 0,96. This is not surprising as they are based on similar principles.

Up to this point, deviant response research and bias research have been following different paths, but there seems to be an application for deviant response indices in bias research. A deviant response index could be used to determine the degree to which "minority" groups respond to a test in a similar way to "majority" groups. It might be possible to partition minority members into a subgroup which responds like majority members and a subgroup which responds in a deviant way relative to the norms of the majority group. Bias research done on the three groups (majority group, non-deviant minority group, and deviant minority group) would indicate whether bias is present for the whole minority group or whether bias is present only in that part of the minority subgroup which responds in a deviant manner. Apart from reducing the problem of bias to a smaller set of people, this approach might also help researchers to understand the nature of bias.

Van den Berg (1985) has written a computer program to detect deviant responses. The program is based on van der Flier's (1977, 1982) index, which is related to the Mann-Whitney U statistic, and is designated U'''. This index is expressed mathematically as follows:

$$U''' = (\log Pmax - \log P(X))/(\log Pmax - \log Pmin),$$

where P(X) is the probability of right-wrong item response pattern X and Pmax and Pmin are the probabilities of (respectively) the least and most deviant pattern yielding the same test score.

It will be remembered from section 3.3 that van der Flier and his co-workers used an iterative method to remove bias from the total score of a test by successively removing the biased items. An analogous approach could be used in deviant response analysis. In this application, the iterative procedure would be directed at individuals rather than items. Any majority group will contain individuals who respond in a deviant fashion. These individuals should be removed in order to obtain a more accurate representation of the pattern of non-deviant responding in the majority group; the purpose of this procedure, then, is to "purify" the sample from the majority group in order to obtain a clear picture of non-deviant responding in this group. The response pattern obtained from the purified sample can then be used as the yardstick against which the response patterns of other groups or individuals can be evaluated.

The "cleaning up" of the majority sample can only be done through the use of iterative methods which remove the "worst offending" individuals on the deviance index, recalculate deviance scores for the remainder of the sample, remove the "worst offending" of those still included in the sample, and so on. As in the van der Flier iterative process, an excluded entity can be readmitted in a subsequent iteration, although this will presumably not happen frequently.

4.2 RESPONSE STYLES

In the previous section we examined the possibility of comparing a group's or individual's pattern of responses with that of the majority group. This approach is heavily centred on the majority group and regards response patterns which differ from that of the majority group as "deviant". A case can be made for this type of approach, because most tests were originally developed on samples drawn from the majority group (in this country, the white population group). If (as could easily be the case), the psychometric properties of a test change as response patterns change, a case can be made for regarding the response pattern of the majority group as canonical and other patterns as deviant.

A different approach is required if the aim is to identify different kinds of response patterns rather than merely to classify them into "normal" and "deviant" categories. Patterns of responses can tell the investigator what styles of information processing an individual or group employed in doing the items of the test (see Taylor, 1987). It is true that conventional tests were not designed to give information on styles of information processing; nevertheless much valuable information can be gathered by investigating patterns of responses.

It was mentioned earlier in this report that researchers often have difficulty in identifying the source of bias in test items. Conventional bias detection techniques concentrate on a single item at a time; these techniques do not use information on response patterns and are therefore "blind" to any information which might be forthcoming from that quarter. The nature of bias might be more understandable and interpretable in a given group if the pattern of responses to the whole test, rather than isolated items, is studied.

Clearly it is not viable or reasonable to accord each pattern of responses the status of a separate style of information processing. For any test more than a few items long, there is an astronomically large number of unique patterns of right and wrong answers. (And if one also distinguishes different types of wrong answers based on which distractors were endorsed, the number becomes considerably larger.)

Substantial numbers of individuals representing each response style are needed in order to do a scientifically respectable evaluation of the implications for bias of the use of different styles.

It is therefore necessary to identify groups of individuals whose styles are similar in many respects.. The most effective way of doing this is to use some cluster analysis procedure, such as

Ward's (1963) method, to identify groups of like-responding testees. Inspection of the protocols of individuals classified into each cluster should enable researchers to determine fundamental characteristics of different response styles. The a posteriori groups formed by the cluster analysis could then be used in a bias analysis. The results from such a study could throw light on the relationship of bias to response styles.

5.0 SUMMARY AND CONCLUSION

In this report I made a case for regarding the detection and elimination of item bias as the most pressing responsibility facing the test constructor at present. Other issues in the domain of comparability and bias should also engage the test constructor's attention; more research will have to be done on construct validity across cultures and groups (i.e., functional equivalence) and some "preliminary" or "illustrative" research into predictive bias will have to be performed. This type of research can only be done effectively, however, once item bias has been removed from tests.

The main purpose of this report is to evaluate item bias detection techniques and suggest a strategy which will minimize the number of items falsely identified as biased and falsely identified as unbiased. All bias detection methods are fallible, just as tests are fallible in predicting performance on a criterion. In both cases we are referring to the validity of the instrument or technique.

Let us pursue this analogy further. One may speak of the validity of a single test, but one may also speak of the validity of a whole selection strategy. This strategy may consist of, inter alia, a biographical questionnaire, tests, an interview, and a decision procedure in which all the above information is integrated and an employment decision made. In this process, tests are embedded in a larger framework, which is the selection strategy. The validity of the selection strategy can be evaluated just as one evaluates the validity of a single test.

An organization which makes use of a fully fledged employment strategy does so in part because it can achieve greater validity than it could through the use of a single test. In a similar way, greater validity can be achieved in item bias detection if a strategy, rather than a single bias detection method, is applied. An examination of the bias detection literature quickly reveals that most researchers are concerned with finding the single best or most valid bias detection method. Although this type of research is important, it is only the first step; the next step is to develop and evaluate bias detection strategies.

In Chapter 3 we proposed a strategy, which may be called a "multimethod-multisample" strategy. We suggested that at least two bias detection methods be used and that these should be based on different sets of assumptions. There are two main classes of bias detection methods -- the conditional and unconditional methods. In the former, group comparisons are made at different ability levels; in the latter, the data is tested for interactions between items and groups. We recommended that a method be taken from each of these two classes. The TID method appears to be the best of the unconditional methods; the iterative logit method the best of the conditional methods (if one takes both practical usability and validity into account).

With regard to sampling, we suggested that several samples be drawn from each group and that these samples vary widely on certain salient characteristics (e.g., occupation, geographical location). It was also suggested that the samples representing the majority group be used for calibration purposes. This procedure enables the researcher to establish bounds within which values of bias indices commonly fall when a test is used on samples drawn from the population for which it was initially developed.

Some standard then needs to be set to guide decisions on the classification of items as biased or unbiased. An item may, for instance, be designated as biased if its bias indices fall outside bounds on more than 50% of comparisons. The number of comparisons is determined by both the number of samples and number of bias detection methods included in the study.

The strategy described above should be undertaken only after certain more basic analyses have been performed on the data. These include: examination of data for evidence of inappropriate test time limits in any of the groups included in the study; comparisons of the means, standard deviations, and skewnesses of the scores in order to evaluate the suitability of the test for application in different groups; and inspection and comparison of various item analysis indices. These straight-forward procedures should be applied before the sophisticated strategy described above is applied; in some cases the simple analyses will establish that a test is unsuitable for use as a selection instrument in a certain group. The need to apply the more sophisticated and time-consuming analyses is then obviated.

All the procedures described above involve the analysis of _data._ Although these analytic procedures are by far the most important in evaluating tests for bias, there is also a place for the actual _inspection_ of items for signs of bias or culture-loadedness. We have called bias which appears to be present on inspection of items _judged bias_ (see Chapter 2). Even when an item is not biased in a statistical sense, it is inadvisable to have it in a test if it _appears_ to be biased. Inspection of items for bias can never replace statistically-based bias detection methods; inspection can, however, be used to impose an additional requirement which each item must satisfy before it is regarded as fit for inclusion in a test.

All bias research reported in the literature to date is based on the comparison of performance in what we have called a priori groups. Assignment of individuals to groups of this kind can be done easily by applying some criterion which is available before the analysis is done. Examples of these criteria are ethnicity, sex, and SES. Bias research on groups of this kind is important, even if only for socio-political reasons. We accept that bias research on a priori groups should have first call on the resources of the research organization, but we do not believe that this is the only type of bias research which should be done.

The problem with bias research on a priori groups is that the results are usually not informative about why bias is present. Items which are judged to be biased often turn out not to be biased on statistical grounds; also, it is often difficult to say why items are biased when they have been identified statistically as biased.

In the opinion of the author bias can only be fully understood once the researcher has an insight into how testees process the information while solving test items. Total scores and individual item scores give very little information on how the testee processed information. Patterns of responses, on the other hand, are more informative. Although not "purpose designed" for the study of processes, response patterns contain useful information on the individual's approach to the tasks in the test. It for this reason that we suggested that bias research be done on a posteriori groups formed on the basis of response patterns.

Two approaches were suggested. In the first, an iterative procedure is used to form a subset of individuals from the majority group who responded in the most "normal" fashion. The responses of this subset are used to calculate item difficulties which in turn are used to calculate the degree of deviance of a response pattern from the "normal" or most non-deviant pattern. A deviance score for each individual, irrespective of his or her a priori group membership, is calculated. If substantially more minority group members have high deviance scores than majority group members, this indicates that many minority group individuals do not approach the problems in the test in the same way that the "normal" individuals in the majority group approach these problems.

A bias study may then be undertaken with the minority sample split into two samples--those with deviant response patterns and those with non-deviant patterns. If bias occurs only against the "deviant" minority group, this means that bias is not associated with minority membership per se, but with certain styles of processing information. Further research into correlates of group membership (deviant vs. normal) might lead to a greater understanding of the nature of bias.

The second approach involves the formation of groups composed of individuals with similar response patterns. Deviance is not used as a criterion for group membership, and there may be more than two groups; the number of groups will reflect the number of distinct response patterns found. This approach can be used to determine whether bias is related to styles of solving problems in tests.

Bias research in a posteriori groups will, at least initially, be more concerned with investigating theoretical issues than with solving practical problems. This type of theoretical research can easily be justified, however, as a greater understanding of bias on the theoretical level will ultimately lead to a greater competence in dealing with this problem in the practical situation.

I mentioned earlier that the use of procedures to detect and remove biased items should be an integral part of test development. The bias detection strategies recommended in this report can be integrated without difficulty into the test construction process. We also mentioned that it is necessary to investigate existing tests for item bias. The procedures to be followed by the test publisher if bias is detected in items of tests already in use are not as clear-cut as those to be followed when bias is detected in items of tests still under development.

In the case of tests still under development, items found to be biased can simply be removed. The test constructor starts with more items than he ultimately needs because he expects attrition to occur, due to the failure of a number of items to satisfy certain psychometric requirements. If the items of a test are to be investigated for bias, this list of requirements is simply lengthened to include those imposed by the bias detection procedures.

In the case of an existing test, the removal of biased items (and possibly their replacement with other items) will inevitably change the test to the extent that it can no longer be regarded as the same test that it was before this "surgery" was undertaken.

What options are open to the test publisher when bias is detected in a test? Some are listed below.

1. Restrict the use of the test to certain applications where the bias problem does not occur.

2. Warn tests users that the test is not score equivalent in different groups. Encourage users who employ the test for selection purposes to take this into account (e.g., by having separate test-criterion regression lines for different groups).

3. Produce a new version of the test complete with norms, recall copies of the old version and replace these with copies of the new version.

4. Produce a new version of the test, "swop over" to selling this version at a certain date, but do not recall and replace old versions of the test which are in use. Send notification of the change to users of the old test.

5. Withdraw the test without replacing it.

I shall not recommend a single way of handling the problem of bias in existing tests. Specific circumstances must be taken into account when deciding what to do.

The investigation of tests for bias is a long procedure, as bias is a multi-faceted concept and many tests have to be subjected to scrutiny. This document does not give guidelines on how to tackle the whole problem, but it is hoped that it helps to clarify the roles and responsibilities of test user and test publisher, and that it offers a useful strategy for investigating item bias.

6.0 REFERENCES

Ader, H J (1982). BIASIT: Iterative program to select biased items. Amsterdam: Subfaculty of Psychology, Vrije Universiteit.

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 (MD): Johns Hopkins University Press.

Angoff, W H & Ford, S F (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-106.

Baker, F B (1981). Log-linear, logit-linear models: A didactic. Journal of Educational Statistics, 6, 75-102.

Baker, F B & Subkoviak, M J (1981). Analysis of test results via log-linear models. Applied Psychological Measurement, 5, 503-515.

Bock, R D (1975). Multivariate statistical methods in behavioral research. New York: McGraw-Hill.

Bock, R D & Aitkin, M (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

Chemel, C S (1985). Fairness in selection with an emphasis on fairness in testing. Unpublished thesis in part fulfillment for MBL. Pretoria: University of South Africa.

Cleary, T A (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-121.

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

Cole, N S (1981). Bias in testing. American Psychologist, 36, 1067-1077.

Drasgow, F (1982). Biased test items and differential validity. Psychological Bulletin, 92, 526-531.

Drasgow, F (1987). Study of the measurement bias of two standardized psychological tests. Journal of Applied Psychology, 72, 19-29.

Drasgow, F & Guertler, E (1987). A decision-theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. Journal of Applied Psychology, 72, 10-18.

Fienberg, S E (1977). The analysis of cross-classified categorical data. Cambridge (Mass): MIT Press.

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

Frederiksen, N (1977). How to tell if a test measures the same thing in different cultures. In Y H Poortinga (Ed.) Basic problems in cross-cultural psychology. Amsterdam: Swets and Zeitlinger.

Harnisch, D L (1983). Item response patterns: Applications for educational practice. Journal of Educational Measurement, 20, 191-206.

Harnisch, D L & Linn, R L (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 18, 133-146.

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

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.

Intasuwan, P (1979). A comparison of three approaches for determining item bias in cross-cultural testing. Unpublished D Phil Dissertation. Pittsburgh: University of Pittsburgh.

Ironson, G H (1982). Use of chi-square and latent trait approaches for detecting item bias. In R Berk (Ed.) Handbook of methods for detecting test bias. Baltimore (MD): 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.

Ironson, G, Homan, S, Willis, R, & Signer, B (1984). The validity of item bias techniques with math word problems. Applied Psychological Measurement, 8, 391-396.

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

Kane, M T & Brennan, R L (1980). Agreement coefficients as indices of dependability for domain-referenced tests. Applied Psychological Measurement, 4, 105-126.

Kok, F G, Mellenbergh, G J, & van der Flier, H (1985). Detecting experimentally induced item bias using the iterative logit method. Journal of Educational Measurement, 22, 295-303.

Lautenschlager, G J & Mendoza, J L (1986). A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. Applied Psychological Measurement, 10, 133-139.

Linn, R L (1984). Selection bias: Multiple meanings. Journal of Educational Measurement, 21, 33-47.

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

Lord, F M (1977). A study of item bias, using item characteristic curve theory. In Y H Poortinga (Ed.) Basic problems in cross-cultural psychology. Amsterdam: Swets and Zeitlinger.

Lord, F M (1980). Applications of item response theory to practical testing problems. Hillsdale (NJ): Lawrence Erlbaum.

Mc Cauley, C D & Mendoza, J (1985). A simulation study of item bias using a two-parameter item response model. Applied Psychological Measurement, 9, 389-400.

Mellenbergh, G J (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118.

Miller, M D (1986). Time allocation and patterns of item response. Journal of Educational Measurement, 23, 147-156.

Oosterhof, A C, Atash, M N & Lassiter, K L (1984). Facilitating identification of item bias through use of Delta plots. Educational and Psychological Measurement, 44, 619-627.

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

Owen, K (1986). Toets- en itemsydigheid: Toepassing van die Senior Aanlegtoetse, Meganiese Insigtoets, en Skolastiese Bekwaamheidsbattery op blanke, swart, kleurling- en Indier technikonstudente. HSRC Report P-66. Pretoria: Human Sciences Research Council.

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

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

Plake, B S (1981). An ANOVA methodology to identify biased test items that takes instructional level into account. Educational and Psychological Measurement, 41, 365-368.

Poortinga, Y H (1971). Cross-cultural comparison of maximum performance tests: Some methodological aspects and some experiments with simple auditory and visual stimuli. Psychologia Africana, Monograph Supplement No. 6.

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

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

Rigdon, S E & Tsutakawa, R K (1983). Parameter estimation in latent trait models. Psychometrika, 48, 567-574.

Rudner, L M (1983). Individual assessment accuracy. Journal of Educational Measurement, 20, 207-219.

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

Scheuneman, J D (1979). A 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 Th van der Kamp, W F Langerak and D N M Gruijter (Eds.) Psychometrics for educational debates. Chichester: John Wiley and Sons.

Shepard, L A, Camilli, G & Williams, D M (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

Shepard, L A, Camilli, G & Williams, D M (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Stricker, L J (1982). Identifying test items that perform differently 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.

Tatsuoka, K K (1984). Caution indices based on item response theory. Psychometrika, 49, 95-110.

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

Tatsuoka, K K & Tatsuoka, M M (1982). Detection of aberrant response patterns and their effect on dimensionality. Journal of Educational Statistics, 7, 215-231.

Taylor, J M (1986). Psychometric testing: An unfair labour practice? Paper delivered at the IPM International Convention, Johannesburg. R/Pers 742.

Taylor, T R (1987). The future of cognitive assessment. HSRC Report Pers-420. Pretoria: Human Sciences Research Council.

differences: The concept of item bias. Psychological Bulletin, 99, 118-128.

Van den Berg, A R (1985). An investigation into the power of two psychometric models to detect testees who have deviant item response patterns. HSRC Report P-51. Pretoria: Human Sciences Research Council.

Van der Flier, H (1977). Environmental factors and deviant response patterns. In Y H Poortinga (Ed.) Basic problems in cross-cultural psychology. Amsterdam: Swets and Zeitlinger.

Van der Flier, H (1982). Deviant response patterns and comparability of test scores. Journal of Cross-Cultural Psychology, 13, 267-298.

Van der Flier, H & Drenth, P J D (1980). Fair selection and comparability of test scores. In L J Th van der Kamp, W F Langerak,and D N M Gruijter (Eds.) Psychometrics for educational debates. Chichester: John Wiley.

Van der Flier, H, Mellenbergh, G J & Ader, H J (1984). Een Onderzoek naar de effectiviteit van een iteratieve item bias detectie methode bij groupen met een verschillend treknivo. Tijdschrift voor Onderwijsresearch, 9, 61-70.

Van der Flier, H, Mellenbergh, G J, Ader, H J, & Wijn, M (1984). An iterative item bias detection method. Journal of Educational Measurement, 21, 131-145.

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

Ward, J H (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58, 236-244.

Zeidner, M (1987). Test of the cultural bias hypothesis: Some Israeli findings. Journal of Applied Psychology, 72, 38-48.

HSRC PUBLICATION LIST

A complete list of HSRC publications or a list of publications of a particular Institute of the HSRC is available from the Head: Technical Services

HUMAN SCIENCES RESEARCH COUNCIL
RAAD VIR GEESTESWETENSKAPLIKE NAVORSING

President	Dr J.G. Garbers	President
Deputy Presidents	Dr H.C. Marais, Dr J.D. Venter	Adjunk-presidente
Vice-Presidents	Dr K.F. Mauer, Prof. D.J. Stoker	Vise-presidente
Executive Director: Administration	J.G.G. Gräbe	Uitvoerende Direkteur: Administrasie
Chief PRO	Dr G. Puth	Skakelhoof

Functions of the HSRC

The HSRC undertakes, promotes, supports and co-ordinates research in the field of the human sciences. It also determines research priorities, disseminates the findings of human sciences research, facilitates and evaluates the implementation of research findings, stimulates the training of researchers, places the full spectrum of human sciences disciplines at the service of the inhabitants of the RSA and promotes science in general.

Institutes

Institute for Communication Research (ICOMM)

Institute for Educational Research (IER)

Institute for Historical Research (IHR)

Institute for Manpower Research (IMAN)

National Institute for Personnel Research (NIPR)

Institute for Psychological and Edumetric Research (IPER)

Institute for Research Development (IRD)

Institute for Research into Language and The Arts (IRLA)

Institute for Sociological and Demographic Research (ISODEM)

Institute for Statistical Research (ISR)

Bureau for Research Support Services (BRSS)

Administration

Head office

Private Bag X41, Pretoria 0001
Republic of South Africa
Telegrams RAGEN
Tel. (012) 202-9111
Telex 3-20893 SA

NIPR

P.O. Box 32410, Braamfontein 2017
Republic of South Africa
Telegrams NAVORSPERS
Tel. (011) 339-4451
Telex 4-25459 SA

Regional offices

Western Cape, Private Bag X5, Roggebaai 8012
Tel. (021) 419-2572/3/4/5 Telex 5-22260 SA

Natal, P.O. Box 17302, Congella 4013
Tel. (031) 815970 Telex 6-28567 SA

NIPR Eastern Cape, P.O. Box 1124, Port Elizabeth 6000
Tel. (041) 53-2131 Telex 2-43203 SA

Funksies van die RGN

Die RGN onderneem, bevorder, ondersteun en koördineer navorsing op die gebied van die geesteswetenskappe, bepaal navorsingsprioriteite, versprei die resultate van geesteswetenskaplike navorsing, vergemaklik en evalueer die implementering van die resultate van navorsing, stimuleer die opleiding van navorsers, stel die volle spektrum van dissiplines in die geesteswetenskappe ten diens van die inwoners van die RSA en bevorder die wetenskap in die breë.

Institute

Instituut vir Geskiedenisnavorsing (IGN)

Instituut vir Kommunikasienavorsing (IKOMM)

Instituut vir Mannekragnavorsing (IMAN)

Instituut vir Navorsingsontwikkeling (INO)

Instituut vir Opvoedkundige Navorsing (ION)

Nasionale Instituut vir Personeelnavorsing (NIPN)

Instituut vir Psigologiese en Edumetriese Navorsing (IPEN)

Instituut vir Sosiologiese en Demografiese Navorsing (ISODEM)

Instituut vir Statistiese Navorsing (ISN)

Instituut vir Taal- en Kunstenavorsing (INTAK)

Buro vir Ondersteunende Navorsingsdienste (BOND)

Administrasie

Hoofkantoor

Privaatsak X41, Pretoria 0001
Republiek van Suid-Afrika
Telegramme RAGEN
Tel. (012) 202-9111
Teleks 3-20893 SA

NIPN

Posbus 32410, Braamfontein 2017
Republiek van Suid-Afrika
Telegramme NAVORSPERS
Tel. (011) 339-4451
Teleks 4-25459 SA

Streekkantore

Wes-Kaap, Privaatsak X5, Roggebaai 8012
Tel. (021) 419-2572/3/4/5 Teleks 5-22260 SA

Natal, Posbus 17302, Congella 4013
Tel. (031) 815970 Teleks 6-28567 SA

NIPN Oos-Kaap, Posbus 1124, Port Elizabeth 6000
Tel. (041) 53-2131 Teleks 2-43203 SA

ISBN 0 7969 0551 7

1560*9/87