SA Journal of Industrial Psychology, 2004, 30 (4), 1-7
SA Tydskrif vir Bedryfsielkunde, 2004, 30 (4), 1-7

ASSESSMENT IN MULTICULTURAL GROUPS: THE SOUTH AFRICAN CASE

A.J.R. VAN DE VIJVER*
Department of Psychology, Tilburg University, the Netherlands and North-West University (Potchefstroom Campus)
S. ROTHMANN
WorkWell: Research Unit for People, Policy and Performance, North-West University (Potchefstroom Campus)

ABSTRACT

It is argued that the 1998 Employment Equity Act, in which the onus of the proof to demonstrate the adequacy of psychometric instruments is on psychology as a profession, creates daunting tasks, but also creates unique opportunities. Recent developments in the assessment of multicultural groups are described, with an emphasis on procedures to enhance the validity of measures for all groups involved and on procedures to examine validity. Bias and equivalence are treated as key concepts in multicultural assessment. Four kinds of procedures for dealing with multicultural assessment (namely, establishing equivalence of existing instruments, defining new norms, developing new instruments, and studying validity-threatening factors in multicultural assessment) are described and illustrated.

OPSOMMING

Daar word geredeneer dat die Wet op Billike Indiensneming, 1998, waarvolgens die onus om die geskiktheid van psigometriese instrumente te bewys na psigologie as professie verskuif, nie net oorweldigende take nie maar ook unieke geleenthede skep. Onlangse ontwikkelings rakende die evaluering van multikulturele groepe word beskryf, met die klem op prosedures om die geldigheid van metings vir alle groepe te verhoog en op prosedures om hierdie geldigheid te ondersoek. Sydigheid en ekwivalensie word as sleutelkonsepte in multikulturele evaluering behandel. Vier soorte prosedures om multikulturele evaluering te hanteer (te wete bepaling van die ekwivalensie van bestaande instrumente, definiëring van nuwe norme, ontwikkeling van nuwe instrumente en 'n studie van faktore wat die geldigheid van multikulturele evaluering bedreig) word beskryf en geïllustreer.

In the last decade the multicultural nature of populations has become more prominent in many countries. Processes of migration and globalisation are often associated with these changes. Take for instance the Netherlands. Labour migration and, more recently, refugee streams have added to the cultural pluralism. In the past (im)migrating groups were often small and each year there were more emigrants than immigrants. In the last 30 years the pattern has changed and the immigration stream is larger than the emigration stream. Whereas in the past immigrants always adapted fully to the Dutch culture and became fully assimilated, it seems now that the recent groups of migrants, mainly from Morocco and Turkey, show more cultural adherence and are less inclined to fully adopt the Dutch culture. This multicultural nature is a novel feature of the Dutch society; not surprisingly, it has received much attention, both in scientific and public discourse. However, there are very few countries in which this change has taken a more visible form and has had more implications than in South Africa. The current article focuses on the implications for psychological assessment of this change.

The cultural appropriateness of psychological tests and their usage were placed in the spotlight with the promulgation of the new Employment Equity Act 55 of 1998, Section 8 (Government Gazette, 1998), which stipulates that: "Psychological testing and other similar assessments 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; and (c) is not biased against any employee or group."

This law differs in two aspects with legislation in various other countries. First, in most countries the legislator adopts the opposite perspective by stipulating that discrimination and unfair treatment in psychological assessment are forbidden. The latter position assumes the fairness of psychological tests, unless proven otherwise. The South African law, on the other hand, requires psychologists to be proactively involved by requiring evidence that tests are fair and unbiased. Second, in various countries issues of bias and fairness are not primarily enacted in national laws, but in codes defined by professional organizations of psychologists and enforceable on their members. Although many countries have both legal and professional regulations, their enforcement shows considerable cross-cultural variation. For example, whereas in South Africa court cases are the main option available to plaintiffs, in a country like the Netherlands the ethics committee of the national association of psychologists is more likely to see a complaint being filed than is one of the courts.

The question can be raised as to whether psychology as a profession in South Africa is ready for the challenge implied by the Equity Act. It is probably fair to say that the law is ahead of the daily practice here and that, to date no single country can live up to the expectations and demands raised by the Act. One of the main goals of the assessment profession in South Africa is (and should be) to bring current practice in line with legal demands, for example by developing new instruments and validating existing instruments for use in multicultural groups. On the short term the Act may be seen as a threat to the profession; on the longer term the Act may enhance the professional level of psychological practice by putting multicultural assessment on the agenda of the profession and by stimulating the development of new tests and even new testing practices.

The current article describes a methodological framework for the development of new instruments and the validation of existing instruments in multicultural groups. The first part describes the context of multicultural assessment in South Africa and presents some examples of methodological studies carried out here. The second part provides a methodological framework for multicultural assessment, based on the concepts of bias and equivalence. The third part contains a description of procedures that can be employed to enhance the quality of multicultural assessment. Conclusions are drawn in the last part.

Context of multicultural assessment in South Africa

Psychological assessment tools are often used for selection and development purposes in South Africa. It is believed that these tools

Requests for copies should be addressed to: AJR Van der Vijver, Department of Psychology, Tilburg University, PO Box 90453, 5000 LE Tilburg, The Netherlands

can contribute to the efficiency of selection, placement and management of human resources (Van der Merwe, 2002). To protect the public against abuse, the use of psychological assessment tools is legally specified (Medical, Dental and Supplementary Health Service Professions Act, 1974). Various authors (e.g. Huysamen, 2002; Roodt, 1998) stressed the importance of responsible use of psychological assessment procedures.

The use of psychological tests in South Africa has largely followed international trends. At the beginning of the 1900s tests were imported from abroad and applied in all sectors of the community (Foxcroft, 1997). Psychological testing in South Africa was originally initiated with white test takers in mind (Huysamen, 2002). Psychological tests were initially developed separately for Afrikaans and English-speaking groups (Claassen, 1997), but excluded the speakers of African languages, who comprise the largest population group.

According to Abrahams and Mauer (1999, p. 76), members of historically disadvantaged groups in South Africa had suffered similar patterns of discrimination as had minority groups in the United States of America, in so far as:

- they tend to be unfamiliar with the material used in psychological tests;
- psychological tests measured different constructs from those which tests had been designed and standardised for, and
- all groups in the multicultural society are not adequately represented in the standardisation samples used to derive norm tables.

Biesheuvel's (1943, 1954) early work in South Africa focused on the empirical investigation of potential bias problems associated with cross-cultural assessment. He emphasised the importance of home environment, schooling, and nutrition and other factors on cognitive test performance in a multicultural society. Schepers (1974) reported that urban subjects, when compared with rural subjects, had a slightly greater differentiated intellect, with education playing the biggest role in the differentiation process.

Between 1960 and 1984 little research was conducted regarding the equivalence and bias of assessment instruments because of the apartheid policy in South Africa (Claassen, 1997; Owen, 1992). In the 1980s there was a growing interest in comparing cultural groups on cognitive tests. In one of the first thorough studies of comparability of test results, Owen (1986) investigated test and item bias using the Senior Aptitude Test, the Mechanical Insight Test and the Scholastic Proficiency Test. A number of cognitive ability tests have shown bias when population groups were compared (Claassen, 1993; Holburn, 1992; Owen, 1986; Owen, 1989; Taylor & Radford, 1986). Owen (1991) found that language was a potential source of bias in the Junior Aptitude Test (JAT). However, according to him, there is a greater resemblance between the cognitive structures of different cultural groups than is generally believed. Schaap (2001) carried out an item bias analysis of the Potential Index Batteries (PIB) and concluded that the instrument does not appear to discriminate unfairly against race groups.

Personality tests are widely used in South Africa. However, few studies have been conducted on the comparability of the results of different cultural groups. Spence (1982) found that the South African Personality Questionnaire (SAPQ) yielded poor alpha coefficients for Black guidance teachers. White (1982) used a number of instruments of American origin to assess job satisfaction, anxiety and job tension. Item analyses and deletion of invalid items failed to yield scales with acceptable internal consistency. Taylor and Boeyens (1991) investigated the psychometric properties of the SAPQ using two Black and two White groups of participants. They found modest support for the construct comparability between the groups, but the majority of items failed to meet the no-bias criteria which had been set.

More recently, Abrahams and Mauer (1999) studied the impact of home language on responses to the items of the Sixteen Personality Factor Questionnaire (16PF) in South Africa. They found that problems existed as far as the comparability of items across groups were concerned. Heuchert, Parker, Strumf, and Myburg (2000) applied a commonly applied measure of the Big Five, the NEO-Personality Inventory-Revised (NEO-PI-R), to college students and found a clear five-factor solution for both Black and White students. Taylor (2000) investigated the construct comparability of the NEO PI-R for Black and White employees in a work setting and found that it did not work as well for Blacks as it did for Whites.

Since the first democratic elections in 1994 the country has been regulated by a new constitution in which basic human rights and equality of individuals are guaranteed. This change has had a major impact on psychological assessment. Demands on the cultural appropriateness of psychological tests and their usage were placed in the spotlight with the promulgation of the new Employment Equity Act 55 of 1998, Section 8 (Government Gazette, 1998).

According to Owen (1991) and Maree (2000), the majority of South Africans regard the use of separate tests for different cultural groups as unacceptable. Sibaya, Hlongwane and Makunga (1996) expressed concern regarding the relevance and effectiveness of some of the assessment tools used in South Africa. The question arises whether construct-irrelevant variance such as that due to language deficiencies or cultural factors, rather than a poor standing on the construct of interest accounts for poorer performance of some groups (Huysamen, 2002).

Research regarding equivalence and bias of assessment tools in South Africa is still in its infancy stage. Clearly, much more research is needed on the equivalence and bias of assessment tools used in South Africa before psychology as a profession can live up to the demands implied in the Equity Act. The next section describes a methodological framework for multicultural assessment.

Methodological framework for multicultural assessment

This section defines the two key concepts of multicultural assessment, bias and equivalence, and discusses a taxonomy of types of bias and equivalence.

TABLE 1
DEFINITIONS OF TYPES OF BIAS AND EQUIVALENCE (FROM VAN DEN VIJVER, 2003)

Concept	Definition
Bias	Nuisance factors that threaten the comparability of scores
Construct bias	Construct measured is not identical across groups
Method bias	Nuisance factors, resulting from such factors as sample incomparability (sample bias), instrument characteristics (instrument bias), tester effects and communication problems (administration bias)
Item bias	Nuisance factors at item level
Equivalence	Comparability of test scores across cultures
Construct inequivalence	Comparisons lack a shared attribute ("comparing apples and oranges")
Structural or functional	equivalence Instrument measures the same construct in the groups studied
Measurement unit equivalence[a]	Measurement scales have the same units of measurement but a different origin (like the Celsius and Kelvin scales in temperature measurement)
Scalar or full score equivalence[a]	Scores are fully comparable across cultures (same origin and measurement unit in all cultures; this does not imply that means are identical across cultures)

[a]Assumes interval- or ratio-level measurement in all cultures studied.

Bias

Bias refers to the presence of nuisance factors in cross-cultural measurement (Poortinga, 1989). If scores are biased, their psychological meaning is not invariant across cultures and differences between cultural groups in assessment outcome are influenced by cultural or measurement artefacts. For example, differences in scores on self-esteem obtained by Chinese and American participants may be influenced by a norm of modesty, which is stronger in China than in the U.S.A., and may be contaminated by social desirability.

Table 2
Sources of bias in cross-cultural assessment (after Van der Vijver & Tanzer 1997)

Type of bias	Source of bias
Construct bias	• only partial overlap in the definitions of the construct across cultures • differential appropriateness of the behaviours associated with the construct (e.g., skills do not belong to the repertoire of one of the cultural groups) • poor sampling of all relevant behaviours (e.g., short instruments) • incomplete coverage of all relevant aspects/facets of the construct (e.g., not all relevant domains are sampled)
Method bias	• incomparability of samples (e.g., caused by differences in education, motivation)[a] • differences in environmental administration conditions, physical (e.g., recording devices) or social (e.g., class size)[b] • ambiguous instructions for respondents and/or guidelines for administrators[b] • differential expertise of administratorsb • tester/interviewer/observer effects (e.g., halo effects)[b] • communication problems between respondent and tester/interviewer (including interpreter problems and taboo topics)[b] • differential familiarity with stimulus material[c] • differential familiarity with response procedures[c] • differential response styles (e.g., social desirability, extreme scoring, acquiescence)[c]
Item bias	• poor item translation and/or ambiguous items • nuisance factors (e.g., item may invoke additional traits or abilities) • cultural specifics (e.g., incidental differences in connotative meaning and/or appropriateness of the item content)

[a]Sample bias [b]Administration bias [c]Instrument bias.

Construct bias occurs when the construct measured is not identical across groups (examples of sources of bias can be found in Table 2). Construct bias precludes the cross-cultural measurement of a construct with the same measuring instrument. An example can be found in work on filial piety (defined as the psychological characteristics associated with being perceived as a good son or daughter) (Ho, 1996). The Chinese concept, according to which children are supposed to assume the role of caretaker of elderly parents, is broader than the Western concept, which is more restricted to showing love and respect. An inventory of filial piety based on the Chinese conceptualisation will cover aspects unrelated to the concept among Western subjects, while a Western-based inventory will not touch on some important Chinese aspects.

Method bias is a label for all sources of bias emanating from the method and procedure of a study, including such factors as sample incomparability, instrument differences, tester and interviewer effects, and the mode of administration. There are three types of method bias. The first, called sample bias, refers to confounding sample differences. Research on cognitive differences between literate and illiterate individuals for example has been plagued by sample bias, because a comparison of literates and illiterates is almost always a comparison between schooled and unschooled persons. So, a study of the influence of literacy then almost inevitably becomes a study of the influence of schooling. Sample bias can be expected to increase with the cultural distance between the samples.

Administration bias, a second source of method bias, can be caused by differences in the procedures or mode used to administer an instrument. For example, when interviews are held in participants' homes, physical conditions (e.g., ambient noise and presence of others) are difficult to control. Participants and clients are more prepared to answer sensitive questions in (anonymous) self-completion questionnaires than in the "shared" discourse of an interview (Kalgraff Skjåk & Harkness, 2003). Another source of administration bias can be ambiguity in the questionnaire instructions and/or guidelines, or a differential application of these instructions (e.g., answers to open questions are considered to be ambiguous and require follow-up questions). The effect of test administrator or interviewer presence on measurement outcomes has been empirically studied (Kane & Macaulay, 1993). Larger effects have been found in interview studies of attitudes than in standardised testing of cognitive abilities (Jensen, 1980). Deference to the interviewer has been reported; subjects were, for instance more likely to display positive attitudes to a particular cultural group when they are interviewed by someone from that group (Weeks & Moore, 1981). A final source of administration bias is constituted by communication problems between tester and participant or client. Language problems may be a potent source of bias when the participants differ in proficiency in the testing language, which is not uncommon in multicultural studies, in which a test or interview is administered in the second or third language of the participants. Knowledge of the dominant language of a country can be a critical issue in multicultural assessment, even when linguistic skills are not assessed. The level of proficiency to answer items of personality instruments is often quite advanced.

The third source of method bias is *instrument bias*, which involves general features of an instrument that give rise to unintended cross-cultural differences. Individuals from different cultures do not deal in the same way with Likert rating scales, such as five- or seven-point scales for assessing agreement. For example, compared to European Americans, Hispanics have been found to show more extreme scores on a five-point scale, but this tendency disappeared when a ten-point scale was used (Hui & Triandis, 1989).

In addition to the construct and the method, items themselves can be the source of bias. Item bias, also known as *differential item functioning*, is a generic name for all disturbances at item level (Berk, 1982; Camilli & Shepard, 1994; Van de Vijver & Leung, 1997). According to a definition that is widely used in psychometrics, an item is biased if respondents with the same standing on the underlying construct (e.g., being equally intelligent) but who come from different cultures do not have the same mean score on the item (e.g., Holland & Wainer, 1993). For instance, if a geography test administered to pupils in the Netherlands and South Africa contains the item "What is the capital of South Africa?", South African pupils can be expected to show higher scores on this item than Dutch students, even when pupils with the same total test score are compared. The item is biased because it favours one cultural group across all test score levels. Several psychometric techniques are available to identify item bias (see, for example, Camilli & Shepard, 1994). The most common sources of item bias are poor item translation, ambiguities in the original item, low familiarity or appropriateness of the item content in certain cultures, and the influence of cultural specifics such as nuisance factors or connotations associated with the item

wording. For example, in a translation of the word "aggression" (as used by Americans) it is difficult or even impossible to maintain the combined meaning of violence ("an aggressive predator") and enterprising energy ("an aggressive salesperson") of the original.

Equivalence

Equivalence refers to the comparability of test scores obtained in different cultural groups; it involves the question as to whether scores obtained in different cultures can be meaningfully compared (Poortinga, 1989; Przeworski & Teune, 1966, 1970). Equivalence and bias are related. If scores are unbiased (free from nuisance factors), they are equivalent and (assuming that they are metrical) can be compared across cultures.

Four different types of equivalence may be distinguished (cf. Van de Vijver & Leung 1997). *Construct inequivalence* amounts to "comparing apples and oranges". Many examples of inequivalent constructs can be found in the clinical literature on culture-bound syndromes. For example, "Latah" is a syndrome that can be found only among some Asian groups, consisting of a sudden fright, resulting in uncontrollable imitative behaviours (e.g., verbal repetition of obscenities) (cf. Berry, Poortinga, Segall & Dasen, 2002).

An instrument administered in different cultural groups shows *structural equivalence* if it measures the same construct in all these groups. Statistical techniques, notably factor analysis, are employed to examine structural equivalence. If an instrument yields the same factors in different cultural groups, there is strong evidence that the instrument measures the same underlying construct(s). Structural equivalence does not presuppose the use of identical instruments across cultures (Przeworski & Teune, 1970). For example, a measure of depression may be based on partly or entirely different indicators in each cultural group and still show structural equivalence. Structural equivalence has been addressed for various cognitive tests and personality measures (see next section).

The third type of equivalence is called *measurement unit equivalence*. Instruments show this type of equivalence if their scales have the same units of measurement but a different origin (such as the Celsius and Kelvin scales in temperature measurement). This type of equivalence assumes interval- or ratio-level scores (with the same measurement units in each culture). It applies when a source of bias with a fairly uniform influence on the items of an instrument affects test scores of different cultural groups in a differential way. Social desirability and stimulus familiarity may exert this influence. If these factors have a differential influence on the scores obtained in the various cultural groups, observed score differences are hard to interpret due to the confounding of valid score differences and measurement artefacts. Empirical research shows that the role of response sets (in personality and attitude measurement) and stimulus familiarity (in cognitive testing) cannot be neglected. For example, there are country differences in social desirability; more specifically, social desirability is inversely related to national wealth and educational level (Van Hemert, Van de Vijver, Poortinga, & Georgas 2002; Johnson & Van de Vijver, 2003); so, individuals who come from richer countries or who are better educated tend to show lower scores on social desirability.

Only in the case of *scalar (or full score) equivalence* can direct comparisons be made using statistical tests such as the t-test and analysis of variance. This is the only type of equivalence that allows for the conclusion that average scores obtained in two cultures are different or equal. Scalar equivalence assumes that identical interval or ratio scales are applicable across cultural groups.

Structural, measurement unit, and scalar equivalence are hierarchically ordered. The third presupposes the second, which presupposes the first. As a consequence, higher levels of equivalence are more difficult to establish. It is easier to verify that an instrument measures the same construct in different cultural groups (structural equivalence) than to identify numerical comparability across cultures (scalar equivalence).

Validity enhancement in multicultural assessment

Five approaches can be envisaged to enhance the development or testing of assessment instruments for multicultural groups. The first is simple: It is recommended to *document in the test manual* how the test has been made suitable for usage in a multicultural context and to describe in the manual which aspects of the test administration are particularly important when the test is applied in a multicultural context. The "Standards for Educational and Psychological Testing" issued jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999) provide a good example of recommended practices that can be used as a frame of reference elsewhere. Although not all guidelines listed therein will be equally relevant (after all, they were meant only for the U.S.A.), and the guidelines are incomplete in that they involve test characteristics and there is no reference to ethnic groups (e.g., which factors should be taken into account when assessing group X?), they provide a useful starting point.

The second way to enhance the quality of multicultural instruments amounts to the "exportation" (Berry et al., 2002) of existing, almost always Western tests, to the non-Western world and *the study of bias and equivalence* of data obtained in the different countries. Literally hundreds of such studies have been carried out. There is an implicit prototype for these studies: a test has been administered in different countries, factor structures are compared, using either exploratory or confirmatory factor analysis, and if sufficient agreement between the factors obtained in the different countries is found, it is concluded that there is structural equivalence (i.e., the test measures the same construct in all countries involved). As an example, Irvine (1969, 1979) concluded that the structure of tests of inductive reasoning found among western participants with exploratory factor-analytic techniques is usually replicated in non-western samples. More recent comparative studies, based on comparisons of ethnic groups in the U.S.A., have confirmed this conclusion (e.g., Fan, Willson & Reynolds, 1995; Jensen, 1980). Van de Vijver (2002) administered tests of inductive reasoning to pupils from primary and secondary schools in Zambia, Turkey, and the Netherlands. He found strong evidence for the similarity of the factor structures underlying the test performance across the three countries. Major differences in structure (for instance as reported by Claassen & Cudeck, 1985) are unusual.

The third way of dealing with multicultural assessment is the *development of culture-specific* norms or score adjustments of members of minority groups. As an example, Resing, Bleichrodt and Drenth (1986) administered a children's test of intelligence, widely employed in the Netherlands, to the major groups of migrants in order to examine whether the test showed particular sources of bias and to determine norms for the various groups.

Various score adjustment procedures have been proposed, such as within-group norming (i.e., establishing norms per cultural group) and bonus points. As an example of the latter, Mercer (1979) (see also Cronbach, 1984, p. 211ff.) designed a system for "correcting" test scores of a child (such as scores on the WISC) based on information of the socioeconomic background of the child. Scores of White children were typically shifted downwards, while scores of Hispanic children and Black children were boosted by the "correction". Another example is the sliding band (Cascio, Outtz, Zedeck & Goldstein, 1991). The procedure defines score bands, beginning with the top score. The band consists of all scores that do not differ significantly from the top score. Within the particular band all observed scores are assumed to reflect

equal proficiency for the prospective job. Minority applicants within the band are then selected first, followed by the choice of majority group members with the highest score. The band is then "slided" to a lower score (one less than the top score) and the procedure is repeated until the target number of chosen applicants is reached. If, and only if members of minorities have scores within bands of eligible candidates, they get preferential treatment.

The score adjustment tradition has been challenged on psychometric grounds. Thus, Cronbach (1984) criticized Mercer's score corrections because no data were provided indicating that the "corrected" scores showed a higher validity. For example, no data were provided to demonstrate that the "corrected" scores better predicted school performance or more adequately reflected the intellectual abilities of candidates. Cronbach's argument is important because it points to the need to base conclusions on solid psychometric grounds. However, Cronbach's argument is also incomplete. The perceived desirability of correcting (or not correcting) for these differences cannot be justified on psychometric grounds; more importantly, no psychometric rationale is needed. Whether or not a society will legalize and accept affirmative action is the outcome of a complex interaction between relevant stakeholders such as political parties, representatives of the justice system, employers, employees, and minority interest groups. The only contribution of psychometrics to this debate can and should be the specification of models to identify bias and group-dependent scoring schemes. Implicit rules of conduct or explicit legislation, inspired by the public debate, define the typically narrow operational boundaries of psychometric procedures.

The fourth approach to deal with cultural heterogeneity is *the development of new instruments*. Examples of Dutch cognitive tests are the Multiculturele Capaciteitentest voor Middelbaar Niveau (MCT-M) of Van den Berg and Bleichrodt (2001; Bleichrodt & Van den Berg, 1997), the Leertest voor Etnische Minderheden (LEM) (Hessels, 1993, 1996) and the Tilburgse Allochtonen en Autochtonen Reactietijd Test (TAART) (Helms-Lorenz et al., 2003). In each of these tests an attempt is made to maximize the suitability of the test for migrants by, among other things, reducing the cultural loading of item contents, and providing an extensive test instruction.

A fifth approach is *the study of factors that threaten the validity of multicultural assessment procedures with a view toward improving their quality*. Two examples are discussed here: the role of cultural loadings in cognitive tests and of response sets in personality instruments. A seemingly perennial problem of cognitive and educational testing is the influence of cultural loadings on test performance. For a very long time cognitive test performance has been known to be susceptible to schooling effects. This effect may be a consequence of the implicit cultural loading in cognitive tests. One of the questions in a well-known Dutch intelligence test for children asks from which animal bacon is made. Migrant children with an Islamic cultural background (which has a taboo on pork) found the question relatively difficult compared to their mainstream age mates. The difficulty of the item reflects differences in cultural background of the children. Problems of differential cultural loadings that occur in isolated items may well be identified by item bias techniques. The question can be raised, however, whether differential cultural loadings will not affect a substantial proportion of the items, in particular when the cultural distance between the groups assessed is very large (which is typically the case in South Africa).

Spearman's hypothesis, formulated by Jensen (1985), predicts larger score differences between African Americans and European Americans on more pure measures of "g" (Spearman's general intelligence factor). A test of higher cognitive complexity tends to have a higher g-saturation, usually expressed by its loading on the first factor of the intercorrelation matrix of a battery of tests. So, Spearman's hypothesis holds that cross-cultural performance differences on cognitive tests of African and European Americans are larger on complex tasks than on simple tasks. Support for Spearman's hypothesis has been derived from a number of studies (Jensen, 1998). It has also been used in interpreting ethnic differences in cognitive test scores in the Netherlands (Te Nijenhuis & Van der Flier, 1997). Helms-Lorenz, Van de Vijver and Poortinga (2003) questioned the equivalence of *g* loadings as an index of (inborn) intellectual capacity. They argued that these loadings are likely to tap additional factors such as cultural practices, cognitive algorithms learned at school, and especially knowledge of the (Dutch) language of the test designer. Twelve tests were administered to 6-12 year old Dutch school children, including a sample children of autochthonous descent and a sample of second-generation Turkish migrants. Structural equivalence could be demonstrated. The migrants showed on average a lower average than the mainstream sample. The nature of the difference was further explored. The complexity level (Carroll, 1993; Fischer, 1980), verbal loading and cultural loading of each test were rated by advanced psychology students. Two factors were extracted from these ratings. The first, labelled "aggregate *g*", showed high loadings for complexity and Jensen's measure of g; the second, called "aggregate *c*", (*c* for culture) had high loadings for verbal loading and cultural loading. Culturally more entrenched tests (i.e., tests with a higher loading on the c factor) showed a larger score difference in the two ethnic samples, while the *g* factor smaller differences. This indicates that, contrary to Spearman's hypothesis, performance differences did not increase with *g* – saturation; rather intergroup performance differences were better predicted by c than by *g*. It was concluded that familiarity with the Dutch language and culture was an important source of intergroup differences. It is clear that the development of new instruments or the modification of existing ones can benefit from insights in the nature of cultural loadings on mental test performance.

Personality questionnaires have their own source of nuisance factors that can threaten the validity of assessment in multicultural groups: response sets, such as acquiescence, and social desirability. In particular social desirability has received attention in cross-cultural studies. Van Hemert et al. (2002) carried out a meta-analysis on the Eysenck Personality Questionnaire which, among other things, contains a social desirability scale. The main finding was the strong negative correlation of social desirability with Gross National Product and other related economic variables. This strong social desirability effect can be interpreted in two ways, in line with two existing models. First, it can be seen as method bias. According to this explanation differences in social desirability should be considered as a kind of response bias, in which participants deliberately portray an incorrect picture of themselves (such as the well-known underreporting of alcohol intake or substance use by addicts). Second, the results can be interpreted as reflecting differences in social-psychological functioning. This can be explained as a reflection of ecocultural conditions, such as affluence. A sociocultural explanation of differences in social-psychological functioning can be considered as well. Ross and Mirowsky (1984) argue that less powerful groups are more prone to socially desirable responding, as less powerful groups are often less affluent groups. People from these groups are forced to behave according to social norms because they depend more on the approval of other people and cannot "afford" to be independent. In this line of interpretation social desirability can be interpreted as 'social naïveté' or conformity (see Eysenck & Eysenck, 1975; Furnham, 1986; Ones, Viswesvaran & Reiss, 1996). The cross-cultural record suggests that social desirability varies across cultural groups according to a meaningful pattern (with more affluent, powerful groups showing less social desirability) and that cross-cultural differences in social desirability can be substantial.

Therefore, it is fair to conclude that treatment of social desirability as "a red herring" (Ones et al., 1996) might be productive or at least innocuous in monocultural research, but is counterproductive in multicultural assessment.

CONCLUSION

Multicultural assessment is a new branch of the tree of psychological assessment. In recent years the branch even seems to get new shoots such as cross-cultural neuropsychological assessment (e.g., Ferraro, 2002; Nell, 2002). The relative novelty may incorrectly convey the impression that the topic is of relevance for an esoteric group of experts. The opposite is true. The advent of multicultural assessment is mainly inspired by a growing societal need; it is a response to the perceived need to deal with a multitude of cultures in assessment without the a priori designation of a single culture as the target or model for other cultures. The development in which growing numbers of members of various ethnic minorities ask for culture-specific and culture-informed psychological practices will gain momentum in the coming years. Multicultural assessment has come to us not long ago, but it is fair to assume that it will stay with us for quite some time. It is in our own professional interest that we take care of it in an adequate way.

REFERENCES

Abrahams, F. & Mauer, K. F. (1999). Qualitative and statistical impacts of home language on responses to the items of the Sixteen Personality Questionnaire (16PF) in South Africa. *South African Journal of Psychology*, *29*, 76-86.

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

Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting item bias*. Baltimore: John Hopkins University Press.

Biesheuvel, S. (1943). *African intelligence*. Johannesburg: South African Institute of Race Relations.

Biesheuvel, S. (1954). The measurement of occupational aptitudes in a multi-racial society. *Occupational Psychology*, *52*, 289-196.

Bleichrodt, N. & Van den Berg, R. H. (1997). *Voorlopige handleiding MCT-M. Multiculturele Capaciteiten Test voor Middelbaar Niveau*. Amsterdam: Stichting NOA.

Berry, J. W. Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (2002). *Cross-cultural psychology: Research and applications* (2nd ed.). Cambridge: Cambridge University Press.

Camilli, G. & Shepard, L. N. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Carroll, J. B. (1993). *Human cognitive abilities. A survey of factor-analytic studies*. Cambridge: Cambridge University Press.

Cascio, W. F., Outtz, J., Zedeck, S. & Goldstein, I. L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance*, *4*, 233-264.

Claassen, N. C. W. (1997). Culture differences, politics and test bias in South Africa. *European Review of Applied Psychology*, *47*, 297-307.

Claassen, N. C. W. & Cudeck, R. (1985). Die faktorstruktuur van die Nuwe Suid-Afrikaanse Groeptoets (NSAG) by verskillende bevolkingsgroepe. *South-African Journal of Psychology*, *15*, 1-10.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York: Harper & Row.

Eysenck, H. J. & Eysenck, S. B. G. (1975). Manual of the Eysenck Personality Questionnaire. London: Hodder and Stoughton.

Fan, X., Willson, V. L. & Reynolds, C. R. (1995). Assessing the similarity of the factor structure of the K-ABC for African-American and White children. *Journal of Psychoeducational Assessment*, *13*, 120-131.

Ferraro, R. F. (Ed.) (2002). *Minority and cross-cultural aspects of neuropsychological assessment*. Lisse, the Netherlands: Swets & Zeitlinger.

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review*, *87*, 477-531.

Foxcroft, C. D. (1997) Psychological testing in South Africa: Perspectives regarding ethical and fair practices. European *Journal of Psychological Assessment*, *13*, 229-235.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences*, *7*, 385-400.

Government Gazette, 1998, *Republic of South Africa*, Vol. 400, no. 19370, Cape Town, 19 October 1998.

Helms-Lorenz, M., Van de Vijver, F. J. R. & Poortinga, Y. H. (2003). Cross-cultural differences in cognitive performance and Spearman's Hypothesis: G or C? *Intelligence*, *31*, 9-29.

Hessels, M. G. P. (1993). *Leertest voor Etnische Minderheden. Theoretische en empirische verantwoording*. Ph.D. Thesis. Rotterdam: RISBO.

Hessels, M. G. P. (1996). Ethnic differences in learning potential test scores: Research into item and test bias in the Learning Potential Test for Ethnic Minorities. *Journal of Cognitive Education*, *5*, 133-153.

Heuchert, J. W. P., Parker, W. D. Strumf, H. & Myburg, C. P. H. (2000). The five-factor model for African college students. *American Behavioral Scientist*, *44*, 112-125.

Ho, D. Y. F. (1996). Filial piety and its psychological consequences. In M. H. Bond (Ed.), *Handbook of Chinese psychology* (pp. 155-165). Hong Kong: Oxford University Press.

Holburn, P. (1992). Test bias in some HSRC tests. *Proceedings of the Congress on Psychometrics for psychologists and personnel practitioners*, Pretoria.

Holland, P. W. & Wainer, H. (Eds.) (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hui, C. H. & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20*, 296-309.

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.

Irvine, S. H. (1969). Factor analysis of African abilities and attainments: Constructs across cultures. *Psychological Bulletin*, *71*, 20-32.

Irvine, S. H. (1979). The place of factor analysis in cross-cultural methodology and its contribution to cognitive theory. In L. Eckensberger, W. Lonner, & Y. H. Poortinga (Eds.), *Cross-cultural contributions to psychology* (pp. 300-343). Lisse, the Netherlands: Swets & Zeitlinger.

Jensen, A. R. (1980). Bias in mental testing. New York: Free Press.

Jensen, A. R. (1985). The nature of Black-White difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences*, *8*, 193-263.

Jensen, A. R. (1998). *The g factor. The science of mental ability*. Westport, CT: Praeger.

Johnson, T. P. & Van de Vijver, F. J. R. (2003). Social desirability in cross-cultural research. In J. A. Harkness, F. J. R. Van de Vijver & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 195-204). New York: Wiley.

Kalgraff Skjåk, K. & & Harkness, J. (2003). Data collection methods. In In J. A. Harkness, F. J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 179-193). New York: Wiley.

Kane, E. W. & Macaulay, L. (1993), Interviewer gender and gender attitudes. *Public Opinion Quarterly*, *57*, 1-28.

Maree, J. G. (2000). 'n Oorsig van die faktore rakende die kompleksiteit van psigometriese toetsing in multikulturele Suid-Afrika. *Tydskrif vir Geesteswetenskappe*, *400*, 318-329.

Mercer, J. R. (1979). *System of Multicultural Pluralistic Assessment. Technical manual*. New York: Psychological Corporation.

Nell, V. (2000). *Cross-cultural neuropsychological assessment: Theory and practice*. Mahwah, NJ: Erlbaum.

Ones, D. S., Viswesvaran, C. & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81*, 660-679.

Owen, K. (1986). *Test and item bias: Administration of the Senior Attitude test, Mechanical Insight test and the Scholastic Proficiency Battery to White, Asian, Black and Colored Technikon students*. Pretoria: Human Sciences Research Council.

Owen, K. (1989). *Test and item bias: the suitability of the Junior Aptitude Test as a common test battery of White, Indian and Black pupils in Standard seven*. Pretoria: Human Sciences Research Council.

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

Owen, K. (1992). *Test-item bias: Methods, findings and recommendations*. Pretoria: Human Sciences Research Council Group: Education

Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology*, *24*, 737-756.

Prezworski, A. & Tuene, H. (1966). Equivalence in cross-national research. *Public Opinion Quarterly*, *30*, 551-568.

Prezworski A. & Teune H. (1970). *The logic of comparative social inquiry*. New York: Wiley.

Prinsloo, C. H. & Ebersöhn, I. (2002). Fair usage of the 16PF in personality assessment in South Africa: A response to Abrahams and Mauer with special reference to research methodology. *South African Journal of Psychology*, *32*, 48-57.

Resing, W. C. M., Bleichrodt, N. & Drenth, P. J. D. (1986). Het gebruik van de RAKIT bij allochtoon etnische groepen. *Nederlands Tijdschrift voor de Psychologie*, *41*, 179-188.

Roodt, G. (1998). Challenges in psychological assessment. *People Dynamics*, *16* (11), 30-34.

Ross, C. E. & Mirowsky, J. (1984). Socially-desirable response and acquiescence in a cross-cultural survey of mental health. *Journal of Health and Social Behavior*, *25*, 189-197.

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.

Schepers, J. M. (1974) Critical issues which have to be resolved in the construction of test for developing groups. *Humanitas RSA*, *2*, 395-406.

Sibaya, P. T., Hlongwane, M. & Makunga, N, (1996). Giftedness and intelligence assessment in a third world country. *Giftedness Education International*, *11* (2), 107-113.

Spence, B. A. (1982). *A psychological investigation into the characteristics of black guidance teachers*. Unpublished master's dissertation, University of South Africa, Pretoria.

Taylor, T. R. & Boeyens, J. C. A. (1991). The comparability of the scores of blacks and whites on the South African Personality Questionnaire: An exploratory study. *South African Journal of Psychology*, *21*, 1-10.

Taylor, J. M. & Radford, E. J. (1986). Psychometric testing as an unfair labour practice. *South African Journal of Psychology*, *16*, 79-96.

Te Nijenhuis, J. & Van der Flier, H. (1997). Comparability of GATB scores for immigrants and majority group members: Some Dutch findings. *Journal of Applied Psychology*, *82*, 675-687.

Van de Vijver, F. J. R. (2002). Inductive reasoning in Zambia, Turkey, and The Netherlands: Establishing cross-cultural equivalence. *Intelligence*, *30*, 313-351.

Van de Vijver, F. J. R. (2003). Culture and mental measurement. In C. D. Spielberger (Editor in Chief), *Encyclopaedia of applied psychology*. San Diego, CA: Academic Press. (in press).

Van de Vijver, F. J. R. & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Newbury Park, CA: Sage.

Van den Berg, R. & Bleichrodt, N. (2001). Het meten van cognitieve vaardigheden bij allochtone volwassenen. In N. Bleichrodt & F. J. R. Van de Vijver (Eds.), *Het gebruik van psychologische tests bij allochtonen* (pp. 119-141). Lisse, the Netherlands: Swets & Zeitlinger.

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

Van Hemert, D. A., Van de Vijver, F. J. R., Poortinga, Y. H. & Georgas, J. (2002). Structural and functional equivalence of the Eysenck Personality Questionnaire within and between countries. *Personality and Individual Differences*, *33*, 1229-1249.

Weeks, M. F., & Moore, R. P. (1981). Ethnicity-of-interviewer effects on ethnic respondents. *Public Opinion Quarterly*, *45*, 245-249.

White, D. H. (1982). *The effects of job stress in the South African mining industry*. Unpublished doctoral thesis, University of South Africa, Pretoria.