SA Journal of Industrial Psychology, 2004, 30 (4), 44-51
SA Tydskrif vir Bedryfsielkunde, 2004, 30 (4), 44-51

STATISTICAL AND EXTRA-STATISTICAL CONSIDERATIONS IN DIFFERENTIAL ITEM FUNCTIONING ANALYSES

G K HUYSAMEN
Department of Psychology
University of the Free State

ABSTRACT

This article briefly describes the main procedures for performing differential item functioning (DIF) analyses and points out some of the statistical and extra-statistical implications of these methods. Research findings on the sources of DIF, including those associated with translated tests, are reviewed. As DIF analyses are oblivious of correlations between a test and relevant criteria, the elimination of differentially functioning items does not necessarily improve predictive validity or reduce any predictive bias. The implications of the results of past DIF research for test development in the multi-lingual and multi-cultural South African society are considered.

OPSOMMING

Hierdie artikel beskryf kortliks die hoofprosedures vir die ontleding van differensiële itemfunksionering (DIF) en verwys na sommige van die statistiese en buite-statistiese implikasies van hierdie metodes. 'n Oorsig word verskaf van navorsingsbevindings oor die bronne van DIF, insluitend dié by vertaalde toetse. Omdat DIF-ontledings nie die korrelasies tussen 'n toets en relevante kriteria in ag neem nie, sal die verwydering van differensieel-funksionerende items nie noodwendig voorspellingsgeldigheid verbeter of voorspellingsydigheid verminder nie. Die implikasies van vorige DIF-navorsingsbevindings vir toetsontwikkeling in die veeltalige en multikulturele Suid-Afrikaanse gemeenskap word oorweeg.

The concept of item bias was introduced to the psychometric world in the United States in the 1960s. A.P. Schmitt (1988) cites a research bulletin of the Educational Testing Services (ETS) which shows that this organisation has performed such analyses on their Scholastic Aptitude Test (SAT) since 1964. Nowadays major test publishers in the United States of America routinely perform DIF analyses as part of their test development process (Cole & Zieky, 2001). Originally, interest in such analyses was prompted by the concern that cognitive-ability tests may discriminate against African-American and Hispanic examinees on account of their different cultural backgrounds (Angoff, 1993). Since then, the scope of such analyses has broadened and the term *item bias* has been superseded by *differential item functioning* (DIF). DIF is said to occur if different subgroups, who are of equal standing on the construct the test is designed to measure, display different probabilities of passing an item (in tests of maximal performance) or of endorsing an item (in tests of typical performance). In other words, DIF occurs when an item is not equally difficult (in maximal performance tests) or equally popular (in typical performance tests) for groups that have been matched in terms of the construct being measured. Suppose an item in a mathematical aptitude test requires examinees to identify which of two rugby teams has won a match if the one team scored only two converted tries and the other scored only four penalties. Differential item functioning will occur if such an item is passed statistically significantly more often by men than by women of the same mathematical aptitude.

DIF should be distinguished from differences in the mean item performances of entire groups. If such a difference is due solely to a difference in the construct being measured, we are dealing with adverse impact (Ackerman, 1992). DIF analyses involve a comparison of the performances of subgroups that have been matched in terms of the relevant construct and hence do not require equal (total) test scores for the groups involved. (However, Hambleton, Clauser, Mazor & Jones [1993, p. 16] pointed out that if there are large differences in group means, the identification of DIF has a greater probability of being confounded with Type I Errors.)

Following the terminology proposed by Roussos and Stout (1996), the term dimension will be used here to denote any aspect of an item that can affect the probability of passing or endorsing the item. For almost the past quarter of a century it has been recognised that DIF is caused by multi-dimensionality in an item (Linn, Levine, Hastings & Wardrop, 1980), that is, that performance on items depends not only on the construct that the test is designed to measure, referred to as the *primary* (or target) dimension, but also on one or more other dimensions, known as *secondary* dimensions. DIF comes about when different subgroups that are matched in terms of the primary dimension, differ in their standing on a secondary dimension, such as familiarity with the content in terms of which the items are formulated. In the example used earlier, mathematical aptitude was the primary dimension, and familiarity with the point-scoring system in rugby represented a secondary dimension. Items that measure the relevant construct devoid of any content are extremely rare. (A mathematics test that contains only completely abstract tasks such as those involving factoring or exponents would represent some of the rare examples in this category.) DIF analyses are directed at identifying items that are affected by such secondary dimensions or, stated differently, that measure different, additional aspects in different subgroups. As such, DIF analyses are neutral in terms of the sources of DIF so that the term *item bias* with its suggestion of a discriminatory origin has become inappropriate. Nowadays this term is invoked only if items have been identified, by statistical means, as differentially functioning, *and* if the reason for this can be attributed to construct-irrelevant properties of the item (Penfield & Lam, 2000). DIF may be said to be a necessary but not sufficient condition for item bias.

The purpose of this article is to present the main statistical strategies underlying DIF analyses, the implications of these methods, some research findings concerning the sources of DIF, and the distinction between DIF, differential validity and differential prediction. Finally, the implications of some earlier research findings for South African testing practices will be pointed out.

THE METHODOLOGY OF DIF IDENTIFICATION

In the DIF literature it has become customary to distinguish between the *reference group* whose performance serves as the basis of comparison, and the *focal group* whose performance is being compared against that of the reference group. Typically the distinction between these groups has been based on

Requests for copies should be addressed to: GK Huysamen, Department of Psychology, University of the Free State, PO Box 339, Bloemfontein, 9300

demographic variables such as sex (males vs. females) or culture/ethnicity (whites vs. blacks). In the case of tests translated from one language into another, the group-membership variable of interest is language group. Due to the development of DIF analyses in the United States of America, the reference group typically was taken to be the majority group, or the group for which the test was originally intended. However, the designation of one group as the reference group and the other as the focal group is psychometrically an arbitrary decision. More properly, either group may serve as a point of reference for the other.

Clauser and Mazor (1998) and Millsap and Everson (1993) presented various statistical procedures for the analysis of DIF. Penfield and Lam (2000) reviewed DIF procedures for polytomous items. According to Cook, Schmitt and Brown (1999) the most popular methods for investigating DIF are Dorans and Kulicks's standardisation procedure, the Mantel-Haenszel (MH) procedure, the logistic regression approach and Shealy and Stout's Simultaneous Item Bias (SIBTEST) procedure. Of these methods, the Holland and Thayer (1988) adaptation of the MH procedure has generally been regarded as the most popular with dichotomous items, whereas procedures based on logistic regression are recommended in the case of polytomous items (Sireci & Geisinger, 1998). To provide a clearer conceptualisation of DIF, the MH procedure and those based on item response theory (IRT) and logistic regression, will be described briefly. Finally, in view of its useful explanation of the occurrence of DIF, Roussos and Stout's (1996) elaboration of Shealy and Stout's (1993a) multi-dimensional IRT model will be introduced.

The Mantel-Haenszel chi-square procedure

The MH procedure is an extension of the traditional two-way chi-square test of independence (between two variables) to the situation in which three variables are completely crossed, namely, group membership (e.g., men or women; black or white examinees, etc.), performance on the item (e.g., correct or incorrect) and any number of levels of the attribute the test is designed to measure. The latter variable is also known as the matching or stratifying variable as examinees (or respondents) from the reference and focal groups are matched or stratified in terms of it. The null hypothesis tested by means of the MH procedure then is that the probability of answering the item correctly over all levels of the matching variable is the same for the reference and focal groups being studied. If one group shows a higher probability of passing or endorsing any particular item than does another group, the item is said to be functioning differentially (with regard to the group-membership variable).

This procedure proceeds by partitioning the two groups into several (discrete) subgroups in terms of the matching variable (so that those in any particular subgroup have more or less the same score on that variable). In lieu of a more appropriate index of the construct to be measured, the total score on the test is typically used. If different subsections of a test are directed at quite independent skills or aptitudes, it may be advisable to perform different DIF analyses for the different subsections, with each subsection's total score being used as the matching variable for that particular subsection. Next, the procedure calculates for every item the ratio of the odds of success of, say, the reference group over that of the focal group, over all of the subgroups representing the various levels of the matching variable. Finally, it averages these ratios across these subgroups and transforms this average ratio into a DIF index with values ranging from negative infinity to positive infinity. A positive DIF value identifies DIF that favours the reference group; a DIF value of zero represents absence of DIF; and a negative DIF value reveals DIF that benefits the focal group. Items with sizeable DIF that are regarded as unfairly related to group membership are typically eliminated from the total score and the analyses are repeated with the new "purified" total score (Zieky, 1993) as matching variable.

Procedures based on item response theory

In terms of item response theory (IRT), an examinee's probability of passing or endorsing an item is regarded as a function of his or her standing on the construct being measured where the latter is viewed as a continuous variable. The graph of this function, which is plotted by using the values of the construct, θ, being measured as the *X* axis and the probability of passing or endorsing the item as the *Y* axis, has the form of a horizontally stretched-out S. (Cf. Figures 2 and 3 in the article on DIF analyses of the Learning Potential Computerised Adaptive Test in this issue.) This means that as scores on the construct being measured become higher, the probability of passing or endorsing the item initially rises at an increasingly steeper rate up to the point of inflection after which it tapers off again at an increasingly faster rate. The resulting monotonically increasing graph is known as an item characteristic curve (ICC). Several models that vary in terms of their underlying assumptions and the number of parameters that are required to define the curve are available to represent this function. In the two-parameter ICC, the discrimination parameter, *a*, is proportional to the slope of the curve at its point of inflection, and the difficulty, or preference, parameter, *b*, is the value on the *X* axis at the point of inflection. If the ICCs for an item are different for two groups, the item is regarded as functioning differentially for such groups. In the case of uniform DIF, the curve of the one group consistently falls higher (in terms of the Y axis) than that of the other, suggesting that the probability of passing or endorsing the item is uniformly higher for one group than for the other. In the case of non-uniform DIF, the curves cross at some point, implying that the item is more discriminating for one group than for the other. (Figures 2 and 3 in the article referred to above are examples of uniform and non-uniform DIF, respectively.)

There are several procedures for estimating the θ value for each examinee and the item parameters *b* and *a* from examinees' responses to the test items. These include joint, marginal and conditional maximum likelihood and joint and marginal Bayesian estimation procedures (Hambleton, Swaminathon & Rogers, 1991). The test whether the ICCs for two groups are the same involves, for example, Lord's chi-square test of the hypothesis that the *b*s of the ICCs for two groups are the same and that the *a*s for them are the same. Another approach focuses on the size of the area that falls between the two curves, if different.

Logistic regression

In terms of this approach, a logistic regression equation is established for each group separately. In terms of this equation, performance on the studied item is predicted in terms of the (continuous) total score, group membership and the interaction between these two predictors. DIF is absent if the logistic regression equations for the two groups are the same. If the terms for group membership are different (so that the intercepts of the corresponding regression lines are different), uniform DIF may be inferred. If the interaction terms differ (so that the slopes are different), non-uniform DIF may be present. According to Clauser and Mazor (1998), empirical studies have shown that this procedure is comparable to the MH procedure for testing uniform DIF but that it is superior to MH for detecting non-uniform DIF.

Shealy and Stout's SIBTEST procedure and model for studying DIF

Shealy and Stout's (1993a) multi-dimensional model forms the basis of their SIBTEST procedure. This model allows one to formulate hypotheses, based on previous research findings or theoretical or substantive considerations, about items or bundles of items that are likely to display DIF. Such an item bundle that is hypothesised to measure the primary as well as some secondary dimension, is referred to as the studied subtest, whereas the matching subtest contains the items considered to measure only the primary dimension. The hypotheses are tested statistically (by means of these authors' SIBTEST procedure) and the obtained results may provide feedback for the test development process. For example, content dimensions with large DIF (say knowledge of the game of rugby in a mathematical aptitude test used with both men and women) may be replaced by equally valid content dimensions that have been shown to be less likely to produce DIF.

The present model assumes that the primary dimension, θ, and the secondary dimension, η, are bivariate normally distributed, and it also requires the weak assumptions of equal standard deviations and correlations (between the two dimensions) for the reference (R) and focal (F) groups. Under these assumptions, Roussos and Stout (1996), using conventional multivariate statistical methodology, showed that the expected difference in the means of the secondary dimension for individuals from the reference and focal groups with a fixed value on the primary dimension is equal to:

(1) $$E_R(\eta /\theta) - E_F(\eta/\theta) = (\bar{\eta}_R - \bar{\eta}_F) - \rho(\bar{\theta}_R - \bar{\theta}_F),$$

where the bar denotes mean values. This equation states that the probability of DIF occurring is equal to the difference between the two population means on the secondary dimension minus the corresponding difference between the two population means on the primary dimension where the latter difference is multiplied by the correlation, ρ, between these two dimensions. (In the case of cognitive abilities, the primary and secondary dimensions are typically correlated positively.)

In terms of this model, the occurrence of DIF depends on the size and sign of the two subgroup differences on the primary and secondary dimensions respectively, and on the correlation between these two dimensions. More specifically, this model shows that DIF is more likely to occur, firstly (Roussos & Stout's Case A), if these two subgroup differences are of opposite signs (because the minus sign between them prevents them from cancelling each other). Secondly (Roussos & Stout's Case B), DIF is likely to arise if the reference and focal groups are equally strong in terms of the primary dimension (so that the second subgroup difference is zero) but differ on the secondary dimension (implying a nonzero value for the first subgroup difference and hence for the grand difference). In the next section it will be shown how this model provides a useful explanation of some ostensibly paradoxical DIF results.

General features of the above procedures

As in all hypothesis testing, the smaller the size of DIF, the larger the sample size required for rejecting the null hypothesis of no DIF, i.e., of detecting DIF. By the same token, whereas the amount of DIF in any single item may be too small to be detected statistically, the amount aggregated in a bundle of such items (measuring a common secondary dimension) may have a greater probability (power) of being detected statistically. Shealy and Stout's (1993a) SIBTEST procedure takes advantage of this principle as it allows for the investigation of differential bundle functioning (DBF), that is, the different probabilities of passing (or endorsing) sets of several items.

According to Clauser and Mazor (1998), samples of 200 to 250 per group have been shown to be sufficient for use with the MH, logistic regression and SIBTEST procedures, and that larger samples are required for the two- and three-parameter IRT models. Due to their iterative nature, the IRT-based procedures require substantially more computer time than does, for example, the MH procedure. Because the DIF detection procedures typically assume that a single primary dimension is being measured, several authors have recommended that the dimensional structure of a test be investigated prior to any DIF analyses. Various sources, for example, DeAyala and Hertzog (1991) and Nandakumar (1994) have reviewed such methods of dimensionality analyses.

RESEARCH INTO THE SOURCES OF DIF

Research into the sources of DIF may proceed at two different levels. On one level, research may be directed at identifying *which* kinds of content are more likely than others to give rise to DIF. At a different level, the question is pursued *why* certain kinds of content are more likely than others to result in DIF. Only one study, namely that of Stricker and Emmerich (1999), could be identified that deals with the second question. Although this study postdates most of the research on the former question, it will be covered first as its findings present a convenient introduction to the discussion of the findings on the former topic. Stricker and Emmerich had each of the 1000 multiple-choice items of the Advanced Placement Psychology Examination rated on familiarity, appeal and unpleasantness by samples totalling 265 boys and 452 girls. For every item a *d* value, i.e., the standardised mean difference between the boys' and girls' ratings on each of the three variables was obtained. These *d* values were correlated with the items' MH DIF values as determined in the regular American College Board testing programme. All the correlations were statistically significant. The correlation between the items'*d* values and their familiarity, appeal and unpleasantness ratings were 0,24, 0,39 and -0,37 respectively, all of which qualify as effect sizes of medium magnitude in terms of Cohen's (1988) proposed classification. This means that the more familiar or the more interesting the items were to girls (as opposed to boys), the greater their DIF in favour of girls tended to be, whereas the more unpleasant they were perceived to be by the girls, the greater their DIF to the disadvantage of the girls tended to be. These findings suggest that DIF may arise not only because of examinees' differences in their familiarity with item content but also their affective responses towards items (their interest in item content and their experience of items as being offensive).

Studies in which substantive experts and psychometricians attempted to identify the kinds of content that are likely to function differentially and studies in which statistical procedures (such as those discussed in the preceding section) were used to identify such items have tended to yield contradictory results. Bond (1993) facetiously recalled how he and a student had come up with (to them) convincing reasons why a set of flagged items were functioning differentially, only to be informed later that the wrong set of items had been so identified. However, as will be shown below, some progress has been made on this score in the meantime.

O'Neill and McPeek (1993) analysed the items that were identified as differentially functioning in two or more cognitive-ability tests to check whether there were any consistencies in their content. In reading comprehension passages, women, as compared with matched groups of men, performed less well on items cast in a science-related context but better on items formulated in terms of content associated with the social sciences and humanities. This finding is an example of Roussos and Stout's (1996) Case A (in which the subgroup differences in the primary and secondary dimensions are of opposite signs). Women are superior in terms of reading comprehension (the primary dimension) but they perform more poorly than men in the reading passages that contain technical aspects of science (secondary dimension). In this case, the first difference in Equation 1 favours men but the second difference benefits women so that the (positive) grand difference between these two differences represents DIF in favour of men.

As far as discrete, verbal items that are not based on reading passages are concerned, women performed better on antonyms and analogies based on aesthetics/philosophy or human relationships, but less well if the antonyms and analogies were dealing with science or practical affairs. Another consistent finding was that women performed better on algebra items than did a group of matched men, but more poorly than such men on geometry and mathematical problem-solving items.

In comparisons between matched groups of African American and white examinees, the former group performed better on analogy items if the context was human relationships but less well when the context was science. The former, paradoxical finding is an example of Roussos and Stout's (1996) Case B (where some content, that in terms of conventional wisdom should not favour any particular group, favours the very group that is known to be performing poorly in terms of the primary dimension). It is well-known that African Americans score lower than whites on the SAT verbal section (primary dimension). However, one would have expected whites and blacks to be of comparable standing in terms of analogy and antonym items dealing with human relationships (secondary dimension). In terms of Roussos and Stout's multi-dimensional model (Equation 1 above), the two populations have approximately equal means on the secondary dimension (resulting in the first subgroup difference being zero), but the reference group is known to excel on the primary dimension (so that the second subgroup difference is positive). As a result, the grand difference between the two subgroup differences is negative, which translates into the unexpected result of DIF favouring blacks.

If examinees complete a test in a language other than their home language or the language in which they are most proficient, or if a test is translated from one language (known as the source language) into another (the target language), there is a clear need for DIF analyses. Firstly, there is the possibility of phenomena such as homographs (words that are spelled alike but have different meanings in different languages such as the word *spring* in Afrikaans and English). The presence of such words may change the difficulty of the item for different language groups. In the case of translated tests there is also a distinct possibility that some words in the source language may have no counterparts with precisely the same meaning in the target language. Even if the original item and its translated version are equivalent in linguistic meaning, there may be cultural differences in different language groups' reactions to it. As a result, it comes as no surprise that the proportion of items exhibiting DIF in translated tests is much higher than in the case of same-language tests used with men and women, or with white and black examinees. Gierl and Khaliq (2001) refer to examples where up to 52 % of the items of translated cognitive tests displayed DIF. (Elaborate guidelines for the adaptation [the term preferred to *translation*] of tests from the source language for use in target languages are provided by Hambleton [1994].)

In one study, A.P. Schmitt (1988) used the DIF results found by means of the standardisation procedure for the SAT results of 278 166 white and 6 193 Hispanic students to formulate hypotheses about the sources of DIF. Among others, she hypothesised that items containing true cognates (words with a common root in English and Spanish such as *music* in English and *musica* in Spanish) would show DIF in favour of Hispanic examinees. In a subsequent study (with a different group of 285 885 white and 6 840 Hispanic examinees and a different form of the SAT) this hypothesis, among others, was confirmed. In the same pair of studies she also found that content of special interest to Hispanics, such as a reading passage about Mexican-American women, favoured this group.

Ellis (1989) studied DIF in a German translation of a 251-item American intelligence test (the Career Ability Placement Test) and a 145-item German intelligence test (the WILDE Intelligenz Test). The participants were 205 German high school graduates and 217 American first- and second-year university students. Although she found only eight differentially functioning items, the explanations she advanced for their occurrence were informative. Three of the items of the American test were more difficult for the German participants than for the American participants; three items of the same test and two of the German test were more discriminating for the German participants than for their American counterparts; and two items from the American test were both more discriminating and easier for the American examinees than for the German examinees. Ellis concluded that in the majority of cases, DIF was due to translation problems. For example, in two of the items the comparative of the English word *heavy*, namely *heavier*, was involved. Although in both English and German the comparative is formed by adding *er*, the intricacies of the German language (masculine nouns, nominative cases) require the comparative form (of *schwer*) to be *schwererer* (rather than *schwerer*). Although *schwererer* may be the linguistically correct form, it is uncommon in spoken German and hence the items using this term were more difficult for the German examinees than for their American counterparts taking the test in English. In other cases differences in cultural knowledge or experience was thought to explain the occurrence of DIF. For example, one item required examinees to view poodles and retrievers as different races of dogs – a distinction that the American examinees indeed made. However, German examinees were more likely to fail to make this distinction because they were familiar with the poodle's historical background as a waterfowl *retriever* in Germany.

Allalouf, Hambleton and Sireci (1999) investigated the sources of DIF in three forms of the original Hebrew version of the Israeli Psychometric Entrance Test (an Israeli university admissions test) and a Russian translation of it. Each of the three test forms was completed by between 5 837 and 7 150 Hebrew examinees and its translated version by between 1 485 and 2 033 Russian examinees. Five Hebrew-Russian translators examined the items with a view to suggesting which of them were likely to function differentially for these two groups, which language group would be favoured by them, and what the explanation for such DIF might be. Afterwards the five translators and three Hebrew-speaking researchers compared the results obtained with the statistically (MH) identified DIF results and came up with four sources of translation DIF, namely, translating a Hebrew word or sentence into a Russian word or sentence that differs in difficulty from the Hebrew one, changes in content (e.g., the translation of a word that has a single meaning in Hebrew into a word that has more than one meaning in Russian), changes in format (e.g., the translated sentence is longer than the sentence in Hebrew), and differences in cultural relevance (e.g., the situation is more relevant or more familiar to one language group than to the other).

Gierl and Khaliq (2001) followed the same approach to identify sources of translation DIF in the English and French versions of the Mathematics and Social Studies Achievement Test for Grade 6 and those for Grade 9 pupils in Alberta, Canada. Using the differentially functioning items from the 1996 administration of this test (3 000 English and 2 115 French students per grade), an 11-member committee of testing specialists arrived at four such sources that showed some resemblance to those identified in the Allalouf, Hambleton and Sireci (1999) study. One, namely, changes in format, was the same in both studies. Gierl and Khaliq went a step further by requesting two certified translators to identify item bundles from the 1997 administration of the tests that were most likely to be representative of these sources of translation DIF. This meant that there were 16 possible DIF categories (4 potential sources of translation DIF x 2 language groups that could be favoured by them x 2 grade levels) into which suspect items could be sorted. Eight and 13 item bundles were sorted in the case of the Mathematics and Social Sciences tests, respectively. Next, each of these item bundles was used as the studied item bundle for the application of the SIBTEST procedure. The translators correctly predicted the group that would be favoured for seven of the eight Mathematics item bundles and for eight of the 13 Social Sciences item bundles. The sources of translation DIF in the majority of these cases involved differences in the words, expressions or sentence structures of items that, according to the authors, could be overcome by proper translation.

DIF is not restricted to tests of maximal performance. Familiarity with the content of an item in a test of typical performance may also give rise to DIF. For example, an item in an anxiety questionnaire that deals with the respondent's reaction to a late-night phone call draws on the experiences of households that have access to telephones in their homes. Moreover, differences in attitudes or opinions may lead to DIF. Ellis and Kimmel (1992) investigated DIF in English (N = 98 Americans), German (N = 205) and French (N = 191) versions of a 186-item Likert-type scale to measure attitudes toward mental health. The ICC for each group was compared with that for the total, combined group. Four items exhibited DIF in the comparison between the American group and the total group, and two in each of the corresponding comparisons involving the French and the German groups. For example, in comparison with the combined group, Americans were more likely to agree (with the statement) that mentally healthy persons inherently trust human nature, Germans were less likely to agree that mentally healthy individuals were known for bringing up their children to become well-adjusted citizens, and the French were more likely to agree that mentally healthy persons needed several good friends to feel happy.

THE DIF PREVENTATIVE ROLE OF EXPERT JUDGEMENT IN TEST DEVELOPMENT

From the section on the statistical identification of DIF it follows that the basis of comparison for determining DIF is performance on the total test or on the remaining items in a test, or some function of it (as in IRT procedures). Because this is a criterion that is internal to the test, only differential functioning relative to the latter items, rather than "absolute" unfairness (Cole & Zieky, 2001) can be identified. If all items in a test are judged to be equally unfair in terms of some or other extra-statistical consideration (e.g., differences in familiarity, appeal or offensiveness), none of them will be identified as functioning differentially. If all the items in a cognitive-ability test are formulated in terms of the rules of rugby, men are likely to have a higher mean on the total test than women. If men maintain this advantage to the same extent on each and every item, none of the items will be identified as functioning differentially for men and women. As a matter of fact, Roussos and Stout's (1996) Case B referred to earlier suggest that if in such a situation there are a few items that favour neither men nor women, these items may be flagged as functioning differentially to the advantage of women.

To prevent such pervasive bias from occurring, test developers have to be familiar with the research findings on the contents that favour different groups. Some contexts may give an advantage to men, whereas others may benefit women; some may favour white examinees, whereas others may suit black examinees. Test developers have to take care that an acceptable mix of contexts or situations is represented in their tests and that a majority of items does not come from a context that favours some groups over others. Humphreys (1986) emphasised and Roznowski (1987) reiterated that the only way to hold in check the contribution to total test-score variance of individual, secondary dimensions (or non-trait determinants as they called them), is to choose items that represent a diversity of the latter. As the number of such items measuring both primary and secondary dimensions increases (assuming a fixed total number of items), the less any individual secondary dimension will contribute to total test variance. Towards this end, cognitive tests are typically generated in terms of a two-way (content-by-skill) matrix where each skill is to be tested by means of several kinds of content. For example, the skills included in the verbal tests of the SAT are reading comprehension, sentence completion, analogies and antonyms. Within each of these skills, several content categories are delineated. For example, the last three skill areas are subdivided into four content areas, namely aesthetics/philosophy (music, art, architecture, literature, and philosophy), practical affairs (money, tools, mechanical objects, sports, and historical topics), social science and physical science (Roussos & Stout, 1996). Unless such a diversity of content areas is represented in such tests, the relative superiority of women (as compared to men) in tests of reading comprehension, or the relative superiority of men (as compared to women) in tests of logical reasoning could be called into question and be attributed to a predominance of content favouring the better-performing group.

Stricker and Emmerich's (1999) research reviewed earlier suggests that not only differential familiarity with test content but also differences in the interestingness or offensiveness of test content to different groups should be taken into consideration. For this purpose, major test publishers, such as ETS, have so-called sensitivity reviews that are designed to ensure that topics "that acknowledge the contributions of women and minority group members" (Linn, 1993, p.356) are not underrepresented and that those that may offend or patronise such groups are avoided.

THE DISTINCTION BETWEEN DIF AND PREDICTIVE BIAS

Differential item functioning (earlier known as item bias) is concerned with the interrelationships between test components (items or bundles of items) and test composites and with differences in the performance of matched subgroups from different subpopulations on such components. Such analyses ignore correlations with any external criteria. By contrast, differential validity and differential prediction (also known as predictive bias) involve the relationships between total test scores and scores on relevant criteria. Shealy and Stout (1993b) make the distinction that in differential prediction the criterion is regressed on the test, whereas in DIF analyses the total test score is regressed on the items contained in the test. Drasgow (1987) emphasised that a DIF study is only the first part of a study to investigate potential bias; the second study that is required should investigate the possibility of differential prediction.

Differential validity deals with the question whether the test-criterion correlation in one subpopulation is equal to the corresponding correlation in another subpopulation. For example, is the correlation between scores on an academic aptitude test and matriculation marks equally large for boys and girls? However, equal predictive validity for two groups in the sense of such equal test-criterion correlations does not preclude the possibility of predictive bias as the latter also requires equal criterion-on-test regression intercepts for such groups. For example, predictive bias may deal with the question whether the regression line for predicting matriculation performance on the basis of aptitude test performance is the same for boys and girls.

In Figures 1 and 2, the comparable, elliptical shape of the scatter diagrams for two groups implies no differential validity as there is no difference in the correlation between a test (X) and a criterion (Y) for the two groups. Furthermore, in Figure 1 no predictive bias is present because the mean of the actual criterion scores agrees with the criterion score y predicted by any predictor score x for members of both groups. However, for any predictor score (e.g., x) in Figure 2, the regression line for Group A will systematically under-predict the criterion performance of individuals from Group B because the mean of the actual criterion scores (y_2) of this group will consistently be higher than the mean of those predicted for them (y_1).

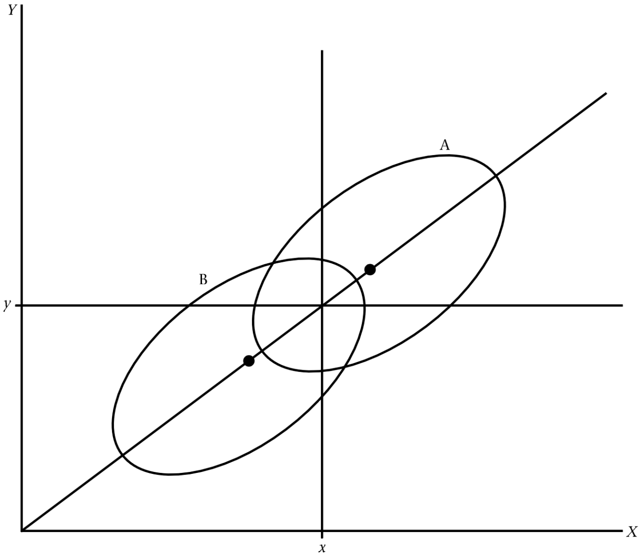

Figure 1: No predictive bias is present

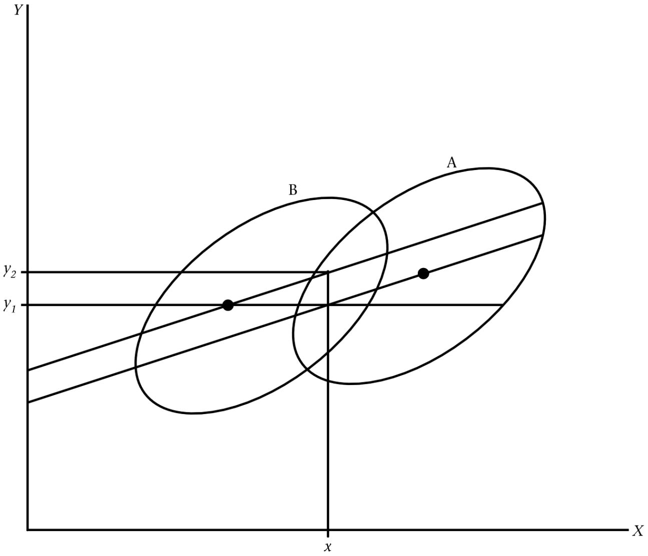

Figure 2: Predictive (intercept) bias present

It is tempting to think that test validity (e.g., the correlation between total test scores and scores on a relevant criterion) will necessarily improve and differential prediction will necessarily decrease upon the removal of items that are differentially functioning owing to, for example, differential familiarity with the item content. (Stated more briefly, the assumption is easily made that a test consisting of unbiased items will be predictively unbiased.) Statements to this effect are not uncommon. However, item bias and test bias are theoretically independent and in practice opposite results may be obtained for these kinds of analyses (Humphreys, 1986).

Roznowski's (1987) empirical research findings demonstrate that a lack of familiarity with some contexts does not necessarily reduce the test-criterion correlation. She compiled two composites of several widely different, narrowly focused subtests of general knowledge for 10th-grade boys and girls in the 1960 Project TALENT. The 20 subtests favouring girls included those dealing with music and home economics, whereas the 20 subtests catering for boys encompassed topics such as the military and farming. Most of the subtests included in the composite favouring males showed a mean difference of at least one standard deviation in favour of males; those included in the composite favouring females displayed a comparable difference in favour of girls. (In terms of Cohen's [1988] classification of effect sizes, these differences would have qualified as large effects.) The tests were chosen so as to share little intertest covariance. The criterion was a measure of general intelligence used in Project TALENT and included reading comprehension, arithmetic reasoning and abstract reasoning. There was a (negligible) difference of 0,05 standard deviation units in favour of boys on this criterion.

The test-criterion correlations (with those corrected for unreliability given between parentheses) for the composite favouring boys were 0,81 (0,84) for boys and 0,82 (0,86) for girls; for the composite favouring girls, these correlations were 0,71 (0,76) for boys and 0,75 (0,80) for girls. After three poorly discriminating subtests were removed from the composite for girls, the resulting composite correlated 0,81 for boys and 0,83 for girls. These results show that the test-criterion correlations were not practically lower when the two groups completed the subtests not favouring their own group. (Although the separate regression lines for boys and girls for both the male and the female advantage composites are not given, intercept bias would have been expected because each group had a lower mean when completing the composite favouring the other group.)

When the two composites (20 subtests favouring boys plus 20 favouring girls) were combined, the test-criterion correlations were slightly higher, namely 0,83 for males and 0,84 for girls (and 0,90 and 0,91 respectively, if corrected for unreliability). Moreover, the regressions for predicting the intelligence test scores on the basis of the combined composites were virtually the same for boys and girls. These results indicate that the inclusion of a diverse array of components, some favouring one group and others favouring another group (i.e., possible candidates for removal by state-of-the-art DIF methods), is not necessarily detrimental to test validity and is not necessarily conducive to differential prediction.

Admittedly, in practice one would not compile a test by including only components that are widely differentially familiar to the reference and focal groups but one would rather strive for components that are equally familiar to them. Nevertheless, the exclusion of differentially functioning items would not necessarily eliminate predictive bias in the resulting tests. Suppose that for most of the lower total test scores (matching variable), the subgroup from the reference group was smaller in size than the matched subgroup from the focal group and that for most of the higher total test scores, the subgroup from the reference group was larger than the matched subgroup from the focal group. (Although the matched subgroups differed in size, their probabilities of passing the items retained were the same as dictated by the non-DIF criterion.) Now, in such a case the mean on the total test of the reference group will be higher than that of the focal group, although none of the items that are contained in the test is functioning differentially for the two groups. If, in such a situation, the standardised difference in the criterion means of the two groups is smaller than the standardised difference in their test means, predictive bias as depicted in Figure 2 may be obtained.

CONCLUSION

N. Schmitt (1999, p. 550), an industrial psychologist, concluded that studies "have repeatedly found that the number of items that display dif is close to chance, and that there is no consistency in terms of the content or type of items that typically display dif." This statement may be applicable to the industrial-psychological field, but judging from the contents of journals such as the *Journal of Educational Measurement*, DIF analyses have merit, particularly when translated tests are involved. Moreover, if the same kind of content consistently produces even a small degree of DIF, this is something that should be taken into account in test construction. Finally, as shown above, some degree of progress has been made in the identification of potential sources of DIF.

Results of DIF analyses performed elsewhere have implications for South African test developers particularly as far as the use of tests in a multi-lingual and multi-cultural society is concerned. If the DIF-inducing capacity of differences in the familiarity, appeal and offensiveness of item content found elsewhere is viewed against the backdrop of the widely differing cultural experiences of various South African groups, the need for DIF analyses is clearly evident. White South Africans may be more familiar with content associated with the game of rugby, whereas black South Africans may be more conversant with the rules of soccer. Items in a reading comprehension test that are formulated in terms of a paragraph that describes the history of the Bollywood movie industry may have a greater appeal among examinees from Indian descent than among the rest of the population. An item dealing with the identity of the member of a previous parliament who stated that the death of Steve Biko had left him cold may be more offensive to black than to white examinees. Because psychological tests have fallen into disrepute among some sections of South African society, even a small number of such differentially functioning items may cause tremendous harm to psychological tests and testing. Such little leaks may sink the ship.

Although very few South African tests have been translated from English into any of the indigenous, black-African languages, the review of DIF results in translated tests has important implications for South African test translation efforts. There is certainly a greater correspondence between the French and English languages (both belonging to the same family of languages) than between English and any of the indigenous black-African languages in South Africa. Simply in terms of sentence length – an aspect that consistently has proven to be a source of DIF elsewhere – one may expect DIF in tests that have been translated from English into the latter languages. Moreover, the cultural differences between the white, English-speaking and the black-African language groups may also be expected to be greater than those between, for example, the English and French Canadians. In view of this, more profound DIF results are likely to occur in locally translated tests than those found by Gierl and Khaliq (2001) in a Canadian context.

By using (untranslated) tests in English for use with individuals whose home language is not English with a view to avoiding DIF due to translation, one does not solve the present problem but rather compounds it. In such instances, test performance is not only affected by the individuals' standing on the construct involved, but also by their familiarity with the English language and with their familiarity with the situations depicted in the test. Test adaptations with the concomitant DIF analyses have yet to begin in all earnest in South Africa.

REFERENCES

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

Allalouf, A., Hambleton, R.K., & Sireci, S.G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, *36*, 185-198.

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

Bond, L. (1993). Comments on the Neill & McPeek Paper. In P.W. Holland & H. Wainer (Eds), *Differential item functioning* (pp. 277-279). Hillsdale, NJ: Erlbaum.

Clauser, B.E. & Mazor, K.M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*, 31-44.

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

Cole, N.S. & Zieky, M.J. (2001). The new faces of fairness. *Journal of Educational Measurement*, *38*, 369-382.

Cook, L., Schmitt, A. & Brown, C. (1999, May). Adapting achievement and aptitude tests: A review of methodological issues. Paper presented at the International Conference on Test Adaptation, Washington, DC.

DeAyala, R.J. & Hertzog, M.A. (1991). The assessment of dimensionality for use in item response theory. *Multivariate Behavioral Research*, *26*, 765-792.

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

Ellis, B.B. (1989). Differential item functioning: Implications for test translations. *Journal of Applied Psychology*, *74*, 912-921.

Ellis, B.B. & Kimmel, H.D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology*, *77*, 177-184.

Gierl, M.J. & Khaliq, S.N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement*, *38*, 164-187.

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, *10*, 229-244.

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

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of item response theory*. Sage: Newbury Park.

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

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

Linn, R.L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P.W. Holland & H. Wainer (Eds), *Differential item functioning* (pp. 349-364). Hillsdale, NJ: Erlbaum.

Linn, R.L., Levine, M.C., Hastings, C.N. & Wardrop, J.L. (1980). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, *5*, 159-173.

Millsap, E.E. & Everson, H.T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297-334.

Nandakumar, R. (1994). Assessing dimensionality of a set of items – Comparisons of different approaches. *Journal of Educational Measurement*, *31*, 17-35.

O'Neill, K.A. & McPeek, W.J. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland & H. Wainer (Eds), *Differential item functioning* (pp. 255-279). Hillsdale, NJ: Erlbaum.

Penfield, R.D. & Lam, T.C.M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, *19* (3), 5-15.

Roznowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: Implications for measurement bias. *Journal of Applied Psychology*, *72*, 480-483.

Roussos, L. & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20* (4), 355-371.

Schmitt, A.P. (1988). Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *25*, 1-13.

Schmitt, N. (1999). Relevant variables for testing diverse groups. [Review of the book Test interpretation and diversity]. *Contemporary Psychology*, *44*, 550-552.

Shealy, R.T. & Stout, W.F. (1993a). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159-194.

Shealy, R.T. & Stout W.F. (1993b). An item response theory model for test bias and differential item functioning. In P.W. Holland & H. Wainer (Eds), *Differential item functioning* (pp. 197-239). Hillsdale, NJ: Erlbaum.

Sireci, S.G. & Geisinger, K.F. (1998). Equity issues in employment testing. In J. Sandoval, C.L. Frisby, K.F. Geisinger, J.D. Scheuneman & J.R. Grenier, (Eds), *Test interpretation and diversity* (pp.105-140). Washington, DC: American Psychological Association.

Stricker, L. E. & Emmerich, W. (1999). Possible determinants of differential item functioning: Familiarity, interest, and emotional reaction. *Journal of Educational Measurement*, *36*, 347-366.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H. Wainer (Eds), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.