SA Journal of Industrial Psychology, 2007, 33 (1), 102-117
SA Tydskrif vir Bedryfsielkunde, 2007, 33 (1), 102-117

CONFESSIONS, SCAPEGOATS AND FLYING PIGS: PSYCHOMETRIC TESTING AND THE LAW[1]

CALLIE THERON
Department of Industrial Psychology
University of Stellenbosch

ABSTRACT

The use of psychometric tests in personnel selection has been regarded with an extraordinary degree of suspicion and scepticism. This is especially true when selection occurs in respect of a diverse applicant group. Concern is expressed about the seemingly uncritical embracing of specific tenets related to the use of psychometric tests in personnel selection in the absence of any systematic coherent psychometric argument to justify these beliefs. The absence of such a supporting psychometric rationale seems unfortunate in as far as it probably would inhibit the independent critical evaluation of the psychometric merits of these generally accepted beliefs. Specific beliefs related to selection fairness, measurement bias and adverse impact are critically examined.

Key words
Measurement bias, employment equity, selection fairness, prediction bias, adverse impact

Selection, as it is traditionally interpreted represents a critical human resource intervention in any organisation in as far as it regulates the movement of employees into, through and out of the organisation. As such selection firstly represents a potentially powerful instrument through which the human resource function can add value to the organisation (Boudreau, 1991; Cascio, 1991b; Cronshaw and Alexander, 1985). However, selection secondly also represents a relatively visible mechanism through which access to employment opportunities is regulated. Because of this latter aspect, selection, more than any other human resource intervention, has been singled out for intense scrutiny from the perspective of fairness and affirmative action (Arvey & Faley, 1988; Milkovich & Boudreau, 1994). More specifically the use of psychometric tests in personnel selection has been regarded with an extraordinary degree of suspicion and scepticism. This is especially true if selection occurs in respect of a diverse applicant group. In South Africa this seems to be true not only for labour representatives and government officials, but also for quite a number of human resource management professionals. The problem is not that the use of psychometric tests in personnel selection is being challenged as such. Rather the concern lies in the seemingly uncritical embracing of specific tenets regarding the use of psychometric tests in personnel selection in the absence of any systematic coherent psychometric argument to justify these beliefs. The absence of such a supporting psychometric rationale seems unfortunate because it prevents the independent critical evaluation of the psychometric merits of these generally accepted beliefs and it most likely would stifle an open-minded, creative search for effective and equitable selection practices. Efficient and equitable personnel selection in respect of a diverse applicant pool is a complex present-day human resource management problem that requires a mature, creative and innovative response from the Industrial Organisational Psychology fraternity in South Africa that acknowledges the intricacies and complexities inherent to the problem. In addition, the danger exists that the manner in which the Industrial Organisational Psychology fraternity in South Africa responds to the challenge in the popular press, academic literature and conference papers (*mea culpa*) could perpetuate and reinforce the somewhat superficial, black box, non-analytical approach one typically finds regarding the problem.

The following seems to be some of the more prominent beliefs that seem to have developed in South Africa as psychometric dogma that apparently guides the day-to-day responses of many human resource management professionals in their use of psychometric tests in the work place.

- It is possible to assure selection fairness solely through the judicious choice of selection instruments. Or in its alternative formulation, it is possible to avoid unfair discrimination in personnel selection solely through the use of reliable, valid and unbiased selection instruments (i.e., instruments that are free from measurement bias);
- It is possible to avoid biased assessments/measures through the judicious choice of properly developed selection instruments;
- It is possible to avoid adverse impact through the judicious choice of assessment/selection instruments. Or in its alternative formulation, it is possible to grade selection instruments in terms of the degree of adverse impact they create;
- Adverse impact should be equated with unfair discrimination; and
- It is possible to certify assessment techniques as Employment Equity Act (Republic of South Africa, 1998) compliant.

Informal observation seems to suggest that a significant number of human resource management professionals in South Africa would endorse all of the above claims. It seems as if in the mind of many human resource management professionals there exists the belief that if they were sufficiently cautious and fastidious in their choice of selection instruments they could gain psychometric salvation and immunity from the Employment Equity Act (Republic of South Africa, 1998). More specifically the belief seems to be that selection procedures will not discriminate unfairly against members of previously disadvantaged groups nor will they create adverse impact against such groups as long as the selection instruments used in these procedures are valid and provide unbiased measures of the intended latent variable (Sehlapelo & Terre Blanche in Bredell, van Eeden & van Staden, 1999; Van der Merwe, 1999; Van der Merwe, 2002; Visser & De Jong, 2000). Humphreys (1986, p. 327) makes a similar observation in the context of the USA:

> A civil rights activist who looks at this literature and listens to psychologists at meetings might well conclude that minority problems in admission to higher education, hiring in industry, and classification in military services will be solved when bias is eliminated from tests.

Although Humphreys (1986) refers to both measurement bias and predictive bias in this observation, he nonetheless then goes on (p. 327) to comment:

1 The insightful and valuable comments and suggestions for improvement to this manuscript made by prof Gert Huysamen are gratefully acknowledged. Liability for the views expressed in this manuscript, however, remains solely that of the author.. The thorough review and constructive suggestions made by two anonymous reviewers are also gratefully acknowledged.

> Many have implicitly assumed that a test composed of unbiased items will also be unbiased in the first (predictive) sense, but the two types of bias can frequently be quit independent or even opposite to each other.

The Employment Equity Act (Republic of South Africa, 1998) seems to echo the foregoing conviction by prohibiting the use of psychological tests unless it can be shown that the tests are valid and not biased against any employee or group (i.e., without measurement bias). Specifically the Employment Equity Act (Republic of South Africa, 1998, p.14) prohibits unfair discrimination by stating that:

> No person may unfairly discriminate, directly or indirectly, against an employee, in any employment policy or practice, on one or more grounds, including gender, sex, pregnancy, marital status, family responsibility, ethnic or social origin, colour, sexual orientation, age, disability, religion, HIV status, conscience, belief, political opinion, culture, language and birth.

At the same time, however, paragraph 2(b) of the Employment Equity Act (Republic of South Africa, 1998, p. 14) could be interpreted to mean that it does not constitute unfair discrimination to use selection instruments that demonstrate predictive validity to distinguish between, exclude or show preference for any applicant:

It is not unfair discrimination to-
a) take affirmative action measures consistent with the purpose of this Act, or
b) distinguish, exclude, or prefer any person on the basis of an inherent requirement of a job.

Under a construct orientated approach to personnel selection (Binning & Barrett, 1989) selection instruments demonstrate predictive validity if inferences about reliable and valid measures of job performance can permissibly be made from valid and reliable measures of person attributes that determine the level of job success that will be achieved (Guion, 1998; Messick, 1989). In this sense those attributes that correlate with job performance could be regarded as inherent requirements of the job. In paragraph 8 of the Employment Equity Act (Republic of South Africa, 1998, p. 16) this position is reiterated and qualified by requiring that all selection instruments should be valid[2] while at the same time their measures should not be biased against members of any of the previously cited protected groups:

Psychological testing and other similar assessments of an employee are prohibited unless the test or assessment being used-
a) has been scientifically shown to be valid and reliable;
b) can be applied fairly to all employees;
c) is not biased against any employee or group.

Presumably the prohibition of biased psychological tests is seen to serve the objective of the Act of "promoting equal opportunity and fair treatment in employment through the elimination of unfair discrimination" (Republic of South Africa, 1998, p. 12). When referring to tests or assessments that are not biased against any employee or group, moreover, the Act is referring to measurement bias. Although not necessarily all studies have been precipitated by the Act, the argument that the elimination of measurement bias would necessarily prevent unfair discrimination nonetheless seems to have inspired a number of bias studies in South Africa (Abrahams & Mauer, 1999; Schaap, 2001; Schaap, 2003; Schaap & Basson, 2003; van Zyl & Visser, 1998). This line of reasoning also quite often seems to form the essence of the argument in terms of which the necessity of measurement bias analysis in South Africa is motivated (Kanjee, 2001). In terms of this psychometric test view it would, moreover, not be inappropriate if test publishers and distributors would certify instruments as EEA compliant. In fact it would probably be welcomed as a very useful guide in the choice of selection instruments (Lopes, Roodt & Mauer, 2001). The seal of approval is after all meant to communicate the assurance that use of the test in question would serve the objective of the Act of "promoting equal opportunity and fair treatment in employment through the elimination of unfair discrimination" (Republic of South Africa, 1998, p. 12). As a case in point a HSRC test catalogue (2003) has recently awarded the LPCAT with an EEA compliant seal of approval, presumably because of the commendable rigor with which item bias analysis has been performed using latent trait theory (De Beer, 2000).

There finally exists the belief that the origin of adverse impact resides in the selection instruments used for personnel selection or in the differences in the latent trait being assessed. As an expression of the former belief Sackett and Ellingson (1997, p. 707) for example, report (italics added):

> An ongoing concern in the field of personnel selection is the search for selection systems with high validity and low adverse impact (i.e., similar selection ratios for majority and minority groups). A longstanding source of tension in this area results from certain types of predictors emerging as valid indicators of performance, but also exhibiting substantial group differences. For example, extensive research has demonstrated a strong relationship between general cognitive ability and job performance for multiple jobs (Hunter, 1986; Re & Earles, 1991). *However, cognitive tests traditionally demonstrate adverse impact against racial minorities* (Hartigan & Widor, 1989; Jensen, 1980).

Maxwell and Arvey (1993) also seem to subscribe to this point of view when they define the standardised difference in mean predictor performance between protected and non-protected groups ($(\mu_{XNP} - \mu_{XP})/\sigma_X$) as an index of adverse impact. Moreover the belief exists that selection instruments differ in terms of the adverse impact that they impose on protected groups and thus can be graded in terms of their relative degree of adverse impact. The extremely influential and highly respected Uniform Guidelines on Employee Selection Procedures published by the Equal Employment Opportunity Commission (EEOC) endorses this position by requiring that:

> Where two or more selection procedures are available which serve the user's legitimate interest in efficient and trustworthy workmanship, and which are substantially equally valid for a given purpose, the user should use the procedure which has been demonstrated to have the lesser adverse impact (Equal Employment Opportunity Commission, 1978, p. 38297).

The conviction that adverse impact is fundamentally determined by differences in mean predictor performance resulted in the investigation of various strategies to reduce these subgroup differences in mean predictor scores in an effort to increase the representation of members of protected groups without sacrificing predictive accuracy (Sackett, Schmitt, Ellingson & Kabin, 2001). These include the use of valid, non-cognitive predictors (Sackett & Ellington, 1997; Sackett et al., 2001; Schmitt, Rogers, Chan, Sheppard & Jennings, 1997), identification and removal of culturally biased items in the predictor (Humphreys, 1986; Sackett et al., 2001), the use of alternative modes of presenting predictor stimuli (Chan & Schmitt, 1997; Pulakos & Schmitt, 1996; Sackett et al., 2001) and the use of coaching or orientation programmes (Sackett et al., 2001).

The question is whether the broad psychometric stance outlined above, in which the predictor, or some combination of predictors, is the primary villain responsible for most if not all of the evils associated with personnel selection from a diverse applicant pool, is a psychometrically justified one that best

2 Logically the EEA must thereby refer to the permissibility of criterion construct inferences made from predictor measures (i.e., predictive validity) rather than the permissibility of predictor construct inferences (i.e., construct validity), although the latter needs to be demonstrated to convincingly establish the former.

serves the interests of all stakeholders involved? More to the point, will it assist in achieving the extremely laudable vision formulated by then president Mandela in the preamble to the Employment Equity Bill (Republic of South Africa, 1996, p. 5)?

> What we are against is not the upholding of standards as such but the sustaining of barriers to the attainment of standards; the special measures that we envisage to overcome the legacy of past discrimination are not intended to ensure the advancement of unqualified persons, but to see to it that those who have been denied access to qualifications in the past can become qualified now, *and that those who have been qualified all along but overlooked because of past discrimination, are at last given their due.*

The objective of this article is to critically reflect on the psychometric tenability of the viewpoint outlined above. More specifically, the intention is to identify specific flaws in the foregoing argument and to outline the implication of these flaws for the two-pronged employment equity objective of the Employment Equity Act (Republic of South Africa, 1998) reflected in the preamble to the Employment Equity Bill quoted earlier. It is hoped that the argument presented here will elicit an open and frank debate amongst South African human resource management professionals. To paraphrase Guion (1998, p. 470), fair selection, measurement bias and adverse impact are topics too important to ignore or bury under popular rhetoric.

THE FUNDAMENTAL LOGIC UNDERLYING PERSONNEL SELECTION

Assuming that only a limited number of vacancies exist, the task of the selection decision maker is in essence to identify a subgroup from the total group of applicants to allocate to the accept treatment (Cronbach & Gleser, 1965), based on limited but relevant information about the applicants. The subgroup, furthermore, has to be chosen so as to maximise the average gain on the utility scale on which the outcomes of decisions are evaluated. The utility scale/payoff and the actual outcomes or ultimate criterion (Austin & Villanova, 1992) are the focus of interest in selection decisions (Bartram, Baron & Kurz, 2003; Ghiselli, Campbell & Zedeck, 1981). In personnel selection decisions, future job performance forms the basis (i.e., the criterion) on which applicants should be evaluated so as to determine their assignment to an appropriate treatment (Cronbach & Gleser, 1965). Information on actual job performance can, however, never be available at the time of the selection decision. Under these circumstances, and in the absence of any (relevant) information on the applicants, no possibility exists to enhance the quality of the decision making over that that could have been obtained by chance. This seemingly innocent, but too often ignored, dilemma points to a key fact that needs to be continually kept in mind when contemplating the psychometric merits of the predictor centred selection model outlined earlier. The crucial point that needs to be appreciated is that the only alternative to random decision making (other than not to take any decision at all), would be to predict expected criterion performance (or expected utility) actuarially (or clinically) from relevant, though limited, information available at the time of the selection decision and to base the selection decision on these criterion-referenced inferences[3]. This implies that in personnel selection the primary focus is on the criterion rather than on the predictor from which inferences about the criterion are made (Schmitt, 1989). This position is formally acknowledged by the APA sanctioned interpretation of validity and especially predictive validity (Ellis & Blustein, 1991; Landy, 1986; Messick, 1989; Society for Industrial and Organizational Psychology, 2003). The position, moreover, underlies the generally accepted regression-based interpretations of selection fairness (Cleary, 1968; Einhorn & Bass, 1971; Huysamen, 2002). Very little if anything of this realisation is, however, evident in the views on psychometric testing and the law put forward by Bonthuys (2002) in a somewhat cynically titled paper[3]. Even though it is logically impossible to directly measure the performance construct at the time of the selection decision, it can nonetheless be predicted at the time of the selection decision if: (a) variance in the performance construct can be explained in terms of one or more predictors (b) the nature of the relationship between these predictors and the performance construct has been made explicit; and (c) predictor information can be obtained prior to the selection decision in a psychometrically acceptable format. The only information available at the time of the (fixed treatment) selection decision (Cronbach & Gleser, 1965) that could serve as such a substitute would be psychological, physical, demographic or behavioural information on the applicants. Such substitute information would be considered relevant to the extent that the regression of the (composite) criterion on a weighted (probably, but not necessarily, linear) combination of information explains variance in the criterion. Thus the existence of a relationship, preferably one that could be articulated in statistical terms, between the outcomes considered relevant by the decision maker and the information actually used by the decision maker, constitutes a fundamental and necessary, but not sufficient, prerequisite for effective and equitable selection decisions.

Measurement data, once obtained, is translated into decisions in accordance to some strategy for decision-making (Cronbach & Gleser, 1965). A decision strategy describes how scores from tests are to be combined with non-test information, and what decision will be made for any given combination of facts. A strategy is thus a rule for arriving at selection decisions used by a decision maker in any possible contingency (Cronbach & Gleser, 1965). It consists of a set of specified conditional probabilities (typically either zero or unity), which reflects the policy of the decision-maker. In the final analysis it is the selection decision strategy that should be evaluated in terms of its predictive validity - in other words in terms of the correspondence that exists between the criterion-referenced inferences made via the decision rule from the available predictor information and the actual criterion performance achieved. Demonstrating that the available predictor variables individually correlate significantly with the criterion thus constitutes insufficient evidence to justify a selection procedure. Even demonstrating that the available predictor variables in combination correlate significantly with the criterion would constitute insufficient evidence to justify a selection procedure if the manner in which the predictors are combined would differ between application and validation. This important realisation often seems to be absent in validation studies, which combine selection information in accordance with a clinical or judgemental strategy (Gatewood & Feild, 1994).

Several selection decision-making strategies exist that range from purely clinical to purely mechanical combinations of data available to the decision maker (Grove & Meehl, 1996; Kleinmutz, 1990; Gatewood & Feild, 1994; Murphy & Davidshofer, 1988). All of these require that the nature of the relationship between the criterion and the substitute information be understood. The two extreme options, however, differ in the way they express their understanding of the criterion-information relationship. Clinical prediction involves combining information from test scores and measures obtained from interviews and observations covertly in terms of

3 This raises an important question on the appropriateness of the extensive use of conventional construct-referenced norms (e.g., percentile ranks, stens, stanines) for the interpretation of assessments in a selection context. Construct-referenced norms shed light on the relative strength of a latent trait (assumed to be in part a determinant of job performance) by interpreting the observed test scores in terms of the relative position of the score in the normative distribution. This, however, still leaves the real question of interest unanswered, namely what level of job performance could be expected given the relative strength of the said latent trait. Criterion-referenced norms are thus required that interpret the observed test score in terms of the expected position of the associated job performance in the criterion distribution. Criterion-referenced norms in turn require that the regression of the criterion on the predictor is accurately understood. **4** A worrisome question is to what extent the views on psychometric testing expressed by Bonthuys (2002) are representative of the legal fraternity in South Africa?

an implicit combination rule imbedded in the mind of a clinician to arrive at a judgment about the expected criterion performance of the individual being assessed (Grove & Meehl, 1996; Gatewood & Feild, 1994; Murphy & Davidshofer, 1988). Mechanical prediction involves using the information overtly in terms of an explicit combination rule to arrive at a judgment about the expected criterion performance of the individual being assessed (Gatewood & Feild, 1994; Murphy & Davidshofer, 1988). An actuarial system of prediction represents a mechanical method of combining information, derived via statistical or mathematical analysis from actual criterion and predictor data sets, to arrive at an overall inference about the expected criterion performance of an individual (Meehl, 1957; Murphy & Davidshofer, 1988). An actuarially derived decision rule should, therefore, more accurately reflect the nature of the relationship that exists between the various latent predictor variables and the criterion construct than a clinically derived selection decision rule. The former would, in all likelihood, also be more consistently applied than the latter.

The accuracy of clinical and actuarial prediction has been studied widely (Dawes & Corrigan, 1974; Dawes, 1971; Goldberg, 1970; Grove & Meehl, 1996; Kleinmutz, 1990; Meehl, 1954; 1957; Murphy & Davidshofer, 1988). These reviews seem to suggest that clinicians very rarely make better predictions that can be made using actuarially derived prediction methods, that statistical methods are in many cases more accurate in predicting relevant criteria than are highly trained clinicians, and that clinical judgement should be replaced, wherever possible, by mechanical methods of integrating the information used in forming predictions (Murphy & Davidshofer, 1988). Grove and Meehl, (1996) for example quite categorically argue in favour of the mechanical combination of selection data.

The decision whether to accept an applicant is based on the mechanically or judgementally derived expected outcome conditional on information on the applicant or, if a minimally acceptable outcome state can be defined, the conditional probability of success (or failure) given information on the applicant. Alternatively, the bivariate distribution could be converted into a contingency table through the formation of intervals on both the predictor and the criterion. The resultant validity matrix (Cronbach & Gleser, 1965) or expectancy table (Ghiselli, Campbell & Zedeck, 1981; Lawsche & Balma, 1966), indicating the probability of a specific criterion state conditional on a specific information category, could then be used as basis for decision-making. Given the objective of human resource management in general and personnel selection in particular to add value, a strict top-down selection decision-rule is furthermore assumed, based on expected criterion performance or the conditional probability of success.

IN SEARCH OF SELECTION FAIRNESS

The question is firstly whether the *selection decision strategy* under investigation is worth implementing in comparison to an alternative (possibly currently existing) strategy. Utility analysis (Boudreau, 1989; 1991; Brogden, 1949a; Cascio, 1991b; Cronbach & Gleser, 1965; Naylor & Shine, 1965; Taylor & Russell, 1939) aims to provide an answer to this question in terms of various indices for judging worth. The question is moreover whether the *decision strategy* that will dictate the categories to which applicants will be assigned (accept or reject) for any given combination of facts, can be considered fair. Stated differently, the question is whether the decision strategy will directly or indirectly put members of specific applicant groups at an unfair, unjustifiable disadvantage. Selection measures are designed to discriminate and in order to accomplish their professed objective they must do so (Cascio, 1991a). However, due to the relative visibility of the selection mechanism's regulatory effect on the access to employment opportunities, the question readily arises whether the selection strategy discriminates fairly. Selection fairness, however, represents an exceedingly elusive concept to pin down with a definitive constitutive definition. The *Standards for Educational and Psychological Testing (Standards)* acknowledges this dilemma (AERA, APA & NCME, 1999). The problem is firstly that the concept cannot be adequately defined purely in terms of psychometric considerations without any attention to moral/ethical considerations. The inescapable fact is that, due to differences in values, one man's foul is another man's fair (Huysamen, 1995). The problem is further complicated by the fact that a number of different definitions and models of fairness exist which differ in terms of their implicit ethical positions and which, under certain conditions, are contradictory in terms of their assessment of the fairness of a selection strategy and their recommendations on remedial action (Petersen & Novick, 1976; Cascio, 1991a; Arvey & Faley, 1988). Three distinct fundamental ethical positions (Hunter & Schmidt, 1976) underpinning views on what constitutes fair selection have been identified. A fairness model, based on any one of these ethical positions (or a variant thereof), formalises the interpretation of the fairness concept and thus permits the deduction of a formal investigative procedure to assess the fairness of a particular selection strategy should such a strategy be challenged in terms of a *prima facie* showing of adverse impact (Arvey & Faley, 1988; Singer, 1993).

A definite stance on what constitutes fair or unfair discrimination in personnel selection nonetheless needs to be taken. Since the Employment Equity Act (Republic of South Africa, 1998) and the Promotion of Equality and Prevention of Unfair Discrimination Act (Republic of South Africa, 2000) both explicitly prohibit unfair discrimination, a definite verdict on the fairness of the criterion inferences made during selection needs to be pronounced. If the equity objective of the Act is to be reached, we must commit to a specific interpretation of selection fairness and stop hiding behind the protest that it is impossible to produce definitive constitutive and operational definitions of selection fairness. The question, however, is, which of the variety of fairness models that have been proposed (Arvey & Faley, 1988; Cascio, 1991a; Huysamen, 1995; Petersen & Novick, 1976) would serve the spirit of the Employment Equity Act (Republic of South Africa, 1998) best.

Influential technical guidelines on personnel selection procedures (Equal Employment Opportunity Commission, 1978; Society for Industrial and Organizational Psychology, 2003; Society for Industrial Psychology, 1998) seem to favour unqualified individualism as the basic ethical point of departure. The basic premise is that applicants with an equal probability of succeeding on the job (being applied for and at the time of the selection decision) should have an equal probability of obtaining the job, irrespective of group membership (AERA, APA & NCME, 1999; Guion, 1966; 1991; Huysamen, 2002). This fundamental premise, moreover, seems to be in agreement with the anti-discrimination objectives of the Employment Equity Act (Republic of South Africa, 1998) as voiced by the previously quoted preamble to the Employment Equity Bill (Republic of South Africa, 1996). To that should probably be added the principle voiced by the Principles for the Validation and Use of Personnel Selection Procedures (AERA, APA & NCME, 1999; Society for Industrial and Organizational Psychology, 2003) that all applicants should receive a uniform treatment in terms of testing conditions, access to training material, feedback and retest opportunities. This latter interpretation seems to correspond with the stance of the Employment Equity Act (Republic of South Africa, 1998, p. 16) that:

Psychological testing and other similar assessments of an employee are prohibited unless the test or assessment being used-
b) can be *applied* fairly to all employees

More specifically technical guidelines on personnel selection procedures (AERA, APA & NCME, 1999; Equal Employment Opportunity Commission, 1978; Society for Industrial and Organizational Psychology, 2003; Society for Industrial Psychology, 1998) seem to favour the regression-based models of selection fairness (Cleary, 1968; Einhorn & Bass, 1971; Huysamen, 1996; Huysamen, 2002). Organised labour and other affirmative action proponents could, however, possibly favour the psychometrically less sound quota models (Huysamen, 1996; Petersen & Novick, 1976; Schmitt, 1989). It would, however, probably be wise not to underestimate the business and intuitive psychometric acumen of organised labour representatives. The regression or Cleary model of selection fairness defines fairness in terms of the absence of differences in regression slopes and/or intercepts across the subgroups comprising the applicant population (Arvey & Faley, 1988; Petersen & Novick, 1976; Cascio, 1991a; Maxwell & Arvey, 1993). According to Cleary (Cleary, 1968, p. 115):

> A test is biased for members of a subgroup of the population if, in the prediction of the criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of unfair, particularly if the use of the test produces a prediction that is too low. If the test is used for selection, members of a subgroup may be rejected when they were capable of adequate performance.

The Cleary model thus argues that *selection decision-making, based on expected criterion performance*, can be considered unfair or discriminatory if the position members of specific groups receive in the rank-order resulting from the decision strategy is either systematically too low or systematically too high for members of a particular group. This would happen if group membership explains variance in the (unbiased) criterion, either as a main effect or in interaction with the predictors, which is not explained by the predictors, and *the selection strategy fails to take group membership into account*. Under these conditions the criterion inferences derived from selection instrument scores, could be said to exhibit predictive bias (Guion, 1991; 1998).

The Cleary model therefore examines the fairness of a selection strategy by fitting a saturated regression equation, shown as equation 1 below, and testing the hypothesis H_{01}: $\beta_2 = \beta_3 = 0$ against the alternative hypothesis H_a: at least one of the two parameters is not zero (Bartlett, Bobko, Mosier & Hannan, 1978; Berenson, Levine & Goldstein, 1983; Kleinbaum & Kupper, 1978).

$$E(Y) = \alpha + \beta_1 X + \beta_2 D + \beta_3 XD \quad (1)$$

In equation 1, X is a single predictor or a (clinically or actuarially) weighted combination of predictors, and D is a dummy variable representing group membership such that D = 0 would indicate membership of a protected group and D = 1 membership of a non-protected group (or vice versa).

Should H_{01} not be rejected it would imply that selection decisions based on expected criterion performance derived from the combined regression equation is fair. Should H_{01}, however, be rejected it would imply that selection decision-making based on expected criterion performance derived from the combined regression equation is unfair because the rank-order resulting from the decision strategy is either systematically too low or systematically too high. The inappropriate placement in the selection rank order will result from the use of the combined regression equation because the rejection of the null hypothesis would imply that the separate regression equations differ in terms of slope and/or intercept (i.e. one would have to conclude that the regression models fitted to the two subgroups do not coincide). Although it is almost instinctive to suspect that predictive bias would systematically and unfairly burden applicants from the previously disadvantaged community this has not generally been the case in the United States (Arvey & Faley, 1988; Huysamen, 1996; Huysamen, 2002). Insufficient local research on predictive bias, however, prevents the formulation of a general position on nature and consequences of predictive bias in South Africa. Nonetheless, to a certain extent the subsequent argument (quite possibly erroneously) assumes that when group membership explains variance in the criterion that is not explained by the predictors, and the selection strategy fails to take group membership into account, applicants from the previously disadvantaged community will be unfairly burdened. The essence of the argument would, however, not be affected if the opposite would be true.

The Einhorn-Bass selection fairness model argues that *selection decision-making, based on the conditional probability of success*, can be considered unfair or discriminatory if the position members of specific groups receive in the rank-order resulting from the decision strategy is either systematically too low or systematically too high. The equal risk or Einhorn-Bass selection fairness model thus operationalises the concept of fairness in terms of differences in the probability of success conditional on predictor performance. In terms of the equal risk model a selection strategy would be considered unfair if the probability of a member of the protected group (D = 0) with a given predictor score ($X = x_c$) displaying a criterion performance equal to or higher than Y_c is different from a member of the non-protected group (D = 1) who received the same predictor score (i.e., $P[Y \geq Y_c | X = x_c; D = 0] \neq P[Y \geq Y_c | X = x_c; D = 1]$) and the *selection strategy fails to take this into account* (Petersen & Novick, 1976; Cascio, 1991a; Einhorn & Bass, 1971). The Einhorn-Bass conceptualisation thus corresponds exactly to the Guion (1966, p. 26) definition of unfair discrimination referred to earlier: The equal risk model would therefore judge any selection strategy unfair should it be considered unfair by the Cleary model. In addition, however, it would also consider the selection strategy unfair if the criterion variance conditional on predictor performance differs across the two applicant subgroups (i.e. $\sigma^2_{y|x}; D_0 \neq \sigma^2_{y|x}; D_1$) (Petersen & Novick, 1976; Cascio, 1991a; Einhorn & Bass, 1971). The critical null hypothesis to be tested in terms of the Einhorn-Bass selection fairness model is therefore H_{02}: $\sigma^2_{y|x}; D_0 = \sigma^2_{y|x}; D_1$.

The first critical point to appreciate is that H_{01} and/or H_{02} can be rejected even though the regression of the criterion on the predictor is significant (i.e., the selection instrument demonstrates predictive validity). The Employment Equity Act (Republic of South Africa, 1998) is correct in describing the use of invalid predictors as an unacceptable practice since it violates the fundamental principle of the unqualified individualism position that applicants with an equal probability of succeeding on the job should have an equal probability of obtaining the job, irrespective of group membership (Guion, 1991). Since the use of a completely invalid predictor is tantamount to random selection, it gives all applicants the same probability of obtaining the job despite the fact that they differ in terms of the probability of succeeding on the job. The use of a predictor that demonstrates predictive validity, however, is not a sufficient condition to ensure that the fundamental principle comprising unqualified individualism is complied with. Even when a predictor demonstrates predictive validity, (indirect) discrimination can still unfairly disadvantage members of specific subgroups if group membership significantly explains

variance in the criterion, which is not explained by the predictor, and if the selection strategy fails to take this fact into account. The position of the Employment Equity Act (Republic of South Africa, 1998, p. 14) that:

> it is not unfair discrimination to .... distinguish, exclude, or prefer any person on the basis of an inherent requirement of a job,

therefore seems questionably lenient. Translated into psychometric terms, the Employment Equity Act (Republic of South Africa, 1998, p. 14) seems to hold the questionable position that it is not unfair discrimination to distinguish between, exclude or prefer any person on the basis of the scores obtained on a valid selection instrument. The very essence of selection is to distinguish between, exclude or show preference for individuals on the basis of measures that are systematically related to the criterion [i.e., valid selection instruments]. The question nonetheless remains whether the criterion-referenced inferences derived from the relevant predictor information does not unfairly burden or disadvantage members of specific subgroups? The definition of discrimination[6] provided by the Promotion of Equality and Prevention of Unfair Discrimination Act (Republic of South Africa, 2000) read in conjunction with the Cleary (Cleary, 1968) interpretation of unfair discrimination attests to the questionable nature of the Employment Equity Act position:

1. "discrimination" means any act or omission, including a policy, law, rule, practice, condition or situation which directly or indirectly-
   a) imposes burdens, obligations or disadvantage on; or
   b) withholds any benefits, opportunities or advantages from, any person on one or more of the prohibited grounds

If group membership does significantly explain variance in the criterion, which is not explained by the predictor, and if the selection strategy fails to take this fact into account, significant systematic group-related prediction errors will occur and the selection decision-rule will therefore discriminate since it will disadvantage members of a specific group by placing them inappropriate low in the selection rank order even though the predictor significantly correlates with the criterion. Moreover it could be argued that the current formulation of the Employment Equity Act (Republic of South Africa, 1998) still leaves a critical loophole, which will undermine the realisation of the vision of former President Mandela (Republic of South Africa, 1996, p. 5):

> .... that those who have been qualified all along but overlooked because of past discrimination, are at last given their due..

The appropriate remedy, should H_0 be rejected, is contingent on the explanation for the rejection of the null hypothesis. The Cleary model's prescription for a diagnosed unfair selection strategy thus depends on whether there exists an equivalent incremental difference in criterion performance across applicants from the two subgroups, regardless of predictor performance (i.e. the interaction parameter b_3 can be assumed zero but the group main effect parameter b_2 is assumed non-zero) or a non-equivalent incremental difference in criterion performance across applicants from the two subgroups, dependent on the ability level of the applicants (i.e. there exists a subgroup x predictor performance interaction effect on criterion performance) (Bartlett et al., 1978; Berenson, Levine & Goldstein, 1983; Kleinbaum & Kupper, 1978). The Cleary solution to the fairness problem thus dictates that the information category entries in the strategy matrix (Cronbach & Gleser, 1965) should be derived from an appropriately expanded multiple regression equation containing the group variable either as a main effect and/or as an interaction effect (Bartlett et al., 1978; Schmitt, 1989). This recommendation, however, is contingent on the expanded regression equation successfully cross-validating on a holdout sample (Bartlett et al., 1978). The need to expand the regression equation through the addition of the group variable either as a main effect and/or as an interaction effect should therefore be maintained in independent samples taken from the applicant population.

The Einhorn-Bass solution to the fairness problem would be to derive the information category entries (i.e. $P[Y \geq Y_c | X_i; D_j]$) in the strategy matrix (Cronbach & Gleser, 1965) from the appropriate regression equation. The appropriate conditional probabilities are obtained by deriving $E[Y | X_i; D_j]$ from the appropriate regression equation and subsequently, transforming Y_c to a standard score in the conditional criterion distribution (assuming normality) by using the appropriate standard error of estimate as denominator (Berenson, Levine & Goldstein, 1983; Kleinbaum & Kupper, 1978; Einhorn & Bass, 1971).

In both cases the systematic, group-related over- and under-prediction of the criterion would thereby be removed. The inappropriate positioning of members of protected and non-protected groups in the selection rank order would consequently be corrected. Moreover, due to the closer correspondence of estimated and actual criterion performance, the predictive validity of criterion inferences would thereby also be enhanced. Finally, since selection utility is a positive linear function of validity (Brogden, 1946; 1949a; 1949b; Cochran, 1951), it would pay to eliminate unfair discrimination in the manner dictated by the regression-based models of selection fairness.

The second important point that should be stressed is therefore that all valid predictors can in principle be used fairly in the regression-based sense of the term. The converse is, however, not true even though the Employment Equity Act seems to endorse it. Using a valid predictor is not sufficient to conclude that selection will be fair. Fair or unfair discrimination, therefore, does not reside in the predictor as such. Fair or unfair discrimination, therefore, also does not reside in differences in mean predictor score (Schmitt, 1989). Cleary (1968, p. 115) somehow seemed to have done us a disservice by referring to test bias in her interpretation of selection fairness in as far as the term tends to suggest that unfair discrimination is caused by the test. Logically it therefore is not possible to ensure selection fairness solely through the judicious choice of selection instruments. Stated more strongly - it is a totally futile exercise to try and identify or develop selection instruments that will immunise the human resource practitioner against discriminatory personnel selection practices, irrespective of how great the yearning for such a simple solution might be. In addition, the practice of endorsing specific instruments as Employment Equity Act compliant and thereby reinforcing and perpetuating the belief that it is possible to achieve legal immunity through the judicious choice of selection tools might be well intentioned, but should nonetheless be rejected as a misleading and groundless marketing strategy.

This raises a third important point. By far the majority of selection decisions in South Africa are probably based on clinically (as opposed to actuarially) derived criterion inferences. The validity and fairness of such clinically derived inferences can quite easily be established utilising conventional validation techniques, provided an appropriate criterion measure and a sufficiently large N are available. However, the ability of a clinical selection strategy to adapt itself in a manner that would eliminate systematic prediction errors, should they be identified, seems doubtful. Given that selection decisions are based on (clinically or mechanically derived) estimates of criterion performance, a critical requirement for effective selection is that the nature of the predictor-criterion relationship should be accurately understood. The literature (Dawes & Corrigan, 1974; Goldberg, 1970; Grove & Meehl, 1996;

6 Discrimination, in terms of this definition, should not be equated with unfair discrimination but rather with adverse impact.

Kleinmutz, 1990; Meehl, 1954; 1957; 1956; Dawes, 1971; Murphy & Davidshofer, 1988; Wiggins, 1973) rather unequivocally considers the mechanical methods of integrating the information used in forming predictions as superior to clinical methods (at least with regards to relative short-term predictions). Actuarially derived mechanical decision rules probably derive their superior performance record through their ability to capture the nature of the relationship that exists between the various latent predictor variables and the criterion construct with greater accuracy and the greater consistency with which the rule is applied (Gatewood & Feild, 1994). The problem thus seems that in some cases an already complex job performance structural model that needs to be understood is made even more complex by the fact that a group membership variable not only affects the latent variables that determine job performance, but also affects job performance directly and possibly moderates the effect of one or more latent variables on performance. The likelihood that the clinical mind will be able to accurately understand the manner in which even a small subset of these latent variables combine to determine criterion performance and be able to consistently apply this understanding, therefore seems even smaller than in cases where group membership need not be considered to accurately estimate job performance.

In too many cases where it is feasible to conduct the rigorous validation research required to develop proper actuarial decision rules, it has sadly enough not been performed. In many cases where selection decisions are currently being made, moreover, it will (seemingly) not be feasible to do so. Unless ingenious ways can be found to circumvent the practical obstacles at present preventing these studies (e.g. synthetic validation, inter-organisational cooperation, bootstrapping), the harsh reality will be that in many cases selection fairness will remain an unattainable ideal. Simply because a need for equitable selection exists does not mean that it will necessarily be easily attainable in each and every case; it might even be unattainable in some cases irrespective of how strong the desire for a fair selection procedure might be.

In the United States of America the remedies for unfair selection proposed by Cleary (Cleary, 1968), and Einhorn and Bass (1971), outlined above, would seemingly not be allowed (Huysamen, 2002). The problem is that section 106 (1) of the 1991 Civil Rights Act (in Guion, 1998, p. 468) prohibits the adjustment of test scores on the basis of group membership:

> It shall be an unlawful practice for an employer, in connection with the selection or referral of applicants or candidates for employment or promotion to adjust the scores of, use different cutoffs for, or otherwise alter the results of employment related tests on the basis of race, color, religion, sex or national origin.

In its (quite justified) effort to prohibit within-group (construct-referenced) norming the Civil Rights Act (1991) seemingly worded the relevant section in such broad terms that it could be interpreted to mean that it also is illegal to attach different criterion-referenced interpretations to the same test score as a function of group membership. The effect of this seems to be that selection unfairness can be evaluated, but once detected cannot be rectified in terms of the logic of the model that was used to detect it. Psychometrically this seems like an internal contradiction. If legislative thinking and psychometric rationality disagrees, should the latter challenge the former or should the legislative constraints simply be passively accepted as part of the rules that govern the manner in which the employment game is played? The argument presented in this paper seems to suggest that some unfortunate discrepancies between legislative thinking, specifically as expressed by the Employment Equity Act (Republic of South Africa, 1998), and psychometric theory also exist in South Africa. Moreover, too few South African psychometric scholars seem to be concerned about this. Questionably worded sections of the Act simply seem to have been passively accepted as part of the new rules that now govern the manner in which the employment game is to be played in the democratic South Africa.

Despite other possible flaws, the Employment Equity Act (Republic of South Africa, 1998) and the Promotion of Equality and Prevention of Unfair Discrimination Act (Republic of South Africa, 2000), however, fortunately seemingly still would permit human resource management professionals to follow the regression-based fairness models to their logical conclusion by attaching different criterion-referenced interpretations to the same test score if the validation data would require it. This position is, however, not generally held nor is it widely practiced in South Africa. It is moreover, ironically, that the practice of attaching different criterion-referenced interpretations to the same test score will most likely be opposed by many in South Africa as an unfair selection practice.

IN SEARCH OF SELECTION FAIRNESS; THE ROLE OF MEASUREMENT BIAS

Surely selection fairness cannot be achieved if the predictor is not free from measurement bias? The use of selection instruments that are biased against members of protected groups in the measurement of the underlying latent variable must surely unavoidably result in unfair discrimination against the members of those groups? Is this not the reasoning behind the Employment Equity Act's (Republic of South Africa, 1998) insistence that biased psychological tests may not be used to distinguish between, exclude or show preference for any applicant?

Bias unfortunately is an emotionally charged term (Humphreys, 1986) that has a negative connotation to it. It probably would not be incorrect to refer to measurement bias as a characteristic of an assessment instrument. It would, however, be more informative to interpret measurement bias (similarly to predictive bias) as a systematic, group-related error in the inferences made from obtained measures. In the case of measurement bias, however, the systematic, group-related error is not in the inferences made with regards to a criterion (or performance) construct (h) but rather with regards to the standing on the latent trait θ (or person construct ξ) being assessed by the selection instrument in question (Millsap & Everson, 1993). With regards to measurement bias (as opposed to predictive bias), a distinction needs to be made between scale bias, item bias and factorial bias (Drasgow & Hulin, 1990; Vandenberg & Lance, 2000).

Assume a continuous predictor scale X measuring a latent trait θ (or ξ) applied to members of two groups γ_1 (D = 0) and γ_2, (D = 1). Scale bias (or differential scale functioning) can be said to exist if $P[X \geq x_c|\theta = \theta_c; D = 0] \neq P[X \geq x_c|\theta = \theta_c; D = 1]$. Scale bias exists when the probability of achieving a specific observed score ($X \geq x_c$) differs for members of protected (D = 0) and non-protected (D = 1) groups when controlling for the latent trait (θ) being measured. Scale bias therefore exists when group membership (Γ) explains variance in the observed scale score X, either as a main effect or in interaction with the latent variable θ (or ξ), X is meant to reflect, which is not explained by that latent variable θ (Drasgow & Hulin, 1990; Millsap & Everson, 1993). Scale bias, therefore exists if the regression of the observed predictor score X on the latent variable θ (or ξ) differs across groups in terms of intercept (i.e. the expected observed score when $\theta = 0$) and/or slope. Item bias (or differential item functioning) would be defined similarly. Assume a dichotomous item X measuring a latent trait θ (or ξ) applied to members of two groups γ_1 (D = 0) and γ_2, (D = 1). Item bias can be said to exist if $P[X = x_c|\theta = \theta_c; D = 0] \neq P[X = x_c|\theta = \theta_c; D = 1]$. Item bias therefore exists when group membership (Γ) explains variance

in the observed item score X, either as a main effect or in interaction with the latent variable θ (or ξ), X is meant to reflect, which is not explained by that latent variable q (Millsap & Everson, 1993). Item bias, therefore exists if the (non-linear) regression of the observed item score X on the latent variable θ (or ξ) differs across groups in terms of intercept (i.e. the difficulty parameter b) and/or slope (i.e., the discrimination parameter a)[7] (Drasgow & Hulin, 1990; Drasgow & Parsons, 1983; Guion, 1998; Humphreys, 1986). Items are combined to determine an observed predictor scale score. The parameters of the scale or test characteristic curve (TSS) are determined by the parameters of item characteristic curves of the items comprising the scale (Guion, 1998). Criterion inferences are derived from the observed predictor scale scores and not individual item scores. The question thus firstly is how differential item functioning on the item level affects bias on the predictor scale level and secondly, if bias should exist on the predictor scale level, whether slope differences in the TCC would have a different effect on the regression of the criterion on the predictor than intercept (i.e., difficulty parameter) differences in the TCC? With regard to the first question there is evidence to suggest that in the United States, at least for cognitive tests, approximately half of differentially functioning items in a scale favour members of the non-protected group whereas the other half is biased against members of the non-protected group (Hunter & Schmidt, 2000; Society for Industrial and Organizational Psychology, 2003). The net effect is no scale bias. The situation locally is unknown.

If, however, scale bias would occur, it does not seem unreasonable to argue that the effect of group-related slope differences in the TCC should have a different effect on the regression of the criterion on the predictor than group-related intercept differences in the TCC[8]. Intercept differences in the TCC would imply that group significantly explains unique variance in the scale scores, not explained by the latent variable as a main effect. The observed predictor scale scores thus vary more (or less, depending on the nature of the latent means and the direction of the bias) than could be expected based only on the variance in the latent variable the scale is meant to reflect. The predictor scale means would therefore differ more (or less) than would have been the case if group had not explained unique variance in X. The movement in the observed predictor means should affect the intercept of the regression of the criterion on the predictor. More specifically it should create intercept differences, increase existing intercept differences or reduce intercept differences. Humphreys (1986) seems to agree. It moreover seems reasonable to argue that slope differences in the TCC would imply that group significantly explains unique variance in the scale scores, not explained by the latent variable as a group x predictor interaction effect. This would imply that the mean/expected observed scale score associated with a fixed latent trait level, increases at a differential rate for members of the protected and non-protected groups. This most probably would also have the effect of increasing observed predictor score variance. More importantly, however, since movement up the latent variable axis is associated with a differential rate of increase in X, differences in the scale discrimination parameter should affect the slope of the regression of the criterion on the predictor in addition to the intercept since it is the latent variable that ultimately determines the level of criterion performance achieved. Again Humphreys (1986) seems to have the same opinion.

If not properly accounted for in the selection decision rule, both forms of predictor scale bias could therefore have the effect of disadvantaging members of a specific group in that they would be positioned too low in the selection rank-order due to systematic group-related prediction errors. The systematic, group-related over- and under prediction of the criterion can, however, be removed by including group in the regression model as a main effect and/or a group x predictor interaction effect (although the scale bias itself would not thereby be removed). Again the assumption is that the criterion measures are reliable, valid and unbiased measures of the criterion construct. The inappropriate positioning of members of protected and non-protected groups in the selection rank order resulting from scale bias can therefore be corrected.

It, moreover, also seems reasonable to argue that the absence of predictor scale bias is no guarantee that discrimination in criterion-referenced selection cannot occur. Assuming a continuous scale X measuring a latent trait θ (or ξ) applied to members of two groups γ_1 and γ_2, a reliable and unbiased criterion measure Y determined (in part) by θ, it could still happen, even though $P(X \geq x_c|\theta = \theta_c; \Gamma = \gamma_1) = P(X \geq x_c|\theta = \theta_c; \Gamma = \gamma_2)$ (i.e., no scale bias), that $P(Y \geq Y_k|X = x_c; \Gamma = \gamma_1) \neq P(Y \geq Y_k|X = x_c; \Gamma = \gamma_2)$. Even though the latent predictor variable is measured without bias it should still in principle be possible that (predictive) bias could exist in the criterion inferences derived from the unbiased predictor measures. Predictive bias exists if the regression of the criterion on the predictor differs across protected and non-protected groups and this difference is not taken into account when deriving criterion estimates. This can easily happen even though no scale bias exists. This seems important since it would suggest that even if the Employment Equity Act (Republic of South Africa, 1998) would be successful in eradicating all forms of measurement bias it would thereby still not have succeeded in ensuring that selection decisions do not disadvantage members of specific groups.

It is consequently not quite clear why the Employment Equity Act (Republic of South Africa, 1998), in its effort to promote "equal opportunity and fair treatment in employment through the elimination of unfair discrimination" (Republic of South Africa, 1998, p. 12), would want to prohibit the use of scale biased psychological tests and other similar assessments (Republic of South Africa, 1998, p. 16). Ensuring that predictors are (predictively) valid and ensuring that predictors are free from item- and scale bias is neither necessary nor sufficient to ensure that the objective of the elimination of unfair discrimination will be reached. Neither will the presence of predictor scale bias necessarily nor unavoidably result in unfair criterion-referenced selection.

The argument presented earlier on the probability of eliminating predictive bias in judgmental decision rules again seems highly relevant here. When criterion inferences are derived clinically from predictor scale scores containing measurement bias, unfair discrimination most likely would occur. The unfair discrimination should, however, ultimately not be blamed on the scale bias existing in the predictor but rather on the inappropriate manner in which criterion inferences are derived from the predictor scale scores.

Factorial (or construct) bias refers to the extent to which the factor structure (Byrne, 1998) or measurement model (Diamantopoulos & Sigauw, 2000; Mels, 2003) is invariant across groups. Factorial equivalence (Byrne, 1998) would be demonstrated if the parameters constituting the measurement model would remain the same across groups. More specifically factorial equivalence (Byrne, 1998) would be demonstrated if (a) the same number of latent dimension(s) are required to explain the covariances observed amongst the items comprising the tests, (b) the loadings of the items on their designated latent dimensions (Λ_X) are invariant across groups, (c) the intercept of the regression of the item scores on the latent variables (τ_X) are invariant across groups, (d) the correlations amongst the latent dimensions are invariant across groups, and possibly, although this might be considered an overly stringent requirement (Byrne, 1998), (e) the measurement error variances and covariances are invariant across groups. In short, factorial equivalence would be indicated if the factor loading matrix (Λ_X), factor correlation

7 From a structural equation modelling perspective, uniform and non-uniform item bias could be said to exist if the vector of intercept parameters τ_X and the factor-loading matrix Λ_X of slope parameters differ across groups (Vandenberg & Lance, 2000). **8** The ideal would be to beyond the speculative verbal arguments presented here and to eventually develop an analytical understanding of the manner in which differences in the TCC parameters affect the regression of the criterion on the predictor.

matrix (Φ) and the variance-covariance matrix of measurement error terms (Θ_δ) and the vector of intercept terms of the regression of the observed item scores on the underlying latent variables (τ_X) (Byrne, 1998; Diamantopoulos & Sigauw, 2000; Vandenberg & Lance, 2000) are invariant across groups.

The important but seemingly neglected question is what the consequences of significant differences in these matrices, individually and collectively, across groups are for the regression of the criterion on the predictor? The previously cited measurement equivalence studies in South Africa do not seem to analyse the relationship between construct bias and equity in any great depth but rather seem to simply accept that lack of structural equivalence in any form one way or another will result in discriminatory selection practices. It probably would be safe to argue that if major differences exist in Λ_X across groups, both in terms of number of factors and factor loadings, that significant differences in predictive validity would probably exist across groups and therefore most likely also significant slope differences. This, however, seems an unlikely event, since it appears to be generally accepted, in the United States at least, that both single group validity and differential validity occur no more than could be expected by chance (Bartlett et al., 1978; Schmidt & Hunter, 1981; Schmitt, 1989). Nonetheless, the Employment Equity Act (Republic of South Africa, 1998) probably would be correct in prohibiting this extreme form of construct bias. The Employment Equity Act (Republic of South Africa, 1998), however, is wrong in as far as it implies that the absence of factorial bias will ensure that discrimination in criterion-referenced selection cannot occur. What the effect of minor, albeit significant differences in factor loadings, phi coefficients or error variances on the regression of the criterion on the predictor might be is not clear. Could variance in the measurement model parameters across groups, apart from the possibility mentioned above, affect the regression of the criterion on the predictor in such a manner that it would preclude the possibility of adapting the prediction model in a way that would prevent group-related prediction errors?

The foregoing is a plea to refrain from motivating research on measurement bias in terms of the simplistic premise that it will necessarily promote "equal opportunity and fair treatment in employment through the elimination of unfair discrimination" (Republic of South Africa, 1998, p. 12). The foregoing argument should not be construed as a plea that bias analysis should not be performed. Although the most recent edition of the Principles (Society for Industrial and Organizational Psychology, 2003) seems rather indifferent towards differential item functioning research in the personnel selection domain, this type of research should nonetheless be regarded as indispensable in the development of both predictor and criterion measures. In the personnel selection domain, hypotheses are developed on the nature of the latent person variables that determine job performance (Guion, 1991; 1998; Landy, 1986). In these hypothesised relationships lies the possibility of estimating job performance. In pursuit of this possibility instruments are subsequently developed (or chosen) to measure these constructs as defined amongst all members of the applicant population. Despite the fact that the measurement of these latent traits is not an objective in and by itself but rather one phase in a larger process, every effort should nonetheless be made to see to it that these instruments do provide reliable, valid and unbiased measures of their target constructs because that is what they were commissioned to do at that stage of the process. The fact that later stages in the process could be adapted to accommodate some of the failures in earlier stages should never be used as an excuse to condone careless test construction[9]. Measurement bias therefore can and should as far as possible be avoided through the judicious choice of properly developed selection instruments. In doing so, however, the danger of systematically disadvantaging members of specific groups in personnel selection would not necessarily have been neutralised.

Although easier said than done (Guion, 1998) measurement bias analysis with regards to the criterion is critically important if valid and credible validity, fairness and utility analyses results are desired (Schmitt, 1989). If measurement bias in the criterion against protected groups is not detected and removed prior to the validity, fairness and utility analyses, unfair discrimination will be invisible and irreversibly built into the selection decision rule.

IN SEARCH OF MINIMUM ADVERSE IMPACT

Adverse impact in personnel selection occurs when a specific selection strategy affords members of a specific group a lower likelihood to be selected than members of another group. Adverse impact is indicated when there is a substantial difference in the selection ratios of groups that work to the disadvantage of members belonging to a certain group (Guion, 1991; 1998). A selection ratio for any group, which is less than four-fifths (4/5) or 80 percent of the ratio of the group with the highest selection ratio would typically be regarded as evidence of adverse impact (Huysamen, 1996; Maxwell & Arvey, 1993). The four-fifth rule is normally interpreted with reference to the predictor distributions (Arvey & Faley, 1988; Guion, 1991; 1998; Hough, Oswald & Ployhart., 2001; Sackett & Ellingson, 1997; Sackett & Wilk, 1994) In the conceptualisation of adverse impact it is, however, critically important to appreciate that the selection ratios for the various groups should ultimately be determined by their expected criterion performance conditional on their test performance (derived fairly, i.e., without systematic prediction bias) and not the selection ratios that would have resulted if selection would have occurred top-down on the predictor. The Maxwell and Arvey (1993) position that the standardised difference on the predictor between protected and non-protected groups should serve as an index of adverse impact therefore is highly questionable[10]. The standardised difference on the criterion (or expected criterion) between protected and non-protected groups should rather serve as an index of adverse impact. The criterion construct is the focus of interest in selection decisions. Predictor measures should be interpreted in terms of expected/predicted criterion performance in personnel selection. Since selection decisions are based on rank ordered expected criterion performance, the selection ratios in question should therefore be calculated on the $E[Y|X_i;D_i]$ distribution. The question thus is whether the selection ratio's based on the predicted criterion performance ($E[Y|X_i; D_i]$), derived fairly via moderated regression analysis from the predictor measures X_i, differ for protected and non-protected groups.

Adverse impact in and by itself does not constitute discrimination. In employment litigation in the United States of America adverse impact is used to make a *prima facie* case for discrimination[11]. Once established, the burden of proof shifts to the defendant (Arvey & Faley, 1988; Dupper, 2002; Guion, 1991). If adverse impact is shown, the burden of proof shifts to the employer to demonstrate the job-relatedness of the selection procedure and that the inferences derived from the predictor scores are fair. Alternatively, the employer could show that no equally valid alternative, with less adverse impact,

9 Differential reliability across protected and non-protected groups could quite possibly be the most prevalent fatal test construction failure in South Africa because in its extreme form it would render validity, item bias, scale bias, measurement and structural equivalence and predictive bias analyses highly questionable. The development of insightful diagnostic hypotheses, derived from measurement theory, however, seems to be a critical prerequisite that needs to be satisfied if local test development initiatives would want to overcome the differential reliability problem quite often found with imported psychometric tests. **10** It only makes sense to do so if selection decisions were inappropriately directly based on predictor scores instead of the expected criterion scores conditional on the predictor scores. Although equity legislation in the United States prohibits differential score interpretation it does not prohibit criterion-referenced predictor interpretation as such. **11** Adverse impact defined in terms of the criterion is, however, not a necessary condition for unfair (criterion-referenced) personnel selection to exist. If, for example the mean criterion performance of protected and unprotected groups would differ significantly but the predictor distributions would coincide, selection based on the predictor scores or based on the regression of the criterion on the predictor would disadvantage the members of the group scoring higher on the criterion. This *de facto* discriminatory procedure, however, seemingly is immune against litigation since it does not create adverse impact and therefore leaves no *prima facie* evidence of discrimination. Moreover, this illustrates the potential danger of trying to ameliorate adverse impact (Hough et al., 2001) by focusing on strategies for reducing subgroup mean differences in the predictor.

exists. Even though the use of the latter line of defence is quite widely advocated (Arvey & Faley, 1988; Cook, 1998; Equal Employment Opportunity Commission, 1978; Gatewood & Feild, 1994; Guion, 1991; Maxwell & Arvey, 1993), it nonetheless seems highly questionable. The remedy proposed by the Uniform Guidelines only makes sense if adverse impact is defined in terms of the predictor distributions. This in turn would make sense if selection decisions would be based on inferences regarding predictor constructs derived from predictor scores. Selection decisions should, however, not be based on predictor construct inferences but should rather be based on criterion estimates derived from the predictor. This is clearly signalled by the APA sanctioned interpretation of predictive validity as the permissibility of *criterion inferences* derived from test scores (Society for Industrial and Organizational Psychology, 2003). The regression-based interpretations of selection fairness (Cleary, 1968; Einhorn & Bass, 1971; Huysamen, 1996; Huysamen, 2002) favoured by the Principles (Society for Industrial and Organizational Psychology, 2003) and the South African Guidelines (Society for Industrial Psychology, 1998) moreover also explicitly reflects the assumption that selection decisions are based on criterion estimates derived from the predictor. In the final analysis the cause of adverse impact in personnel selection therefore resides in systematic differences in criterion distributions. To deny this would be to deny the logic underlying predictive validity and the regression-based interpretations of selection fairness. The ratio of the selection ratio of the protected group to that of the non-protected group (SR[P]/SR[NP]) will necessarily be less than unity in a strict top-down selection strategy based on $E[Y|X_i; D_i]$, to the extent that the mean criterion performance of the protected and non-protected groups differ ($\mu_{YP} < \mu_{YNP}$). Adverse impact in criterion referenced personnel selection can therefore not be avoided by the judicious choice of selection instruments (Huysamen, 1996; Schmidt & Hunter, 1981). Nor can selection instruments be graded in terms of the degree of their adverse impact. Not even an omniscient but "meritocratic" decision-maker would be able to avoid (fair) adverse impact if the mean criterion performance of the protected and non-protected groups differ (i.e., if $\mu_{YP} < \mu_{YN}$). If adverse impact occurs because of differences in predictor performance across groups but which cannot be justified in terms of differences in criterion performance, it would imply that the criterion inferences derived from such test scores are biased (i.e., the selection decision-making is unfair in the Cleary sense of the term). This type of unfair/discriminatory adverse impact can be avoided, however, by eliminating the systematic, grouprelated prediction error. As Schmitt (1989, p. 138) appropriately remarks:

> ... the presence of subgroup mean differences on selection tests is not terribly important if we adopt Cleary's definition of fair test use.

How this stance links up with his subsequent predictor-focused search for strategies to reduce adverse impact (Sackett, Schmitt, Ellingson & Kabin, 2001; Schmitt, Rogers, Chan, Sheppard & Jennings, 1997) is not clear. The results reported by Sackett and Ellingson (1997) on the protected group selection ratio relative to the non-protected group selection ratio for various standardised group differences in mean predictor performance (d) should therefore still be relevant provided that d is now interpreted with reference to the distributions of expected criterion performance rather than the predictor distributions. Their results on the four-fifths ratios for specific non-protected group selection ratios and values of d should therefore also still be relevant again provided that d is interpreted with reference to the distributions of expected criterion performance.

The foregoing argument can be illustrated (rather than formally proven) in terms of the following fictitious dataset (N = 400) comprising a normally distributed criterion (Crit_Y) systematically related to a normally distributed predictor (Pred_X[12]). Half the observations are obtained from members of a protected group (D = 0) and half from members of a non-protected group (D = 1). The criterion distributions of the two groups coincide perfectly as shown in Table 1. Scores on the two variables were generated in SPSS (SPSS, 2005) utilising the normal density function.

TABLE 1
DESCRIPTIVE STAISTICS IN RESPECT OF THE PREDICTOR AND CRITERION DISTRIBUTIONS OF PROTECTED AND NON-PROTECTED GROUPS

	Predictor (Pred_X)			Criterion (Crit_Y)		
D = 0	N	Valid	200	D=0 N	Valid	200
		Missing	0		Missing	0
	Mean		49,039	Mean		146,474
	Median		48,753	Median		146,702
	Mode		19,370	Mode		43,420
	Std. Deviation		10,267	Std. Deviation		33,0796
	Variance		105,403	Variance		1094,259
	Skewness		-0,055	Skewness		0,071
	Std. Error of Skewness		0,172	Std. Error of Skewness		0,172
	Kurtosis		-0,272	Kurtosis		0,109
	Std. Error of Kurtosis		0,342	Std. Error of Kurtosis		0,342
D = 1	N	Valid	200	D=1 N	Valid	200
		Missing	0		Missing	0
	Mean		64,039	Mean		146,475
	Median		63,753	Median		146,702
	Mode		34,370	Mode		43,42
	Std. Deviation		10,267	Std. Deviation		33,0796
	Variance		105,403	Variance		1094,259
	Skewness		-0,055	Skewness		0,071
	Std. Error of Skewness		0,172	Std. Error of Skewness		0,172
	Kurtosis		-0,272	Kurtosis		0,109
	Std. Error of Kurtosis		0,342	Std. Error of Kurtosis		0,342

The predictor distributions, however, differ in terms of location only as indicated in Table 1. A standardised difference in mean predictor performance of d = 1,461 thus exists in this case. The standardized difference is obtained by subtracting the mean predictor score of the protected group (D = 0) from the mean predictor score of the non-protected group (D = 1) and dividing by the within-group standard deviation (Sackett & Ellingson, 1997). The predictor-criterion correlation is 0,743 (p < 0,01) in both groups as shown in Table 2.

TABLE 2
WITHIN-GROUP PREDICTOR-CRITERION CORRELATIONS (N=200)

GROUP			PRED_X	CRIT_Y
D = 0	PRED_X	Pearson Correlation	1	0,743
		Sig. (1-tailed)	.	0,000
	CRIT_Y	Pearson Correlation	0,743	1
		Sig. (1-tailed)	0,000	.
D = 1	PRED_X	Pearson Correlation	1	0,743
		Sig. (1-tailed)	.	0,000
	CRIT_Y	Pearson Correlation	0,743	1
		Sig. (1-tailed)	0,000	.

12 The variable names Crit_Y and Pred_X were arbitrarily chosen to represent the observed criterion and predictor variables in the SPSS analysis.

When selection occurs strict top-down based on the predictor or based on the estimated criterion performance derived from the regression of Crit_Y on Pred_X, serious adverse impact results against the members of the protected group (D = 0). Table 3 depicts the selection ratios for the two groups that would result from an overall selection ratio of 0,20. The ratio of the proportion of selectees from the protected group to the proportion of selectees from the non-protected group amounts to 0,06666, which clearly fails to meet the four-fifths requirement of the Uniform Guidelines. These findings agree with the results Sackett and Ellingson (1997, pp. 710 & 712) report on the effect of mean predictor differences on the selection ratio of the protected group.

TABLE 3
CROSS TABULATION OF GROUP MEMBERSHIP AGAINST SELECTION DECISION

			Decision		Total
			Reject	Accept	
GROUP	D = 0	Count	195	5	200
		% within GROUP	97,5%	2,5%	100,0%
		% within Decision	60,9%	6,3%	50,0%
		% of Total	48,8%	1,3%	50,0%
	D = 1	Count	125	75	200
		% within GROUP	62,5%	37,5%	100,0%
		% within Decision	39,1%	93,8%	50,0%
		% of Total	31,3%	18,8%	50,0%
Total		Count	320	80	400
		% within GROUP	80,0%	20,0%	100,0%
		% within Decision	100,0%	100,0%	100,0%
		% of Total	80,0%	20,0%	100,0%

The adverse impact created against the protected group would be considered unfair by the Cleary-model of selection fairness (Cleary, 1968) because group membership significantly ($p < 0,01$) explains variance in the criterion, which is not explained by the predictor, but the current selection strategy fails to take this fact into account. This results in the significant underprediction of the criterion performance of the members of the protected group. The selection decision-rule will therefore discriminate against members of the protected group by placing them too low in the selection rank order. This is shown in Table 4 and Table 5 and illustrated in Figure 1 and Figure 2.

TABLE4
DIFFERENCE IN MEAN UNSTANDARDISED (Y-E[Y|X]) BETWEEN PROTECTED (D = 0) AND NON-PROTECTED (D = 1) GROUPS

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
D = 0	200	11,681	23,757	1,6798	8,368	14,993	-49,268	73,959
D = 1	200	-11,681	23,757	1,6798	-14,993	-8,368	-72,629	50,598
Total	400	,000	26,452	1,323	-2,600	2,600	-72,629	73,959

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	54573,367	1	54573,367	96,698	0,000
Within Groups	224618,975	398	564,369		
Total	279192,342	399			

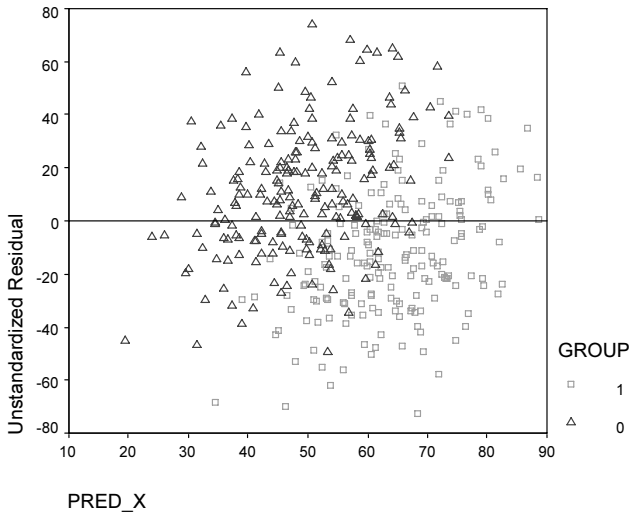

Figure 1: Scatter plot of the unstandardised residuals against the predictor with group as a plot symbol

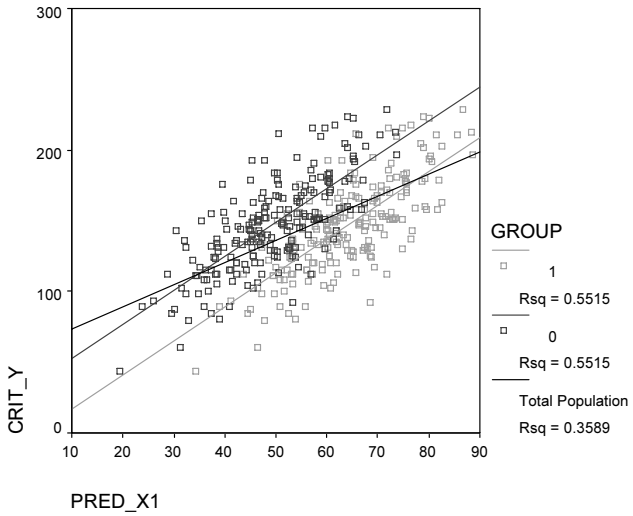

Figure 2: Scatter plot of the criterion-predictor relationship with group membership as plot symbol

The remedy would be to include Group as a main effect in the prediction model. The regression of Crit_Y on Pred_X and Group is shown in Table 5.

TABLE 5
MULTIPLE REGRESSION OF THE CRITERION ON A LINEAR PREDICTOR-GROUP COMPOSITE

R	R Square	Adjusted R Square	Std. Error of the Estimate
0,743	0,551	0,549	22,183

Predictors: (Constant), GROUP, PRED_X
Dependent Variable: CRIT_Y

	Sum of Squares	df	Mean Square	F	Sig.
Regression	240166,655	2	120083,327	244,041	0,000
Residual	195348,530	397	492,062		
Total	435515,185	399			

	UnstandardiSed Coefficients		Standardised Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	29,140	5,538		5,262	0,000
PRED_X1	2,393	0,108	0,920	22,093	0,000
GROUP	-35,891	2,750	-0,544	-13,053	0,000

When regressing Crit_Y on Pred_X, only 0,359 of the variance in the criterion is explained by the predictor or E[Crit_Y|Pred_X] whereas within groups Pred_X explains 0,551 of the variance in Crit_Y (see Figure 2). On the other hand, when regressing Crit_Y on Pred_X and Group, 0,551 of the variance in the criterion is explained by the linear composite of Pred_X and Group or E[Crit_Y|Pred_X; Group]. The multiple correlation between the criterion and the weighted linear composite of the predictor and the group variable is therefore 0,734 (i.e., R[E[Crit_Y|Pred_X; Group],Crit_Y] = 0,734) (see Table 5). By taking group membership into account in the prediction model the systematic group-related under- and over-prediction of criterion performance is eliminated and as a consequence the proportion of criterion variance explained is increased.

When selection occurs strict top-down based on the estimated criterion performance derived from the regression of Crit_Y on Pred_X and Group, adverse impact no longer result against the members of the protected group (D = 0). Table 6 depicts the selection ratios for the two groups that would result from an overall selection ratio of 0,20. The ratio of the proportion of selectees from the protected group to the proportion of selectees from the non-protected group amounts to 1,0, which constitutes perfect compliance with the requirement of the Uniform Guidelines. The fair use of the predictor (in the Cleary sense of the term) totally eliminated adverse impact in this case because the criterion distributions coincide. It could easily be demonstrated that if the criterion distributions had differed in terms of location, the fair use of the predictor would have resulted in fair, acceptable (AERA, APA & NCME, 1999; Huysamen, 1996; Huysamen, 2001) adverse impact.

TABLE 6
CROSS TABULATION OF GROUP MEMBERSHIP AGAINST SELECTION DECISION

			Decision		Total
			0,00	1,00	
GROUP	D = 0	Count	160	40	200
		% within GROUP	80,0%	20,0%	100,0%
		% within Decision	50,0%	50,0%	50,0%
		% of Total	40,0%	10,0%	50,0%
	D = 1	Count	160	40	200
		% within GROUP	80,0%	20,0%	100,0%
		% within Decision	50,0%	50,0%	50,0%
		% of Total	40,0%	10,0%	50,0%
Total		Count	320	80	400
		% within GROUP	80,0%	20,0%	100,0%
		% within Decision	100,0%	100,0%	100,0%
		% of Total	80,0%	20,0%	100,0%

Developing a clear and unambiguous stance on the meaning of adverse impact seems to be important from a South African perspective since the Employment Equity Act (Republic of South Africa, 1998) and the Promotion of Equality and Prevention of Unfair Discrimination Act (Republic of South Africa, 2000) also seem to assume a shifting burden of persuasion model (Arvey and Faley, 1988; Dupper, 2002). In Chapter II of the Employment Equity Act (Republic of South Africa, 1998, p. 16), under the heading "Burden of proof", paragraph 11 states:

> Whenever unfair discrimination is alleged in terms of this Act, the employer against whom the allegation is made must establish that it is fair.

In Chapter 3 of the Promotion of Equality and Prevention of Unfair Discrimination Act (Republic of South Africa, 2000, p. 8), again under the heading "Burden of proof", paragraph 13 states:

1. If the complainant makes out a *prima facie* case of discrimination[13].
   a) the respondent must prove, on the facts before the court, that the discrimination did not take place as alleged: or
   b) the respondent must prove that the conduct is not based on one or more of the prohibited grounds
2. If the discrimination did take place-
   a) on a ground in paragraph (a) of the definition of "prohibited grounds" then it is unfair, unless the respondent proves that the discrimination is fair;
   b) on a ground in paragraph (b) of the definition of "prohibited grounds" then it is unfair-
      i) if one or more of the conditions set out in paragraph (b) of the definition of "prohibited grounds"[14] is established; and
      ii) unless the respondent proves that the discrimination is fair.

The rather intricate nature of the Promotion of Equality and Prevention of Unfair Discrimination Act's (Republic of South Africa, 2000) position of the burden of persuasion resting on the defendant/respondent further underlines the necessity of clarifying in practical terms exactly how a *prima facie* case of (indirect) discrimination will be established. In the case of both acts the question moreover arises how the respondent can prove that a selection procedure that discriminates against individuals from a protected group (i.e., the procedure imposes a burden or disadvantage on such members or it withholds opportunities from them reflected in a lower probability of being selected) is in fact fair? Clarity on neither of these two issues seems to have been reached in the legal fraternity in South Africa (Bonthuys, 2002; Dupper, 2002; Landman, 2002).

Personnel selection procedures would nonetheless want to minimise adverse impact, not only in order to avoid litigation, but to ensure that access to job opportunities are distributed across groups in the labour market in proportion to the size of the various groupings and to optimally utilise the human recourses available in the labour market. In an ideal world one would want to share job opportunities amongst protected and non-protected groups in proportion to their presence in the labour market. It should also be acknowledged that organisations face the very real demand to increase the diversity of their workforce so as to mirror the composition of the community more closely (Sackett et al., 2001). The same is true for institutions of higher learning with regards to the composition of their student bodies.

When the criterion distributions of protected and non-protected groups coincide, it is possible to use a valid predictor fairly to maximise the utility of the selection procedure while avoiding adverse impact. However, when systematic differences in the criterion distributions exist it no longer is possible to achieve all four objectives simultaneously. If selection decisions are fair in terms of the Cleary-interpretation of fairness and selection occurs strictly top-down based on $E[Y|X_1; D_i]$, then utility will be maximised, but adverse impact will now be unavoidable. The objective of minimising adverse impact could be satisfied through quotas or criterion referenced race norming, but only if the utility objective is sacrificed. The sacrifice required by top-down hiring within each group (criterion-referenced race norming) would depend on the magnitude of the difference in the criterion distributions. According to Schmidt and Hunter (1981, p. 1130):

> ... selection systems based on top-down hiring within each group completely eliminates "adverse impact" at a much smaller price in lowered productivity. Such systems typically yield 85% to 95% of the productivity gains attainable with optimal nonpreferential use of selection tests.

13 The definition of discrimination held by the Promotion of Equality and Prevention of Unfair Discrimination Act (Republic of South Africa, 2000) was quoted earlier in the manuscript. **14** "prohibited grounds" are – (a) race, gender, sex, pregnancy, marital status, ethnic or social origin, colour, sexual orientation, age, disability, religion, conscience, belief, culture, language and birth; or (b) any other grounds where discrimination based on that other ground – (i) causes or perpetuates systemic disadvantage; (ii) undermines human dignity; or (iii) adversely affects the equal enjoyment of a person's rights and freedoms in a serious manner that is comparable to discrimination on a ground in paragraph (a) (Republic of South Africa, 2000, p. 5).

Meta-analytic summaries of criterion differences in the United States indicate a 0,30 standard deviation difference in mean protected and non-protected group criterion performance (Sackett & Roth, 1996). To the extent that similar conditions would exist in South Africa criterion-referenced race norming presents itself as a viable strategy to combat adverse impact. Three considerations, however, argue against a blind reliance on within-group top-down selection. A drop in utility of 5% to 15% can be substantial when projected over number of selectees, time and successive cohorts (Boudreau, 1991). More importantly, however, to solely rely on within-group top-down selection would leave the root causes of the performance imbalance, which fundamentally underlies adverse impact, untreated. Moreover, the difference in mean criterion performance amongst protected and non-protected groups in South Africa could be substantially greater than in the United States. Criterion-referenced race norming under these conditions would result in a more severe drop in utility than anticipated by Schmidt and Hunter (1981).

Increasing the weights of the work performance dimensions less susceptible to ethnic or gender differences and decreasing the weights associated with dimensions on which larger differences exist would also reduce adverse impact on the composite criterion (De Corte, 1999; Hattrup, Rock & Scalia, 1997). The weighing of performance dimensions should, however, only reflect the relative importance of the various competencies in achieving the objective for which the job exists. The manipulation of criterion composite weights, therefore, does not offer a meaningful solution to the problem of adverse impact (Sackett et al., 2001).

The realisation that adverse impact in criterion referenced personnel selection cannot be avoided by the judicious choice of selection instruments is by no means a novel insight. Twenty-three years ago Schmidt and Hunter (1981, pp.1131 & 1134) already declared:

> These findings show that tests do not cause "adverse impact" against minorities. The cumulative research on test fairness shows that the average ability and cognitive skill differences between groups are directly reflected in job performance and thus are real. They are not created by tests. ... But the solution to the problem (of adverse impact) cannot begin until the problem is faced in an intellectually honest way. It is not intellectually honest, in the face of empirical evidence to the contrary, to postulate that the problem is biased and/or invalid employment tests.

Although it would not be intellectually honest to ultimately attribute the problem of adverse impact on biased selection instruments and/or unfair selection decision-making (Schmidt & Hunter, 1981) and although performance can be maximised fairly (within the current reality) despite adverse impact, the problem of adverse impact can nonetheless not simply be ignored. How the human resource function should respond to the problem of adverse impact in selection would depend on why the systematic differences in criterion distributions exist. This is a question that is not raised often enough by human resource management professionals when contemplating the appropriate response to the dilemma outlined above. This question is, however, critically important since remedial actions will only succeed if they deal with the root cause of the problem. In the South African context it does not seem unreasonable to attribute at least some part of the systematic group-related differences in criterion distributions to a socio-political system that systematically denied the members of specific groups the opportunity to develop and acquire those crystallised abilities required to succeed on the criterion. Psychological tests that report standardised mean score differences between ethnic groups on especially measures of cognitive abilities should therefore not be characterised as villains responsible for the problem but rather as unbiased messengers relatively accurately conveying the consequences of a tragic social system. The solution therefore is not to be found in strategies to convince the messenger to alter its message as is seemingly suggested by Hough et al. (2001) and Sackett et al. (2001). The difference in criterion distributions observed between protected and non-protected groups reflect *bona fide* differences on numerous critical dispositions and attainments (Schmidt & Hunter, 1981; Saville & Holdsworth, 2000; 2001) required to succeed in the world of work, which have resulted from the systemic denial of access to developmental opportunities. To deny the criterion differences and the differences in the underlying competency potential (Saville & Holdsworth 2000; 2001) is to deny the history that caused it. The solution rather lies in affirmative development interventions aimed at developing those attainments and dispositions needed to succeed on the criterion. This puts the assessment of learning potential centre-stage.

SUMMARY

The objective of personnel selection is to add value to organisations by maximising the performance of employees by regulating the quality of employees moving into, up and out of the organisation. The criterion construct is therefore the focus of interest in personnel selection. Direct information on the criterion construct is, however, not available at the time of the selection decision. Selection decisions are therefore based on expected criterion performance or the conditional probability of success. Such decision-making can be considered fair to the extent that members of protected and non-protected groups with the same probability of success on the job have the same probability of obtaining the job. This will be the case to the extent to which there is no systematic group-related (prediction) bias in the expected criterion performance or the conditional probability of success. Selection fairness therefore cannot be assured solely through the careful development or judicious choice of selection instruments. Measurement bias can be avoided through the careful development or judicious choice of selection instruments. Unfair discrimination in personnel selection, however, cannot be avoided through the use of reliable, valid and (scale) unbiased selection instruments. Fair (i.e., non-discriminatory) selection can in the final analysis only be assured by determining whether group membership systematically affects any of the parameters defining the regression of the criterion on the predictors and appropriately accounting for the group effect in the selection decision rule. Assessment techniques for this reason also cannot be certified as Employment Equity Act (Republic of South Africa, 1998) compliant. Adverse impact, finally, cannot be avoided through the careful development or judicious choice of selection instruments. Selection instruments cannot be graded in terms of the degree of their adverse impact. In the final analysis, adverse impact resides in differences in the criterion distributions of protected and non-protected groups. Adverse impact cannot be equated with unfair discrimination. In as far as unfair discrimination most likely (although not necessarily) will result in adverse impact, the latter can be regarded as prima facie evidence of unfair discrimination. Adverse impact will most likely result from fair selection procedures in South Africa if a strict top-down selection strategy is followed because of systematic differences in the criterion distributions of protected and non-protected groups. Organisations in South Africa can (and probably in the interim have to) choose to avoid adverse impact through quotas because they value work force diversity more than the drop in utility produced by the deviation from strict top-down selection based on fairly derived expected criterion performance. In a country like South Africa where the difference in average criterion performance (i.e. adverse impact) is a legacy of an artificial socio-political situation, it would, however, be a pity not also to address the fundamental causes underlying adverse impact. In the final analysis it is the differences in developmental opportunity and the resultant

differences in the attainments and dispositions that drive performance that should be dealt with. Aggressive investment in affirmative development interventions seems the only truly honest (Schmidt & Hunter, 1981) way of dealing with the labour market legacy of our previous political dispensation. This will present numerous exiting and stimulating challenges to the I/O psychology fraternity in South Africa. First amongst these would probably be to develop a comprehensive performance@learning structural model (Saville & Holdsworth, 2000; 2001) that explicates the manner in which critical learning dispositions and attainments map onto critical learning competencies (Taylor, 1994) and how these in turn relate to job performance dispositions and attainments and ultimately job competencies. Deriving an appropriate affirmative development selection battery from the model to identify those previously disadvantaged individuals that would maximally benefit from an affirmative development opportunity seems to present a second important challenge (Taylor, 1994). Deriving appropriate interventions from the model aimed at maximising transfer of training probably represents a third critical challenge. Additional challenges with regards to training content, learning strategies and training delivery also exist.

The broad psychometric position in which the predictor is the primary villain responsible for most if not all of the evils associated with personnel selection from a diverse applicant pool is therefore not a psychometrically justified one that best serves the interests of all stakeholders involved. More to the point, it will not assure that the commendable vision formulated by then president Mandela in the preamble to the Employment Equity Bill (Republic of South Africa, 1996) will be achieved. It is, moreover, probably a very natural psychological reaction to target an explicit scapegoat to be blamed and sacrificed for the selection sins committed during the pre-equity legislation era in South Africa. However, when the ill-fated scapegoat is erroneously being perceived as the true culprit without any honest confession on the part of the real sinner, more harm is being done than good. It is the decision-maker who must shoulder the final responsibility for what went wrong in the past and for complying with the spirit and the letter of the Employment Equity Act (Republic of South Africa, 1998) in future. And on a more personal note, it is me who must ask myself why I had so little to say about employment equity before it was forced upon me by the newly written Constitution and the legislation that was enacted in terms of it, despite the extensive available literature on the topic (e.g. Bartlett et al., 1978; Cleary, 1968; EEOC, 1978; Petersen & Novick, 1976).

PRACTICAL IMPLICATIONS

It is crucial that human resource management professionals involved in personnel selection should move beyond the popular rhetoric on the use of psychological tests in personnel selection and engage in an open (Louw, 1965), honest and penetrating debate on the interplay between past injustices, measurement bias, selection fairness, adverse impact and selection utility. However, open, honest and penetrating debate, in and by itself, will not achieve the extremely laudable vision formulated by former president Mandela in the preamble to the Employment Equity Bill (Republic of South Africa, 1996, p. 5). The courage to act on the convictions emerging from the debate is what will ultimately bring us closer to realising the vision.

The argument presented above, and the approach to practical psychological assessment it implies, could be criticised as unrealistically empirical and actuarial. Undeniably the approach advocated here would pose severe practical, technical and logistical challenges to the human resource management professional. However, if there is some psychometric merit in the argument outlined above, could the Industrial-Organisational Psychology and Psychology fraternities not rise to the challenge of finding creative and innovative solutions to the obstacles that currently prevent the widespread implementation of an actuarial approach to personnel selection (Mossholder & Arvey, 1984)? The development of a generic individual performance structural model and an accompanying individual performance index, analogous to the Theron, Spangenberg and Henning (2004) unit performance structural model and Performance Index (Spangenberg & Theron, 2004) in conjunction with synthetic validity procedures (Guion, 1998; Mossholder & Arvey, 1984), cross-industry cooperation, validity generalisation analysis (Schmidt & Hunter, 1977) and possibly bootstrapping procedures (Efron & Tibshirani, 1993) could be explored as possible solutions.

REFERENCES

Abrahams, F. & Mauer, R. (1999). The comparability of the constructs of the 16PF in the South African context. *Journal of Industrial Psychology*, *25*, 53-59.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington D.C, NY: Author.

Arvey, R.D. & Faley, R.H. (1988). *Fairness in selecting employees* (Second edition). Reading, MA: Addison-Wesley.

Austin, J.T. & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology*, *77*, 836-874.

Bartlett, C.J., Bobko, P., Mosier, S.B. & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: an alternative to differential analysis. *Personnel Psychology*, *31*, 233-242.

Bartram, D., Baron, H. & Kurz, R. (2003). *Let's turn validation on its head*. Occupational Psychology Conference of the British Psychological Society Bournemouth.

Berenson, M.L, Levine, D.M & Goldstein, M. (1983). *Intermediate statistical methods and applications*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Binning, J.F. & Barrett, G.V. (1989). Validity of Personnel decisions: a conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, *74*, 478-494.

Bonthuys, E. (2002). Counting flying pigs: psychometric testing and the law. *Industrial Law Journal*, *23*, 1175-1194.

Boudreau, J.W. (1989). *Selection utility analysis: a review and agenda for further research*. In M. Smith & I. Robertson (Eds.). Advances in selection and assessment. Chichester: John Wiley.

Boudreau, J.W. (1991). Utility analysis for decisions in human resource management. In M.D. Dunnette & L.M. Hough (Eds.). *Handbook of industrial and organizational psychology* (Second edition; Volume 2). Palo Alto, CA: Consulting Psychologists Press.

Bredell, B., van Eeden, R. & van Staden, F. (1999). Culture as a moderator variable in psychological test performance: issues and trends in South Africa. *Journal of Industrial Psychology*, *25*, 1-7.

Brogden, H.E. (1946). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Education and Psychology*, *37*, 65-76.

Brogden, H.E. (1949a). When testing pays off. *Personnel Psychology*, *2*, 171-185.

Brogden, H.E. (1949b). A new coefficient: application to biserial correlation and to estimation of selective efficiency. *Psychometrika*, *14*, 169-182.

Byrne, B.M. (1998). *Structural equation modelling with LISREL, PRESLIS and SIMPLIS: basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Earlbaum.

Cascio, W.F. (1991a). *Applied psychology in personnel management*. Englewood Cliffs, N.J: Prentice-Hall.

Cascio, W.F. (1991b). *Costing human resources; the financial impact of behavior in organizations*. Boston, MA: PWS-Kent Publishing Company.

Chan, D. & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgement tests: sub-group differences in test performance and face validity perceptions. *Journal of Applied Psychology*, *82*, 143-159.

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

Cochran, W.G. (1951). Improvement by means of selection. In J. Neyman (Ed.), *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (pp.449-470). Berkeley, CA: University of California Press.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, Ill: University of Illinois Press.

Cronshaw, S.F. & Alexander, R.A. (1985). One answer to the demand for accountability: selection utility as an investment decision. *Organizational Behavior and Human Decision Processes*, *35*, 102-118.

Dawes, R.M. (1971). A case study of graduate admissions: application of three principles of human decision making. *American Psychologist*, *26*, 180-188.

Dawes, R.M. & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*, 95-106.

De Beer, M. (2000). *The construction and evaluation of a dynamic computerized adaptive test for the measurement of learning potential*. Unpublished D.Litt et Phil dissertation. University of South Africa, Pretoria

De Corte, W. (1999). Weighing job performance predictors to both maximize the quality of the selected work force and control the level of adverse impact. *Journal of Applied Psychology*, *84*, 695-702.

Diamantopoulos, A. & Siguaw, J.A. (2000). *Introducing LISREL; a guide for the uninitiated*. London: Sage Publications.

Drasgow, F. & Parsons, C.K. (1983). Applications of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, *7*, 189-199.

Drasgow, F. & Hulin, C.L. (1990). Item response theory. In M.D. Dunnette & L.M. Hough (Eds.). *Handbook of industrial and organizational psychology* (Second edition). Palo Alto, CA: Consulting Psychologists Press.

Dupper, O. (2002). The burden of proof in US employment discrimination law: any lessons for South Africa. *Industrial Law Journal*, *23*, 1143-1155.

Efron, B.S. & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman and Hall.

Einhorn, H.J. & Bass, A.R. [1971]. Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin*, *75* , 261-269.

Ellis, M.V. & Blustein, D.L. (1991). Developing and using educational and psychological tests and measures: the unificationist perspective. *Journal of Counseling and Development*, *69*, 550-555.

Equal Employment Opportunity Commission. (1978). *Uniform Guidelines on Employee Selection Procedures*. 29 C.F.R. 1607.

Gatewood, R.B. & Feild, H.S. (1994). *Human resource selection* (3rd ed.). Fort Worth, TX: Dryden.

Ghiselli, E.E., Campbell, J.P. & Zedeck, S. (1981). *Measurement theory for the behavioural sciences*. San Francisco, CA: Freeman and Company.

Goldberg, L. R. (1970). Man versus model of man: a rationale plus evidence for a method of improving on clinical inferences. *Psychological Bulletin*, *73*, 422-432.

Grove, W.M. & Meehl, P.E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: the clinical- statistical controversy. *Psychology, Public Policy, and Law*, *2*, 293-323.

Guion, R.M. (1966). Employment tests and discriminatory hiring. *Industrial Relations*, *5*, 20-37.

Guion, R.M. (1991). Personnel assessment, selection and placement. In M.D. Dunnette & L.M. Hough (Eds.). *Handbook of industrial and organizational psychology* (2nd ed.; Volume 2). Palo Alto, CA: Consulting Psychologists Press.

Guion, R.M. (1998). *Assessment, measurement and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum.

Hattrup, K., Rock, J. & Scalia, C. (1997). The effects of varying conceptualisations of job performance on adverse impact, minority hiring and predictor performance. *Journal of Applied Psychology*, *82*, 656-664.

Hough, L.M., Oswald, F.L. & Ployhart (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: issues evidence and lessons learned. *International Journal of Selection and Assessment*, *9*, 152-194.

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

Hunter, J.E. & Schmidt, F.L. (1976). Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin*, *83*, 1053-1071.

Hunter, J.E. & Schmidt, F.L. (2000). Racial and gender bias in ability and achievement tests: resolving the apparent paradox. *Psychology, Public Policy, and Law*, *6*, 151-158.

Huysamen, G.K. (1995). The applicability of fair selection models in the South African context. *Journal of Industrial Psychology*, *21*, 1-6.

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

Huysamen, G.K. (2002). The relevance of the new APA standards for educational and psychological testing for employment testing in South Africa. *South African Journal of Psychology*, *32*, 26-33.

Kanjee, A. (2001). Cross-cultural test adaptation and translation. In C. Foxcroft & G. Roodt (Eds.). *An introduction to psychological assessment in the South African context*. Cape Town: Oxford University Press.

Kleinbaum, D.G. & Kupper, L.L. (1978). *Applied regression analysis and other multivariate methods*. North Sciituate, MA: Duxbury Press.

Kleinmutz, B. (1990). Why we still use our heads instead of formulas: towards an integrative approach. *Psychological Bulletin*, *107*, 296-310.

Landman, A.A. (2002). Tweaking the scales-reflections on the burden of proof in SA labour discrimination law. *Industrial Law Journal*, *23*, 1133-1142.

Landy, F.J. (1986). Stamp collecting versus science; validation as hypothesis testing. *American Psychologist*, *41*, 1183-1192.

Lawsche, C.H. & Balma, M.J. (1966). *Principles of personnel testing*. New York, NY: McGraw-Hill.

Lopes, A., Roodt, G. & Mauer, R. (2001). The predictive validity of the APIL-B in a financial institution. *Journal of Industrial Psychology*, *27*, 61-69.

Louw, N.P.van Wyk. (1965). *Die oop gesprek*. In E. Botha (Ed.). Afrikaanse essayiste. Cape Town. Human & Rousseau.

Maxwell, S.E. & Arvey, R.D. (1993). The search for predictors with high validity and low adverse impact; compatible or incompatible goals? *Journal of Applied Psychology*, *78*, 433-437.

Meehl, P. E. (1954). *Clinical vs. statistical prediction: a theoretical analysis and a review of the evidence*. Minneapolis, MI: University of Minnesota Press.

Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology*, *4*, 268-273.

Mels, G. (2003). *A workshop on structural equation modeling with LISREL 8.54 for Windows*. Chicago, Ill: Scientific Software International.

Messick, S. (1989). Validity. In R.L. Linn (Ed.). *Educational measurement* (Third edition). New York, NY: American Council on Education and McMillan Publishing Company.

Milkovich, G.T. & Boudreau, J.W. (1994). *Human resource management* (Seventh edition). Homewood, Ill: Richard D. Irwin.

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

Mossholder, K.W. & Arvey, R.D. (1984). Synthetic validity: a conceptual and comparative review. *Journal of Applied Psychology*, *69*, 322-333.

Murphy, K.R. and Davidshofer, C.O. (1988). *Psychological testing: principles and applications*. Englewood Cliffs, NJ: Prentice Hall.

Naylor, J.C. & Shine, L.C. (1965). A table for determining the increase in mean criterion score obtained by using a selection device. *Journal of Industrial Psychology*, *3*, 33-42.

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

Pulakos, E.D. & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance*, *9*, 241-258.

Republic of South Africa. (1996). Employment Equity Bill. *Government Gazette*, *390* (18481). Cape Town, 1 December.

Republic of South Africa. (1998). *Employment equity act. Government Gazette*, *400* (19370), Cape Town, 19 October.

Republic of South Africa. (2000). Promotion of equality and prevention of unfair discrimination Act. *Government Gazette*, *416* (20876), Cape Town, 9 February.

Sackett, P.R. & Ellingson, J.E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology*, *50*, 707-721.

Sackett, P.R. & Roth, L. (1996). Multi-stage selection strategies: a monte carlo investigation of effects on performance and minority hiring. *Personnel Psychology*, *49*, 549-572.

Sackett, P.R. & Wilk, S.L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist*, *49*, 929-954.

Sackett, P.R., Schmitt, N., Elligson, J.E. & Kabin, M.B. (2001). High stakes testing in employment, credentialing, and higher education; prospects in a post-affirmative-action world. *American Psychologist*, *56*, 302-318.

Saville & Holdsworth. (2000). Competency design; towards an integrated human resource management system. *SHL Newsline*, March, 7-8.

Saville & Holdsworth. (2001). Competencies and performance@work, *SHL Newsline*, May,6.

Schaap, P. & Basson, J.S. (2003). The construct equivalence of the PIB/SPEEX motivation index for job applicants from diverse cultural backgrounds. *SA Journal of Industrial Psychology*, *29*, 49-59.

Schaap, P. (2003). The construct comparability of the PIB/SPEEX stress index for job applicants from diverse cultural groups in South Africa. *South African Journal of Psychology*, *33*, 95-102.

Schaap, P. (2001)Determining differential item functioning and its effect on the test scores of selected PIB indexes, using item response theory techniques. *Journal of Industrial Psychology*, *27*, 32-38.

Schmidt, F.L. & Hunter, J.E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*, 529-540.

Schmidt, F.L. & Hunter, J.E. (1981). Employment testing; old theories and new research findings. *American Psychologist*, *36*, 1128-1137.

Schmitt, N, Rogers, W, Chan, D, Sheppard, L. & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology*, *82*, 719-730.

Schmitt, N. (1989). Fairness in employment selection. In M. Smith & I. Robertson (Eds.). A*dvances in selection and assessment*. Chichester: John Wiley.

Singer, M. (1993). *Fairness in personnel selection*. Avebury: Aldershot.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures*. Bowling Green: Author.

Society for Industrial Psychology. (1998). *Guidelines for the validation and use of assessment procedures for the workplace*. Aucklandpark:Author.

Spangenberg, H.H. & Theron, C.C. (2004). Development of a questionnaire for assessing work unit performance. *SA Journal of Industrial Psychology*, *30*, 19-28.

SPSS (2005). *SPSS 13.0 for Windows*. SPSS Inc. http://www.spss.com.

Taylor, H.C. & Russell, J.T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology*, *23*, 565-578

Taylor, T.R. (1994). A review of three approaches to cognitive assessment, and proposed integrated approach based on a unifying theoretical framework. *South African Journal of Psychology*, *24*, 184-193.

Theron, C.C., Spangenberg, H.H. & Henning, R. (2004). An elaboration of the internal structure of the unit performance construct as measured by the performance index (PI). *Management Dynamics*, *13*, 35-52

Van der Merwe, R.P. (1999). Psychological assessment in industry. *Journal of Industrial Psychology*, *25*, 8-11.

Van der Merwe, R.P. (2002). Psychometric testing and human resource management. *SA Journal of Industrial Psychology*, *28*, 77-86.

Van Zyl, E. & Visser, D. (1998). Differential item functioning in the Figure Classification test. *Journal of Industrial Psychology*, *24*, 25-33.

Vandenberg, R.J. & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-70.

Visser, D. & De Jong, A. (2000). Black and white employees' fairness perceptions of personnel selection techniques. *South African Journal of Psychology*, *30*, 17-24.

Wiggins, J.S. (1973). *Personality and prediction: principles of personality assessment*. Reading, MA: Addison-Wesley.