THE CONSTRUCT EQUIVALENCE AND ITEM BIAS OF THE PIB/SPEEX CONCEPTUALISATION-ABILITY TEST FOR MEMBERS OF FIVE LANGUAGE GROUPS IN SOUTH AFRICA

PIETER SCHAAP
THERESA VERMEULEN
Department of Human Resource Management
University of Pretoria
South Africa

Correspondence to: Pieter Schaap
e-mail: pieter.schaap@up.ac.za

ABSTRACT

This study's objective was to determine whether the Potential Index Batteries/Situation Specific Evaluation Expert (PIB/SpEEx) conceptualisation (100) ability test displays construct equivalence and item bias for members of five selected language groups in South Africa. The sample consisted of a non-probability convenience sample (N = 6 261) of members of five language groups (speakers of Afrikaans, English, North Sotho, Setswana and isiZulu) working in the medical and beverage industries or studying at higher-educational institutions. Exploratory factor analysis with target rotations confirmed the PIB/SpEEx 100's construct equivalence for the respondents from these five language groups. No evidence of either uniform or non-uniform item bias of practical significance was found for the sample.

Keywords: Abstract reasoning, psychological instruments, cross-cultural fairness, item bias, test bias

In South Africa, as elsewhere in the world, psychological instruments are often used for selection and development purposes (Van de Vijver & Rothmann, 2004). Psychological tests are commonly used as aids to determine whether employees have the necessary skills for a specific job (Van der Merwe, 1999).

Psychological instruments can be divided into different groups or types, such as cognitive, personality and interest tests. This study focuses on a cognitive test known as the Potential Index Batteries/Situation Specific Evaluation Expert (PIB/SpEEx) conceptualisation (100) ability test. The PIB/SpEEx 100 test was developed by Potential Index Associates specifically to assess job-relevant conceptual-reasoning skills within a cross-cultural context (Erasmus, 2001). The history and development of cognitive tests in general and the use thereof in a cross-cultural context are particularly relevant to the issue at hand.

Cognitive tests have been developed over more than a century and a variety of perspectives about what constitutes intelligence has emerged. Initially, the specific constructs that were measured by cognitive tests were disputed and theories were therefore developed to explain what really constitutes intelligence or cognitive ability as well as how best to measure these concepts and how to measure these constructs across different cultures in a fair and unbiased way (Gregory, 2004).

The possibilities of unfairness and bias in the use of cognitive tests have resulted in extensive research on factors that might affect the fairness of psychometric instruments. Culture and language can be included among these factors (Gregory, 2004).

Initially, there was little or no attempt to assess cognitive competence in a culturally relevant framework (Kendell, Verster & Von Mollendorf, 1988) and early pioneers in the assessment movement largely ignored the impact of cultural background on test results (Gregory, 2004). Early psychometric testing in South Africa mainly followed international trends and, at the beginning of the 1900s, when psychology began to emerge as an independent field of study, tests were imported from Europe and North America and applied in all sectors of the community without any distinction being made (Foxcroft, 1997).

Gradually, however, an increasing need for change in psychometric testing throughout the world in general and in South Africa in particular began to emerge. Cross-cultural issues began to emerge as being problematic in the 1920s (Meiring, Van de Vijver, Rothmann & Barrick, 2005), when studies of diverse-culture assessments became somewhat more systematic and empirically orientated. It began to dawn on practitioners that not all instruments were equally appropriate to all peoples and cultures (Bedell, Van Eeden & Van Staden, 2000).

In the 1940s and 1950s, work in the psychometric domain in South Africa focused rather pragmatically on the educability and trainability of black South Africans. There was some realisation that cultural differences can influence testing outcomes and attempts to create 'culture-free' tests soon became the vogue (Bedell *et al*., 2000). Biesheuvel (1949; 1952) can be considered as one of the pioneers in the development of tests to solve problems associated with the testing of preliterate black populations. The General Adaptability Battery was one of the better measures developed during this period for the testing of blacks with little educational background for occupational suitability.

From 1960 onwards, there was a growing recognition that culture exerts subtle and pervasive effects in the testing domain and that it is not possible to remove culture from the equation. At that time, it was increasingly understood that culture affects behaviour and consequently the psychological constructs that were measured and culture began to be seen as an important moderator of test performance (Kendell *et al*., 1988). From 1960 to 1984, The National Institute for Personnel Research and the Institute for Psychological and Edumetric Research played important roles in the development of measures along cultural and racial lines in the South African context. Both these institutions were later incorporated into the Human Sciences Research Council, which took over the

role of test development. The emphasis at the time was on the development of separate measures for each cultural group and/or the use of group-specific norms (Foxcroft & Roodt, 2005). Examples of measures that were developed during the 1990s are the General Scholastic Aptitude Test and the Paper and Pencil Games (PPG) test (Claassen, 1990; 1996). The PPG is, to date, the only measure available in all 11 official languages in South Africa. Due to problems experienced in the comparison of the scores of tests developed for different groups, the focus since the late 1990s has been on the development of tests that are fair in terms of both language and culture (Foxcroft & Roodt, 2005). Examples of more recent developments along these lines are the Learning Potential Computerised Adaptive Tests and the Ability, Processing of Information and Learning Battery tests (Taylor, 1997 & De Beer, 2000).

A shift towards a closer consideration of any cultural bias inherent in tests also strengthened the notion that culture may constitute a source of systematic error in test results. Kendell *et al.* (1988) point out that test scores often correlate with non-test variables, such as test-taking behaviour, cultural and/or environmental factors and dispositional factors. Test-taking behaviour is influenced by factors such as the level of education, home language, practice or familiarity with tests of the person(s) taking the tests.

Among these factors, the issue of language received much attention in psychological assessment, as it is an overriding consideration that linguistic barriers may inhibit the test performance of minority groups (Gregory, 2004). Language is closely linked to the culture in which a test is developed, as language is almost always used to express the cultural concepts and constructs that need to be measured (McCrae, 2000). However, this is a complex area of study, as not only are there inter-cultural differences in language usage but language itself also evolves and changes over time, even within cultural groupings (Wallis & Birt, 2003).

To compensate for the problems associated with the link between culture and language, some test developers sought to resolve such problems by developing tests in different languages (Bedell *et al.*, 2000). However, the translation of tests into different languages (which was expected to be the answer to the culture dilemma) posed problems of its own. Various practical problems were found with the translation procedure itself. Although there is support for the adaptation of existing tests and the development of culturally appropriate tests and norms, it must be recognised that there are several difficulties in developing and norming tests in a culturally and linguistically diverse society (Foxcroft, 1997).

In the South African context, very specific problems arose in the translation of tests due to three main reasons. Firstly, South Africa has 11 official languages and tests therefore have to be translated into all 11 languages (presenting problems with regard to cost, to the lack of available translators with both language and specialist psychological/human-resource expertise and to a lack of equivalent specialist vocabulary in all the languages). Secondly, among the limited pool of available test administrators, there are not enough administrators speaking the preferred language of test takers who cannot understand English. Thirdly, practitioners have reported problems with regard to the different dialects (of one language) spoken in different areas and to a difference in performance between urban and rural individuals tested in their mother tongue (Bedell *et al.*, 2000).

Apart from the link between culture and language, language is also linked to cognitive processes. Galotti (2004) points out that the use of language in a variety of cognitive tasks raises the following important question: what influence(s) does language have on other cognitive processes? Two extreme positions exist: on the one hand, Chomsky (in Sharrat, 1987) argues that language and other cognitive processes operate completely independently of each other and, on the other, Sharrat (1987) posits that language and other cognitive processes relate completely with one determining the other. Between these two extremes there is considerable middle ground, where language and other cognitive processes are seen as related in some ways but as independent in others (Galotti, 2004). For example: Chomsky (in Sharrat, 1987) states that children manage to acquire language rapidly and efficiently at a stage when cognitive functioning still seems to be relatively undeveloped – thereby implying that language acquisition and cognitive processes are independent processes. Sharrat (1987) argues in favour of the dependency of language and cognitive processes in that the structure of language causes people to think of the world in certain ways. Both arguments appear to be plausible depending on the context within which the arguments are presented.

Owen (1992) studied the content and format of items in tests that function differentially and suggests reasons for bias: language (especially in the case of the black subjects in his study who were tested in English) and cognitive style (subject-related). Language training and problem-solving strategies were recommended, as the differences in mean test performance preclude the use of common norms, while the use of separate norms for the different population groups defeats the purpose of a common test (Owen, 1992).

Therefore, the language in which a test is developed has important consequences because of its relation to both culture and cognitive processes. Culture and language have an effect on cognitive processes and may consequently affect an individual's performance in cognitive tests.

Non-verbal tests: A means to reduce the effect of culture and language proficiency on test performance

In reaction to criticism arguing that intelligence tests are culturally biased, a number of non-verbal tests of intelligence have been published (Owen, 1998). Non-verbal tests were developed to measure fluid intelligence, which is a relatively culture-reduced form of mental efficiency (Gregory, 2004). Fluid intelligence is related to a person's inherent capacity to learn and solve problems and is thus used when a task requires a person to adapt to a new situation (Gregory, 2004).

Historically, test developers have tried to construct non-verbal tests of intelligence to meet the needs of a linguistic minority (in other words, individuals who have limited proficiency in the language of the dominant culture). Typically, in Europe and North America, these individuals are either foreign-born or have hearing problems. The situation in South Africa is somewhat different, in that the 'linguistic minority' may, in fact, be a numerical majority of people who do not belong to a 'foreign' group at all. Increasingly, there is a greater realisation among psychologists that many measuring devices are not entirely appropriate for subjects whose mother tongue or first language is not English, for illiterates and for those with speech and hearing impairments (Gregory, 2004).

According to Kline (1993), non-verbal items include pictorial odd-man-out, pictures with errors that have to be recognised, figure classification in which two figures of a series that belong together have to be selected, embedded figures where a shape embedded in other shapes has to be discovered, the identification of the sequence of shapes in matrix format and other variations of pictorial stimuli. Examples of specific non-verbal tests include the Test of Non-verbal Intelligence (TONI), Cattell's Culture Fair Intelligence Test (CFIT) and Raven's Progressive Matrices (RPM).

The TONI items require the examinee to solve problems by identifying relationships among abstract figures. Many of the items are similar in format to those found in the RPM (Gregory, 2004).

The CFIT is a non-verbal measure of fluid intelligence or the ability to engage in analytic and reasoning activities with abstract and novel materials. This is a widely used test, particularly for examinees with language or cultural deficits. Originally designed by Cattell (1940), this test is a culture-free measure of cognitive aptitude. It consists of items without any verbal content. However, questions have been raised about the extent to which the test is completely free of cultural content (for example, even pictures can be culturally loaded) and the name was later changed from the Culture Free Intelligence Test to the CFIT (Hoge, 1999).

The RPM has a non-verbal construction and does not rely on an examinee's fluency in English or any other language, since it consists only of universal symbols. It is often used when testers require a measure of aptitude and ability that is not biased by a test candidate's educational background, ethnic or racial differences, linguistic ability or cultural deficiencies (Samuda *et al*., 1998). However, reviewers of this test have raised some questions about the construct validity of the instrument, as it is not entirely clear what aspects of cognitive aptitude are assessed. Hoge (1999) states, moreover, that it is clear that RPM scores are not equivalent to the abstract-reasoning scores yielded by an instrument such as the Wechsler Intelligence Scale.

Like the above-mentioned tests, the PIB/SpEEx 100 test is also a non-verbal cognitive-assessment measure.

The development of non-verbal tests is seen as a possible solution to minimising the effect of language proficiency on the comparability of the test scores of different groups.

Challenges associated with cross-cultural testing in South Africa

At present, psychological assessment in South Africa faces many challenges (Foxcroft, Paterson, Le Roux & Herbst, 2004), including the following:

- The creation of tests that can be used without bias across diverse linguistic and cultural backgrounds is a complex process (Huysamen, 1996).
- According to recent legislation, notably the stipulations of the Employment Equity Act, Act 55 of 1998 (Section 8), qualified professionals may use only psychological tests and similar instruments that can be proved to be scientifically valid and reliable and that are not biased against any particular employee or group (Republic of South Africa, 2006).

These challenges encourage industrial psychologists to conduct applied research on the psychometric properties of tests and to explore the fairness of tests that are used (Foxcroft *et al*., 2004).

Although it is reassuring to see the vast interest in cross-cultural studies, it is regrettable that practitioners and academics do not have a well-established and widely adopted practice in cross-cultural research to deal with issues such as instrument feasibility and multiple interpretations (Van de Vijver, 1998). According to Van de Vijver (1998), bias and equivalence are concepts that form the core of a framework attempting to incorporate aspects specific to cross-cultural research.

Previous studies in South Africa report race, the level of education, language and the understanding of English to be the main factors that affect the construct and item comparability of cognitive tests (Meiring *et al*., 2005). Therefore, there is a need to continue to research the issue of bias and equivalence in the culturally diverse South Africa (Meiring *et al*., 2005). Bias and equivalence research would assist in establishing whether assessment instruments are fair to all language or cultural groups.

The objective of this study was to determine the construct equivalence and item bias of the non-verbal PIB/SpEEx 100 test for diverse language groups in South Africa. The key terms 'construct equivalence' and 'item bias' are briefly explained below.

In theory, the concepts of equivalence and bias are the opposite of each other. Thus, scores are equivalent when they are not biased. Nevertheless, in cross-cultural research conducted to date, the two concepts are treated separately and become associated with different aspects of cross-cultural comparisons. Equivalence is associated with the measurement level at which scores obtained in different cultures can be compared and bias is a generic term for all measurement artefacts that threaten the validity of cross-cultural comparisons (Van de Vijver & Leung, 1997).

Construct equivalence (also known as structural equivalence) is at the first-measurement level and indicates the extent to which the same construct is measured across different cultural groups under study. Construct equivalence is a precondition to subsequent measurement levels known as measurement-unit equivalence (ratio level) and scalar equivalence (interval level). Measurement-unit equivalence requires the offset of scales to be similar for groups and scalar equivalence requires scores on the instrument to have the same interval scales across cultural groups (Van de Vijver, 1998). The problem with dichotomous items is that they do not have an origin or a unit of measurement and the concepts of unit and scalar equivalence consequently cannot be applied to dichotomous variables (Eid, Langeheine, & Diener, 2003).

According to Van de Vijver and Leung (1997), item bias refers to measurement artefacts at item level. A few examples from an inexhaustive list of nuisance factors at item level are poor item translation, inappropriate item content and inadequate item formulation (complex wording). Item bias is a measurement problem that, if not attended to, can jeopardise the validity of cross-cultural comparisons.

When a test is biased towards a group, the scores for the group consistently underestimate or overestimate the true values. A test can be said to be biased towards a group when any given score obtained by an individual in that group does not have the same meaning as the very same score obtained by an individual in another group. The two groups in question might be from different racial groups, have different socio-economic backgrounds, or be of different genders or any other biographical category of persons in the general population (Jensen, 1981).

To understand the need for this study, one must take into account the history and development of cognitive tests as well as the many challenges that psychological assessment faces in South Africa. 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 Employment Equity Act (Van de Vijver & Rothmann, 2004). For these reasons, this study aims specifically to investigate the equivalence and bias of the PIB/SpEEx 100 test to ensure that this test is used appropriately in the South African context and measures one construct, namely conceptualisation or conceptual reasoning, in different language groups.

RESEARCH DESIGN

Research approach

In this study, a quasi-experimental design was used. Quasi-experimental designs help researchers test for causal relationships in a variety of situations (Neuman, 1997). According to Van de Vijver and Leung (1997), cross-cultural studies fall into the quasi-experimental category. Within the context of cross-cultural research, quasi-experimental methodological considerations centre on the enhancement of

the interpretability of observed differences in the focal variable and on a reduction in the number of alternative explanations. A substantive step in the process of enhancing the interpretability of observed differences and reducing alternative explanations is the choice of appropriate context variables either to verify or to falsify a particular interpretation (Van de Vijver & Leung (1997). It is evident from the preceding literature discussion that language group as a context variable can be considered a plausible explanation of the observed differences in the focal variable (test score). Consequently, the research method followed in this study is designed to evaluate the (lack of) success of the context variable 'language group' as an alternative explanation for observed score differences in the no-verbal-based PIB/SpEEx 100 test.

The method followed in this study is discussed below with regard to the respondents, the measuring instrument and the statistical procedures used.

Research method

Respondents

TABLE 1
Biographical information on the respondents

	FREQUENCY	VALID %	CUMULATIVE %
Gender			
Female	2 690	43.2	43.2
Male	3 544	56.8	100.0
Total	6 234	100.0	
Unknown	27		
Total	6 261		
Age in years			
16–20	2 890	69.8	69.8
21–25	1 116	27.0	96.8
26–30	87	2.1	98.9
31–35	25	0.6	99.5
36–40	8	0.2	99.7
41–45	7	0.2	99.9
46–50	5	0.1	100.0
Total	4 138	100.0	
Home language			
Afrikaans	1 643	26.2	26.2
English	912	14.6	40.8
North Sotho	1 304	20.8	61.6
Setswana	1 139	18.2	79.8
isiZulu	1 263	20.2	100.0
Total	6 261	100.0	
Education			
Grades 1–7	71	1.2	1.2
Grades 8–12	5 508	89.6	90.8
Occupational certificate	16	0.3	91.1
Tertiary diploma/degree	549	8.9	100.0
Total	6 144	100.0	
Unknown	117		
Total	6 261		
Industry			
Beverage	1 353	21.7	21.7
Medical	269	4.3	26.0
Tertiary institution	4 605	74.0	100.0
Total	6 227		
Unknown	34		
Total	6 261		

A non-probability convenience sample was drawn from three industries and sectors within South Africa, namely the beverage-manufacturing industry, the medical sector and two tertiary institutions from the higher-education sector. The ample included 6 261 participants from five different language groups (Afrikaans, English, North Sotho, Setswana and isiZulu). The participants numbered as follows: 1 643 Afrikaans speakers, 912 English speakers, 1 304 North Sotho speakers, 1 139 Setswana speakers and 1 263 isiZulu speakers. The biographical information of the sample is presented in Table 1.

The sample consisted of 43.2% females and 56.8% males. A further 4.3% of the respondents did not indicate their gender and are therefore indicated as unknown in Table 1. Most of the respondents (89.6%) had completed secondary school up to Grade 8 or Grade 12, while the rest of the sample had obtained a diploma or degree at university or technikon as their highest qualification (8.9%). Only 1.87% of the respondents did not indicate their qualification(s). Most of respondents were enrolled as full-time students at tertiary institutions (74%), while the rest were employed by the beverage (21.7%) and medical (4.3%) industries.

The mean age of the sample was 20.26 years. The youngest respondent was 17 years old and the oldest was 49 years old.

Measuring instrument

The aim of the PIB/SpEEx is to provide a comprehensive assessment package suitable for the assessment and development of human potential in the workplace. The various indices assess human potential relating to specific dimensions or basic competencies. These are identified in the PIB/SpEEx battery manual (Erasmus, 2001) as set out below.

The PIB/SpEEx battery consists of two types of scales, namely the cognitive and the behavioural scales (Erasmus, 2001). PIB/SpEEx (conceptualisation) 100 is one of the cognitive scales, which means that it assesses an element of intellectual potential and, more specifically, conceptual reasoning. PIB/SpEEx 100 is a visual or non-verbal scale and, because it consists of visual or non-verbal items that explore reasoning processes using shapes and figures, it could arguably be administered in any language whatsoever (Erasmus, 2001). It is therefore particularly useful when people with poor English language skills or any other language, for that matter, are assessed (Erasmus & Schaap, 2003).

The PIB/SpEEx 100 test is a normative scale consisting of 30 items. The respondent must complete a pattern through the identification of one or more rules that determine the relationships of parts. The test assesses potential to reason in spatial terms, to see the relationships of parts, to complete a picture, to envisage a whole or an end result and to anticipate outcome. It is a performance test and a time limit of 15 minutes to complete the test is therefore imposed.

In a previous study, the average metrical properties of the PIB/SpEEx 100 were investigated. The sample included different industry sectors and academic institutions. It was reported that the PIB/SpEEx 100 scale obtained a mean Cronbach alpha coefficient of 0.90 (Schaap, 2001).

Research procedure

Data were collected from the existing PIB/SpEEx database, which is used for selection and development purposes in industry and tertiary institutions. The pencil-and-paper version of the PIB/SpEEx 100 test was applied. All the data were acquired under the supervision of registered psychologists and were dealt with in a manner that protects the confidentiality of test results.

Statistical analysis

Construct-equivalence and item-bias analyses were used in this study to evaluate the PIB/SpEEx 100 test's comparability across language groups. The gathered data were analysed by means of scale-level analyses to examine the similarity of the factors underlying the PIB/SpEEx 100 test as well as bias at item level. The SPSS (SPSS Inc, 2006) and MicroFACT 2.0 (Waller, 1995) computer programs were used to perform the required statistical analyses.

Descriptive statistics and reliability analysis. Descriptive statistics were calculated in respect of the test scores of the total sample and in respect of the language groups to provide an understanding of the distribution of scores within and between the groups. A reliability analysis was done for each group. Reliability coefficients can provide valuable clues about the measurement accuracy and the appropriateness of an instrument for cross-cultural comparison (Van de Vijver & Leung, 1997).

Factor analysis. Construct equivalence can be investigated through several techniques, such as factor analysis, cluster analysis and multidimensional scaling or other dimensionality-reducing techniques (Van de Vijver & Leung, 1997). In this study, construct equivalence was examined by means of exploratory factor analysis.

Rogers (1995) explains that the reason for using exploratory factor analysis is to identify a latent subset of psychological characteristics or factors that underlie a specific domain. The basic idea behind the application of this technique is to obtain the structure of each group, which can then be compared across all the groups involved (language groupings, in the case of this specific study). Factor analysis is the most frequently employed technique in the study of construct equivalence (Naudé & Rothmann, 2004). Exploratory factor analysis derives factors that provide the best statistical fit to the data (Murphy & Davidshofer, 2001). According to Rogers (1995), the aim of exploratory factor analysis is to express observed scores as scores on a limited set of unobserved, underlying factors.

Factor analysis is relevant for the establishment of construct equivalence because it decomposes observed scores into unobserved components (Van de Vijver & Leung, 1997). In this study, the factor analysis consisted of a two-step procedure proposed by Van de Vijver and Leung (1997). Firstly, a Principle Axis Factor (PAF) analysis was conducted on the total sample group, which yielded a common matrix of factor loadings. This served as the target matrix for comparison purposes. Secondly, the factor loadings of each language group were compared with the target matrix by means of targeted rotation. The factor loadings were rotated to one target group (total group) to determine the construct equivalence of the factor for the other language groups. Factorial agreement was then estimated with Tucker's coefficient of congruence (Tucker's phi) (Van de Vijver & Leung, 1997).

Item-level analysis. In theory, a test and test items measuring a specific construct are perfectly unidimensional, that is all items of a specific test measure one and the same construct. In practice, however, this absolute is never attained (Rudner, Getson & Knight, 1980). The one-dimensionality assumption is theoretically a prerequisite for item-bias analysis. Item bias refers to the extent to which an item measures a construct differently across different populations.

In this study, item-bias analysis was performed through logistic regression. More specifically, the study made use of binomial (or binary) logistic regression, which is applicable when one or more variables consist of dichotomous scores, in this case correct and incorrect responses to the various items of the test.

In the case of item bias, a distinction can be made between uniform and non-uniform bias. According to Van de Vijver and Leung (1997), uniform bias refers to the influence of bias on scores that are more or less the same for all score levels. Individuals from one cultural group may have higher scores on an item than individuals from another cultural group, even when they have the same total score.

Non-uniform bias refers to a situation where influences are not identical for all score levels and an item discriminates better in one group than in another. An item is taken to show non-uniform bias if the interaction between the score level and culture is significant (Meiring et al., 2005).

Logistic regression is a technique to fit a regression model to data where the dependent variable is dichotomous (Howell, 1997). It is unique in its ability to predict dichotomous variables and, like correlation, provides information about the strength and direction of the relationships across the variables (Marczyk, DeMatteo & Festinger, 2005). In this study, the total test score and language served as the independent variables, while the item score was the dependent variable.

Logistic regression can be used to predict a dependent variable (in this study, the item score) on the basis of independent variables (the test score and language) and to determine the percentage of variance of the dependent variable explained by the independent variables, to determine the relative importance of independent variables, to assess interaction effects and to understand the impact of covariate control variables (Kerlinger & Lee, 2000). In this study, the Chi-square statistic was used to evaluate the statistical significance of the uniform and the non-uniform item bias.

In this study, the Nagelkerke R^2 statistic was used to calculate the effect size (the strength of the relationship) between the dependent variable and the independent variables. The effect size of language (the uniform bias) was determined through the calculation of the difference between the Nagelkerke R^2 of the first step (in which score level was the sole predictor) and that of the second step (in which language, dummy coded, was added as a predictor). In the third step, the interaction of culture and score level was added. The difference between the second and the third step estimates the effect size of the interaction (the non-uniform bias) (Meiring *et al*., 2005).

RESULTS

Descriptive statistics

Table 2 shows the descriptive statistics and Cronbach alpha coefficients of the measuring instrument for the different groups.

Observable differences exist in the mean values, standard deviations (SDs), coefficients of skewness and kurtosis as well as in the alpha (α) coefficients of the five different language groups that were compared. For example, a noticeable difference in the mean, SD, skewness, kurtosis and reliability of test scores exists in respect of the Afrikaans and Northern Sotho groups. The observed score differences between the groups naturally raise questions concerning the construct equivalence and bias

TABLE 2
Descriptive statistics and Cronbach Alpha coefficients

GROUP	MEAN	SD	SKEWNESS	KURTOSIS	ALPHA (α)
Total group	18.89	5.94	-0.36	-0.56	0.88
English	21.25	5.37	-0.22	-0.42	0.87
Afrikaans	21.77	4.49	-0.20	-0.85	0.83
Northern Sotho	16.48	6.00	-0.71	0.52	0.87
Setswana	17.40	5.48	-0.53	0.07	0.85
isiZulu	17.27	6.20	-0.23	-0.68	0.88

of the instrument and add to the importance of conducting appropriate analyses.

Acceptable Cronbach alpha coefficients varying from 0.83 to 0.88 were obtained for the different groups. These alpha coefficients are acceptable if one uses the guideline of $\alpha > 0.80$ suggested by Anastasi and Urbina (1997).

In order to compare a test across cultures meaningfully, its equivalence must be demonstrated in those cultures, in this case different language groups. In this study, this was done at the item level and at the test level as suggested by Kline (1993).

Factor analysis

Exploratory factor analysis was used to determine the construct equivalence of 'conceptualisation' as measured by the PIB/ SpEEx 100 test.

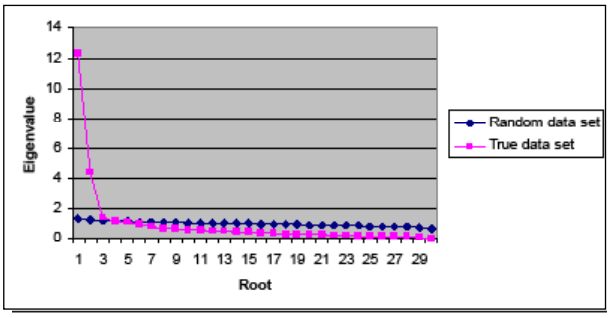

FIGURE 1
Scree plot for the item-level factor analysis in respect of the total group

An item-level PAF analysis based on a tetrachoric correlation matrix was performed in respect of each group. Tetrachoric correlation is used in factor analysis when both variables are dichotomous and are assumed to represent underlying bivariate normal distributions, as is the case when a dichotomous test item is used to measure some dimension of achievement (Van de Vijver & Leung, 1997).

According to the scree plot, the factor analysis yields more factors than expected (see Figure 1). According to Hambleton and Swaminathan (1985), factor analysis based on tetrachoric correlations is inclined to yield too many factors. The difference between the eigen-values of the first two roots and the rest suggests that there might be two significant constructs. A clear break can be observed on the scree plot between roots two and three for all the groups. The eigen-values of the random data set (the broken line) intersect the eigen-values for the true data set (the solid line) at root three, indicating two significant factors (Horn, 1965).

More detailed results of the item-level factor analysis for the total group are depicted in Table 3. The results show that the first factor accounts for up to 60% of the variance of the unrotated factor matrix. This is significantly more than the criterion of Shillaw (1996) of at least 20% variance on the first factor before unidimensionality can be assumed. In addition, the eigen-value of the first factor also needs to be significantly higher than that of the next largest factor. The first factor has a variance of more than three times that of the second factor, which provides strong evidence in favour of assuming unidimensionality.

Due to the fact that the variance for the second factor was relatively high compared to the third factor and the eigen-value significant, the possibility of a second meaningful construct

TABLE 3
Eigen-values and percentage of variance explained (per group) for the unrotated factor matrix

	TOTAL GROUP		ENGLISH		AFRIKAANS		NORTH SOTHO		SETSWANA		ISIZULU	
	Eigen-value	% variance	Eigen-value	% variance	Eigen-value	% variance	Eigen-value	% variance	Eigen-value	% variance	Eigen-value	% variance
1	12.312	60.15	11.694	54.683	10.632	49.511	11.691	56.434	10.887	49.755	11.625	38.751
2	4.420	21.59	4.337	20.281	5.388	25.091	4.341	20.954	4.286	19.588	4.355	14.517
3	1.432	6.99	1.746	8.165	1.624	7.563	1.839	8.877	1.740	7.952	1.672	5.574
4	1.206	5.90	1.451	6.785	1.501	6.989	1.507	7.275	1.511	6.906	1.157	3.857
5	1.100	5.40	1.146	5.359	1.257	5.853	1.338	6.459	1.277	5.836	1.084	3.613
6	0.955	4.70	1.011	4.727	1.072	4.992	0.939	4.533	1.161	5.306	0.957	3.190
7	0.809	4.00	0.977	4.569	0.989	4.606	0.852	4.113	1.019	4.657	0.871	2.903
8	0.698	3.40	0.850	3.974	0.878	4.088	0.820	3.958	0.912	4.168	0.768	2.560
9	0.685	3.30	0.823	3.832	0.778	3.622	0.692	3.340	0.831	3.798	0.714	2.380
10	0.624	3.00	0.807	3.773	0.685	3.189	0.647	3.123	0.768	3.510	0.652	2.173
11	0.583	2.80	0.654	3.058	0.630	2.933	0.621	2.998	0.686	3.135	0.631	2.103
12	0.527	2.60	0.636	2.974	0.568	2.645	0.560	2.703	0.598	2.733	0.604	2.013
13	0.509	2.50	0.556	2.599	0.550	2.561	0.481	2.322	0.521	2.381	0.563	1.877
14	0.437	2.10	0.493	2.305	0.502	2.337	0.446	2.153	0.482	2.203	0.531	1.018
15	0.416	2.00	0.483	2.259	0.441	2.054	0.430	2.076	0.459	2.098	0.467	1.557
16	0.383	1.87	0.382	1.786	0.405	1.886	0.395	1.907	0.419	1.915	0.436	1.453
17	0.363	1.77	0.348	1.627	0.368	1.714	0.348	1.680	0.393	1.796	0.409	1.363
18	0.331	1.62	0.318	1.487	0.348	1.621	0.329	1.588	0.391	1.787	0.361	1.203
19	0.319	1.56	0.258	1.206	0.328	1.527	0.297	1.434	0.329	1.504	0.334	1.670
20	0.281	1.37	0.226	1.057	0.262	1.220	0.251	1.212	0.275	1.257	0.324	1.080
21	0.273	1.33	0.216	1.010	0.196	0.912	0.229	1.105	0.236	1.079	0.291	0.970
22	0.223	1.09	0.193	0.903	0.158	0.736	0.199	0.961	0.201	0.919	0.262	0.873
23	0.219	1.07	0.156	0.729	0.137	0.638	0.180	0.869	0.170	0.777	0.218	0.727
24	0.183	0.9	0.123	0.575	0.119	0.554	0.155	0.748	0.146	0.667	0.205	0.683
25	0.173	0.8	0.067	0.313	0.098	0.456	0.131	0.632	0.130	0.594	0.152	0.507
26	0.155	0.76	0.035	0.164	0.081	0.377	0.116	0.560	0.088	0.402	0.145	0.483
27	0.153	0.75	0.015	0.070	0.005	0.023	0.078	0.377	0.048	0.219	0.084	0.280
28	0.136	0.66	0.000	0.000	0.000	0.000	0.060	0.290	0.035	0.160	0.068	0.227
29	0.073	0.36	0.000	0.000	0.000	0.000	0.030	0.145	0.000	0.000	0.052	0.173
30	0.025	0.12	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.007	0.023

could not be entirely ruled out. Upon further inspection, it was found that a plausible explanation for the second factor was an artefact attributed to the differential item skewness or the difficulty factor of the items. All the items with salient loadings on factor two had a high difficulty value (p-value) in common. A more precise statistical method was therefore needed to determine dimensionality and deal with the effect of item differential skewness.

Consequently, the procedure of Schepers (1992) was applied to determine the dimensionality of the PiB/SpEEx 100 statistically. Schepers developed the procedure to control for factor artefacts that form as a result of items being differentially skewed. The first phase of Schepers's procedure requires an iterative factor analysis and a varimax rotation of the significant factors extracted through the use of Kaiser's criterion (Kaiser, 1961). The PAF analysis (based on tetrachoric correlations) on the total group yielded five factors, which were then subjected to varimax rotation. The items with the highest loading on the respective factors were aggregated to form a new set of variables. The new set of variables was subjected to PAF analysis based on Pearson's correlation matrix. A single factor (explaining the 45% variance) emerged through the use of Kaiser's criterion (see Figure 2). Accordingly, it was confirmed that the PiB/SpEEx 100 test is unidimensional.

Structural (construct) equivalence

The results of the item-level PAF analysis (based on tetrachoric correlations) were used in the structural equivalence analysis. The PAF analysis was repeated for each group and one factor was extracted for comparison purposes. The factor loadings as well as Tucker's congruence coefficients are presented in Table 4.

Factor loadings of 0.30 and higher can generally be considered acceptable (Tabachnik & Fiddell, 1989). Small deviations from the 0.30 criterion were allowed to account for differences in sample homogeneity (Schaap & Basson, 2003). Of the thirty items, only five items showed low factor loadings across the different language groups. Three of these five items (items 3, 4 and 6) consistently displayed low factor loadings across all language groups, while items 18 and 25 displayed a low factor loading for only one language group (Setswana). In total, 25 items had moderate to high factor loadings (> 0.30 with permitted deviations) for all the language groups.

Tucker's phi coefficients for the different language groups are given in Table 4 above. As a general rule, values higher than 0.95 are seen as evidence of factorial similarity, whereas values lower than 0.90 are taken as pointing to non-similarity (Van de Vijver & Leung, 1997). Inspection of Table 4 above shows that Tucker's phi coefficients with values higher than 0.95 are present in all the different language groups. This provides a strong indication of structural equivalence and it can therefore be deduced that the construct is equivalent for all five different language groups.

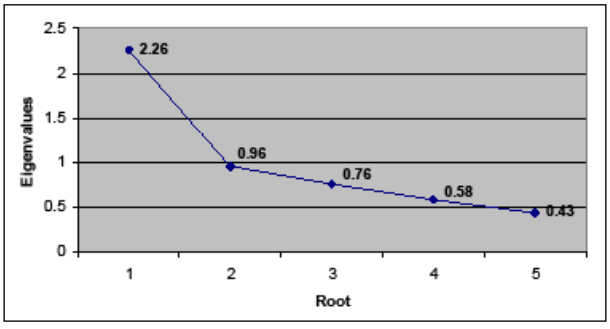

FIGURE 2
Scree plot for the factor-analytical procedure of Schepers (1992) in respect of the total group

Constructs that are equivalent for different cultural groups indicate an absence of construct bias in an instrument (Schaap & Basson, 2003).

Analysis of item bias

The aim of this analysis was not to test for cultural differences but to test whether the item scores were identical for respondents from different language groups with an equal total score level.

Table 5 presents a cross-tabulation of the different language groups and score categories. The cross-tabulation provides information about the cell sizes of the matrix that was used for item-bias analysis.

The respondents were divided into seven groups according to their ability level (test score levels). The various language groups in the seven different ability levels in the table all have more than 50 cases, which can be considered acceptable cell sizes for the purpose of item-bias analysis.

TABLE 4
Factor loadings after the target rotation of the first factor for the different language groups

ITEM	TOTAL GROUP	ENGLISH	AFRIKAANS	NORTH SOTHO	SETSWANA	ISIZULU
1	0.49	0.45	0.46	0.60	0.34	0.29
2	0.52	0.55	0.51	0.53	0.43	0.29
3	0.32	0.28	0.31	0.35	0.24	0.26
4	0.30	0.24	0.25	0.42	0.23	0.20
5	0.55	0.52	0.46	0.62	0.40	0.41
6	0.30	0.23	0.22	0.36	0.30	0.16
7	0.44	0.51	0.36	0.31	0.66	0.55
8	0.41	0.32	0.44	0.47	0.14	0.29
9	0.58	0.56	0.62	0.58	0.49	0.44
10	0.56	0.49	0.60	0.64	0.35	0.39
11	0.46	0.29	0.39	0.52	0.44	0.36
12	0.61	0.46	0.50	0.63	0.52	0.50
13	0.69	0.59	0.67	0.65	0.64	0.61
14	0.58	0.49	0.42	0.65	0.52	0.57
15	0.66	0.61	0.50	0.63	0.60	0.61
16	0.71	0.68	0.60	0.66	0.72	0.65
17	0.80	0.76	0.63	0.75	0.84	0.75
18	0.38	0.30	0.40	0.36	0.23	0.38
19	0.80	0.76	0.78	0.72	0.82	0.82
20	0.83	0.84	0.79	0.76	0.81	0.83
21	0.83	0.83	0.73	0.75	0.84	0.84
22	0.85	0.85	0.77	0.74	0.89	0.85
23	0.76	0.74	0.76	0.66	0.76	0.79
24	0.82	0.83	0.75	0.72	0.90	0.82
25	0.56	0.54	0.59	0.57	0.27	0.48
26	0.61	0.65	0.59	0.55	0.60	0.68
27	0.72	0.74	0.69	0.64	0.78	0.73
28	0.75	0.75	0.67	0.66	0.80	0.87
29	0.59	0.55	0.62	0.53	0.67	0.71
30	0.69	0.73	0.67	0.64	0.70	0.81
Tucker's congruence index		**0.99**	**0.99**	**0.99**	**0.98**	**0.99**

TABLE 5
Total-score cross-tabulation for the different language groups

	TOTAL-SCORE LEVELS (ABILITY LEVELS)							
	1 (Low)	2	3	4	5	6	7 (High)	TOTAL
Language								
Afrikaans	109	170	205	288	280	212	379	1 643
English	130	84	94	121	129	134	220	912
North Sotho	579	178	150	116	104	78	99	1 304
Setswana	408	196	143	113	96	87	96	1 139
isiZulu	493	152	137	125	116	122	118	1 263
Total	1 719	780	729	763	725	633	912	6 261

Table 6 indicates that, when bias is evaluated in terms of the statistical significance of Chi-square, most items reveal statistically significant bias. The criterion of Cohen (1988), according to which the lower threshold for medium-effect size is 0.06, was applied to examine the practical significance of the item bias (this size was chosen because it can be considered significantly large enough to be practically important). Many items show statistical bias but the bias effect size (Nagelkerke R^2) is so slight as to be negligible from a practical point of view.

DISCUSSION

As is mentioned in the introduction to this article, many challenges are faced in the use of different assessment instruments in the South African context today. Obtaining equivalent measures that may be used across a diversity of linguistic and cultural backgrounds is perhaps the most central issue in cross-cultural and cross-language comparative research (Huysamen, 1996).

The purpose of this article is to report on the construct equivalence and item-bias research that was conducted on the PIB/SpEEx (conceptualisation) 100 test for five language groups in South Africa.

Overall, small observable differences in scale reliabilities in respect of the various language groups provide some indication that the construct may be equivalent for the language groups. The scale reliabilities are all well within the range of what is generally considered acceptable for different groups. As it is recognised that differences in scale reliabilities among groups could be considered as preliminary and not as conclusive evidence, factor analysis was required to provide more conclusive evidence (Schaap & Basson, 2003).

Factor analysis of the PIB/SpEEx 100 test yielded a single dominant factor, as expected. The percentage of variance explained by the factors and eigen-values suggests that there is only one significant construct that can be identified as the conceptualisation-ability construct.

Results of the exploratory factor analysis indicate similar response patterns for the different groups on most of the items. The factor congruence coefficients obtained meet the criterion of high agreement and emphasise the structural equivalence of the construct for the different language groups (Meiring *et al.*, 2005).

Although most items show statistically significant bias due to the large sample size, which increases the sensitivity of the Chi-square statistic, the bias is so small as to be negligible from a practical perspective (Cohen, 1988). With regard to the evaluation of item bias, it was found that none of the items show either uniform or non-uniform bias of practical significance.

Overall, it can be concluded that the PIB/SpEEx 100 test appears to be equivalent and that the test items are not biased for the different language groups included in this study. Thus, the non-verbal items of the PIB/SpEEx 100 scale do not appear to be language-sensitive for the language groups included in this study. The assumption made by Erasmus and Schaap (2003) that the non-verbal scales of the PIB/SpEEx are language-free and can be administered in any language can therefore be confirmed for the PIB/SpEEx 100 test for five South African language groups.

TABLE 6
Item-bias statistics of the conceptual-reasoning test for the different language groups

	UNIFORM BIAS		NON-UNIFORM BIAS	
Item	Chi-square	Effect size (Nagelkerke R^2)	Chi-square	Effect size (Nagelkerke R^2)
1	51.625 (4); p = 0.000*	0.021	1.974 (4); p = 0.741	0.001
2	46.753 (4); p = 0.000*	0.021	10.892 (4); p = 0.028*	0.007
3	8.907 (4); p = 0.063	0.002	1.309 (4); p = 0.860	0.000
4	28.712 (4); p = 0.000*	0.007	3.683 (4); p = 0.451	0.001
5	60.890 (4); p = 0.000*	0.014	5.499 (4); p = 0.240	0.001
6	58.073 (4); p = 0.000*	0.012	5.879 (4); p = 0.208	0.001
7	133.154 (4); p = 0.000*	0.024	65.349 (4); p = 0.000*	0.011
8	222.156 (4); p = 0.000*	0.043	26.564 (4); p = 0.000*	0.005
9	18.303 (4); p = 0.001*	0.008	4.204 (4); p = 0.379	0.002
10	198.292 (4); p = 0.000*	0.042	23.836 (4); p = 0.000*	0.005
11	54.446 (4); p = 0.000*	0.014	18.017 (4); p = 0.001*	0.005
12	158.224 (4); p = 0.000*	0.035	9.179 (4); p = 0.057*	0.002
13	69.594 (4); p = 0.000*	0.017	6.218 (4); p = 0.183	0.001
14	13.937 (4); p = 0.008*	0.003	7.703 (4); p = 0.103	0.003
15	79.898 (4); p = 0.000*	0.019	6.594 (4); p = 0.159	0.001
16	39.195 (4); p = 0.000*	0.007	6.249 (4); p = 0.181	0.001
17	27.467 (4); p = 0.000*	0.005	32.318 (4); p = 0.000*	0.006
18	90.591 (4); p = 0.000*	0.017	27.387 (4); p = 0.000*	0.005
19	10.575 (4); p = 0.032*	0.002	18.432 (4); p = 0.001*	0.003
20	21.118 (4); p = 0.000*	0.003	5.902 (4); p = 0.207	0.001
21	31.280 (4); p = 0.000*	0.004	4.374 (4); p = 0.358	0.000
22	58.362 (4); p = 0.000*	0.007	12.926 (4); p = 0.012*	0.002
23	54.047 (4); p = 0.000*	0.006	39.770 (4); p = 0.000*	0.005
24	94.478 (4); p = 0.000*	0.010	16.714 (4); p = 0.002*	0.002
25	37.626 (4); p = 0.000*	0.010	89.703 (4); p = 0.000*	0.023
26	47.240 (4); p = 0.000*	0.010	58.975 (4); p = 0.000*	0.012
27	75.739 (4); p = 0.000*	0.012	46.760 (4); p = 0.000*	0.007
28	77.684 (4); p = 0.000*	0.011	64.610 (4); p = 0.000*	0.009
29	197.626 (4); p = 0.000*	0.032	93.777 (4); p = 0.000*	0.014
30	138.068 (4); p = 0.000*	0.020	90.623 (4); p = 0.000*	0.012

*p < 0.05: item shows significant (non-)uniform bias if followed by an asterisk (*)

The introduction to this study states that a question that needs answering is whether a given psychometric instrument can stand the scrutiny of the Employment Equity Act and its subsections (Republic of South Africa, 2006), in this case a test used to measure conceptual-reasoning ability. This test did not show practical significant bias and the results are consequently encouraging for the equitable use of the PIB/SpEEx 100 test in a multicultural environment like that of South Africa.

Recommendations and suggestions for further research

As discussed, according to Van de Vijver and Leung (1997), there are three kinds of bias: item, construct and method bias. This study does not address all the aspects of test usage but focuses on item and construct bias. Method bias (this refers to problems deriving from instrument characteristics) is not taken into consideration and should be investigated in a separate study.

Multi-sample confirmatory factor analysis (MCFA) procedures suitable for dichotomous variables should be considered, as these provide more options to test for measurement invariance (Skrondal & Rabe-Hesketh, 2005). Compared to PAF procedures, MCFA is a more versatile tool when testing for the hierarchically linked hypotheses of cross-cultural measurement invariance. MCFA allows for the testing of specific pattern coefficients, error variances and factor covariances to determine the specific differences among groups and to understand the aspects of the test structure that differ across groups (Maller & French, 2004).

To ensure the equitable use of the PIB/SpEEx 100 test, the predictive validity and predictive bias of the test can also be considered. Even an unbiased instrument may not work equally well for different language groups. This study does not address the question of whether the cognitive scale can predict future training and job performance in a fair way for all language groups. A final verdict on the cross-cultural suitability of the current test can be given only once the predictive bias is also tested.

In the light of the importance of different language groups in this study, it is recommended that future research include biographical questions that elicit responses on current home language and mother (original) tongue. These questions are highly applicable to South Africa, since many people indicate English as their home language when their mother tongue is not, in fact, English.

Although further investigation is needed, the prospects for the use of the PIB/SpEEx 100 in a multicultural environment seem favourable. The development of new instruments or the modification of existing ones can benefit from insights gained from research on the nature and extent of cultural loadings on cognitive-ability tests.

REFERENCES

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

Bedell, B., Van Eeden, R. & Van Staden, F. (2000). Culture as moderator variable in psychological test performance: Issues and trends in South Africa. *South African Journal of Psychology*, 25(3), 1–7.

Biesheuvel, S. (1949). Psychological tests and their application to non-European peoples. *The Year Book of Education*, 87–126.

Biesheuvel, S. (1952). Personnel selection tests for Africans. *South African Journal of Science*, 49, 3–12.

Cattell, R.B. (1940). A culture-free intelligence test. *Journal of Educational Psychology*, 31, 161–180.

Claassen, N.C.W. (1990). The comparability of general scholastic aptitude test scores across different population groups. *South African Journal of Psychology*, 20, 80–92.

Claassen, N.C.W. (1996). *Paper and pencil games (PPG). Manual*. Pretoria: Human Sciences Research Council.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

De Beer, M. (2000). *Learning potential computerized adaptive test (LPCAT): User's manual*. Pretoria: Production Printers.

Eid, M., Langeheine, R. & Diener, E. (2003). Comparing typological structures across cultures by multigroup latent class analysis: A primer. *Journal of Cross-Cultural Psychology*, 34(2), 195–210.

Erasmus, P.F. (2001). *Situation-specific job profiling and assessment in the workplace for the 21st century*. Johannesburg: Potential Index Associates.

Erasmus, P. & Schaap, P. (2003). *Situation-specific job profiling and assessment short course JP Expert/PIB SpEEx*. Update course presented at the University of Pretoria, Pretoria.

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

Foxcroft, C.D., Paterson, H., Le Roux, N. & Herbst, D. (2004). *Psychological assessment in South Africa: A needs analysis. The test use patterns and needs of psychological assessment practitioners*. Unpublished final report, South Africa.

Foxcroft, C.D. & Roodt, G. (2005). *Introduction to psychological assessment in a South African context* (2nd revised Ed.). Cape Town: Oxford University Press Southern Africa.

Galotti, K.M. (2004). *Cognitive psychology: In and out of the laboratory* (3rd Ed.). Belmont: Wadsworth.

Gregory, R.J. (2004). *Psychological testing: History, principles, and applications* (4th Ed.). Boston: Pearson.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Dordrecht: Kluwer.

Hoge, R.D. (1999). *Assessing adolescents in educational, counseling and other settings*. London: Lawrence Erlbaum Associates.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185.

Howell, D.C. (1997). *Statistical methods for psychology* (4th Ed.). Johannesburg: Duxbury.

Huysamen, G.K. (1996). Fair and unbiased admission procedures for South African institutions of higher education. *South African Journal of Higher Education*, 10(2), 199–207.

Jensen, A.R. (1981). *Straight talk about mental tests*. London: The Free Press.

Kaiser, H.F. (1961). A note on Guttman's lower bound for the number of common factors. *British Journal of Statistical Psychology*, 14(1), 1.

Kendell, I.M., Verster, M.A. & Von Mollendorf, J.W. (1988). Test performance of blacks in South Africa. In S.H. Irvine & J.W. Berry (Eds.) *Human abilities in cultural context*. Cambridge: Cambridge University Press.

Kerlinger, F.N. & Lee, H.B. (2000). *Foundations of behavioral research* (4th Ed.). New York: International Thomson Publishing.

Kline, P. (1993). *The handbook of psychological testing*. London: Routledge.

Maller, S.J. & French, B.F. (2004). Universal nonverbal intelligence test factor invariance across deaf and standardization samples. *Educational and Psychological Measurement*, 64, 647–660.

Marczyk, G., DeMatteo, D. & Festinger, D. (2005). *Essentials of research design and methodology*. Hoboken: Wiley.

McCrae, R.R. (2000). Trait psychology and the revival of personality in culture studies. *American Behavioral Scientist*. [Electronic]. Retrieved August 30, 2005, from the World Wide Web: http://www.epnet.com/ehost/login/html.

Meiring, D., Van de Vijver, F.J.R., Rothmann, S. & Barrick, M.R. (2005). Construct, item, and method bias of cognitive and personality tests in South Africa. *South African Journal of Industrial Psychology*, 31(1), 1–8.

Murphy, K.R. & Davidshofer, C.D. (2001). *Psychological testing principles and applications* (5th Ed.). Englewood Cliffs: Prentice Hall.

Naudé, J.L.P. & Rothmann, S. (2004). The validation of the Maslach burnout inventory human services survey for

emergency medical technicians in Gauteng. *South African Journal of Industrial Psychology*, 30(3), 21–28.
Neuman, W.L. (1997). *Social research methods: Qualitative and quantitative approach* (3rd Ed.). Boston: Allyn and Bacon.
Owen, K. (1992). *Test-item bias: Methods, findings and recommendations*. Pretoria: Human Sciences Research Council Group: Education.
Owen, K. (1998). *The role of psychological tests in education in South Africa: Issues, controversies and benefits*. Pretoria: Human Sciences Research Council.
Republic of South Africa. (2006). The Employment Equity Act, Act 55 of 1998. *Government Gazette* No 19370, 24 June 2006.
Rogers, T.B. (1995). *The psychological testing enterprise: An introduction*. Pacific Grove: Brooks/Cole.
Rudner, L.M., Getson, P.R. & Knight, D.L. (1980). Biased item techniques. *Journal of Educational Statistics*, 5(2), 213–233.
Samuda, R.J., Feuerstein, R., Kaufman, A.S., Lewis, J.E., Sternberg, R.J. & Associates (1998). *Advances in cross-cultural assessment*. Thousand Oaks: Sage.

Schaap, P. (2001). *The psychometric properties of the SpEEx (an updated version)*. Unpublished report. University of Pretoria.
Schaap, P. & Basson, J.S. (2003). The construct equivalence of the PIB/SpEEx motivation index for job applicants from diverse cultural backgrounds. *South African Journal of Industrial Psychology*, 29(2), 49–59.
Schepers, J.M. (1992). *Toetskonstruksie: Teorie en praktyk*. Johannesburg: RAU-Drukpers.
Sharrat, P. (1987). Thinking. In G. A. Tyson (Ed.) *Introduction to psychology: A South African perspective*. Johannesburg: Westro Educational Books.
Shillaw, J. (1996). *The application of the Rasch modelling to yes/no vocabulary tests.* [Electronic]. Retrieved August 30, 2005 from the World Wide Web: http://www.swan.ac.uk/cals/calsres/vlibrary/js96a.htm.
Skrondal, A. & Rabe-Hesketh, S. (2005). *Structural equation modeling: Categorical variables.* [Electronic]. Retrieved August, 2008, from the World Wide Web: http://www.gllamm.org/SEMcat.pdf.
SPSS Inc. (2006). *SPSS for Windows, version 15 (computer software).* Chicago: SPSS Inc.
Tabachnik, B.G. & Fiddell, L.S. (1989). *Using multivariate statistics* (2nd Ed.). New York: Harper Collins.
Taylor, T.R. (1997). *Administrator's manual for APIL battery*. Parktown: Jetline.
Van der Merwe, R.P. (1999). Psychological assessment in industry. *South African Journal of Industrial Psychology*, 25(3), 8–11.
Van de Vijver, F.J.R. (1998). *Towards a theory of bias and equivalence.* [Electronic]. Retrieved October 7, 2006, from the World Wide Web: http://www.gesis.org/publikationen/Zeitschriften_spezial/documents/znspezial3/znspez3_02_vijver.pdf.
Van de Vijver, F.J.R. & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Newbury Park: Sage.
Van de Vijver, F.J.R. & Rothmann S. (2004). Assessment in multicultural groups: The South African case. *South African Journal of Industrial Psychology*, 30(4), 1–7.
Waller, N.G. (1995). *MicroFACT 2.0: A microcomputer factor analysis program for ordered polytomous data and mainframe size problems.* St Paul: Assessment Systems Cooperation.
Wallis, T. & Birt, M. (2003). A comparison of native and non-native English-speaking groups' understanding of the vocabulary contained within the 16PF (SA92). *South African Journal of Psychology*, 33(3), 182–190.