Construct equivalence of the OPQ32n for Black and White people in South Africa

Authors:
Deléne Visser[1]
Rian Viviers[1]

Affiliations:
[1]Department of Industrial and Organisational Psychology, University of South Africa, South Africa

Correspondence to:
Rian Viviers

email:
vivieam@unisa.ac.za

Postal address:
Department of Industrial and Organisational Psychology, University of South Africa, Gauteng 0003, South Africa

Keywords:
construct equivalence; Occupational Personality Questionnaire; personality; racial and cultural differences; test bias

Dates:
Received: 18 Aug. 2008
Accepted: 30 Sept. 2009
Published: 27 May 2010

How to cite this article:
Visser, D., & Viviers, R. (2010). Construct equivalence of the OPQ32n for Black and White people in South Africa. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde, 36*(1), Art. #748, 11 pages. DOI: 10.4102/sajip.v36i1.748

This article is available at:
http://www.sajip.co.za

SA Journal of Industrial Psychology
Article #748

ABSTRACT

Orientation: The construct equivalence of the Occupational Personality Questionnaire (OPQ32n) for black and white groups was investigated.

Research purpose: The objective was to investigate the structural invariance of the OPQ32n for two South African population groups.

Motivation for the study: The OPQ32n is often used for making a variety of personnel decisions in South Africa. Evidence regarding the suitability of personality questionnaires for use across South Africa's various population groups is required by practitioners for selecting appropriate psychometric instruments.

Research design, approach and method: Data were collected by means of a questionnaire and the results were analysed using quantitative statistical methods. The sample consisted of 248 Black and 476 White people from the SHL (South Africa) database. Structural equation modelling was used to examine the structural equivalence of the OPQ32n scale scores for these two groups.

Main findings: A good fit regarding factor correlations and covariances on the 32 scales was obtained, partially supporting the structural equivalence of the questionnaire for the two groups. The analyses furthermore indicated that there was structural invariance, with the effect of the Social Desirability scale partialled out.

Practical/managerial implications: The present study focused on aspects of structural equivalence only. The OPQ32n therefore passed the first hurdle in this particular context, but further investigation is necessary to provide evidence that the questionnaire is suitable for use in personnel decisions comparing the population groups.

Contribution: Despite the positive findings with regard to structural equivalence and social desirability response style, it should be borne in mind that no assumptions regarding full scale equivalence can be made on the basis of the present findings.

INTRODUCTION

No practice in modern psychology has been assailed more than psychological testing, because test bias and fairness have become controversial topics internationally in the broader contexts of cultural and sexual bias (Gregory, 2007). As a result of the globalisation and migration of the workforce, the multicultural nature of populations has become more prominent in many countries worldwide, particularly during the past two decades. These phenomena pose challenges to the practice of psychological assessment (Van de Vijver & Rothmann, 2004). Anastasi and Urbina (1997) indicated that, internationally, the design of selection strategies for fair test use with cultural minorities has emerged as a new focal point. Decision models are being proposed that have the effect of selecting larger proportions of persons from lower-scoring groups (Cascio & Aguinis, 2005). Such decision models have as their goal that which is generally designated by terms such as 'affirmative action' or the reduction of 'adverse impact' in the selection process.

The cultural appropriateness of psychological tests and their usage were placed in the spotlight in South Africa with the promulgation of the Employment Equity Act No. 55 of 1998, specifically Section 8 (Republic of South Africa, 1998). Since the Act was promulgated, the issues of the culture fairness and test bias of psychological instruments became points of continuous concern (Van de Vijver & Rothmann, 2004). Instead of resting with potential complainants, the onus of proof has shifted to psychologists using psychological instruments to prove that those instruments adhere to the regulations of the Employment Equity Act. The South African law requires psychologists to be proactively involved by providing evidence that tests are unbiased and can be used in a fair manner (Van de Vijver & Rothmann, 2004). Therefore, there is a need for measuring instruments that meet the specified requirements so that psychological tests can be used for all cultural and language groups in South Africa. One of the main goals of the assessment profession in South Africa is (and should be) to endeavour to align current practice with legal demands, through the development of new instruments and the validation of existing ones for use in the multicultural society (Foxcroft, 2004; Meiring, Van de Vijver, Rothman & Barrick, 2005; Van de Vijver & Rothmann, 2004).

Crocker and Algina (1986) referred back to the 1960s, when issues involved in using tests to select minority applicants for jobs began to receive attention. The possibility of bias in test scores was an issue for test developers and users only. Since then, these issues have begun to receive much more attention and the matter has become a burning issue within psychological testing.

Currently several documents exist that provide guidelines for assessing the psychometric soundness of psychological tests, such as the American Psychological Association (APA) *Standards for educational* *and psychological tests* (1999), the Society for Industrial and Organisational Psychology of South Africa (SIOPSA) *Guidelines for the validation and use of assessment procedures for the workplace* (2005), *Psychological test use in South Africa* (Mauer, 2002) and *Applied psychology in human resource management* (Cascio & Aguinis, 2005). From these sources it is clear that psychologists have to consider the indicators and guidelines and that every endeavour should be made to address scientifically the psychometric bias properties of tests and the fairness of the uses of tests. In a discussion of the APA standards, Huysamen (2002) pointed out the conceptualisation of construct validity as the primary objective in test validation. Mauer (2002) emphasised the possible juridical and professional consequences if psychometric requirements for tests are ignored. He also stressed that the procedures used in any form of adjudging, appraisal, assessment, evaluation, valuation, grading, ranking, classifying, categorising, placing, positioning or rating, insofar as it deals with employees, should be shown scientifically to be reliable, valid and unbiased. Again, the importance of establishing sound psychometric evidence is emphasised in this reminder.

Fairness, bias and equivalence

Gregory (2007) distinguished clearly between 'test fairness' and 'test bias', but pointed out that the two terms are often wrongly considered to be interchangeable. This is a common misconception, because test fairness is a broad concept that recognises the importance of social values in test usage (a values concept), whereas test bias refers to objective statistical indices that examine the patterning of test scores for relevant subpopulations (a statistical concept). Test developers can therefore control test bias, but they cannot control test fairness, because the fair use of tests and the decisions taken as a consequence of testing are in the hands of test users.

The various selection strategies for fair test use for addressing affirmative action or the reduction of adverse impact referred to above cannot be realised solely by producing unbiased tests. Although it is true that any form of bias, including lack of construct equivalence between groups, may, and probably will, result in discriminating personnel decision making, the converse unfortunately does not hold true. Sections 15 and 20(3) of the South African Employment Equity Act No. 55 of 1998 (Republic of South Africa, 1998) define affirmative action measures as the means employed to 'ensure the equitable representation of suitably qualified people from the designated groups'. Such measures call for selection decision models that are not dictated by the inherent psychometric properties of measuring instruments. By using meticulously constructed tests one therefore cannot ensure compliance with the goals of the Employment Equity Act. The judicious use of reliable, valid and unbiased tests is a necessary, but not sufficient, prerequisite for fairness in testing. Because the focus of the present paper is on a specific psychometric aspect of bias, namely construct equivalence, further reference to fairness or culture fairness in testing is avoided.

Cole and Moss (1989, p. 205) defined test bias as being present 'when a test score has meanings or implications for a relevant, definable subgroup of test takers that are different from the meanings or implications for the remainder of the test takers'. This definition of bias implies that test scores obtained for various subgroups of a given population cannot be interpreted in the same way across the groups. Cole and Moss (1989) proposed that bias is differential validity in the case of a given interpretation of test scores for specific subgroups of a population. Gregory (2007) agreed with this interpretation by equating test bias with differential validity. He distinguished between three different types of bias, namely bias in content validity, bias in predictive or criterion-related validity and bias in construct validity, when comparisons between specific subgroups of populations are being made.

To illustrate the existence of various theoretical viewpoints regarding test bias, the definition of Cascio and Aguinis (2005) for differential validity deserves mention. On considering test bias regarding employment decisions, they held a somewhat more restricted view on differential validity than that advocated by Cole and Moss (1989) and Gregory (2007). Cascio and Aguinis (2005) described differential validity as a form of test bias that is the result of differences in the magnitudes of the criterion-related validity coefficients for the various subgroups being compared. For a proper assessment of bias, they recommended that the possible presence of predictive bias (or differential prediction) should rather be investigated (Cascio & Aguinis, 2005). This entails an examination of possible differences in standard errors of estimate for the subgroups, and in the slopes and intercepts of the subgroups' regression lines, an approach also supported by Geisinger (1994).

The present study deals specifically with bias in construct validity and it is acknowledged that construct validity is a broad concept. The definition offered by Reynolds appears to be logically acceptable, namely, bias with regard to construct validity exists

> *when a test is shown to measure different hypothetical traits (psychological constructs) for one group than for another; that is, differing interpretations of a common performance are shown to be appropriate as a function of ethnicity, gender, or of another variable of interest.*
>
> (Reynolds, 1998, cited in Gregory, 2007, p. 274)

Essential criteria for the non-bias of a test that follow from this definition are that there should be an equal number of underlying factors for the various subgroups and that the item or subscale loadings should be similar for the population subgroups, that is, factorial invariance across the groups is required (Gregory, 2007).

Recent research by Poortinga, Van de Vijver and others (Poortinga, 1989; Van de Vijver & Leung, 1997; Van de Vijver & Leung, 2000; Van de Vijver & Poortinga, 1997; Van de Vijver & Tanzer, 1997) has suggested a taxonomy of bias and equivalence that provides a framework for examining bias that is more comprehensive and less simplistic than the approaches mentioned earlier.

Van de Vijver and Tanzer (1997) and Van de Vijver and Leung (1997) noted that bias (or non-equivalence) is present when there are score differences between subgroups on the measurements of a particular construct (such as the items of a test) that do not correspond to differences between the subgroups in the underlying trait or ability. Bias is defined as the opposite of equivalence, although the term bias generally tends to refer to nuisance factors in cross-cultural comparisons between groups, whereas equivalence is generally associated with a hierarchy of measurement levels regarding cross-cultural score comparisons (Van de Vijver & Leung, 1997). Equivalence, therefore, indicates the measurement level at which the scores obtained for different groups can be compared.

Equivalence and bias are the fundamental concepts when comparisons between subgroups of populations or cross-cultural comparisons are made, because inferences based on biased (or non-equivalent) scores are invalid. Measuring instruments that are used for various cultural groups, such as those found in South Africa, should therefore be assessed in terms of bias and equivalence for score comparisons between the groups. It is important to note that the concepts bias and equivalence do not refer to properties inherent in any particular measuring instrument. These concepts deal with the characteristics of an instrument in a (specific) comparison between groups (such as groups from different cultures), rather than with the intrinsic properties of the measuring instrument (Van de Vijver & Tanzer, 1997).

Three kinds of bias are distinguished in the taxonomy, namely construct, method and item bias (differential item functioning) (Van de Vijver & Leung, 1997; Van de Vijver & Tanzer, 1997). The definition for construct bias is similar to that proposed by Reynolds (1998, cited in Gregory, 2007) and occurs when the construct measured is not identical across the various subgroups being compared. A comprehensive evaluation of bias for a particular comparison requires an integrated and extensive examination of all aspects of bias. There are many procedures and statistical techniques that can be used for this purpose before claims can be made about a lack of all types of bias.

The hierarchy of three different levels of equivalence deals with the level of measurement implicit in any specific comparison between groups (Van de Vijver & Leung, 1997; Van de Vijver & Tanzer, 1997). Direct comparisons between the descriptive statistics of groups are in order only when the scores of the various groups are on the same measurement scale and when the same construct is measured in the groups. When using common psychometric tests in the employment domain across population groups, as is usually the case in South Africa, the overall goal is to use tests that yield directly comparable results.

Construct equivalence

At the bottom of the hierarchy we find the level of construct equivalence, also labelled structural invariance, structural equivalence or functional equivalence. Construct equivalence exists when the same construct is measured in the various groups being studied, whereas construct inequivalence occurs when an instrument measures different constructs in the groups, or when the measured construct overlaps only partially across the groups. Construct equivalence is often assessed by means of exploratory factor analysis with target rotation, by determining the similarity of exploratory factor analysis results by means of the coefficient of congruence, or by structural equation modelling. These equivalence concepts should be distinguished from the concept construct validity, which is the extent to which a measure shows a pattern of high correlations with measures that are expected to measure the same construct (convergent validity), as well as low correlations with measures of other constructs (discriminant validity). Construct validity may be assessed, *inter alia*, by means of examining patterns of correlations, as indicated before, by using the multi-trait–multi-method approach, by experimental means, or by using exploratory and confirmatory factor analysis.

Measurement unit equivalence

Measurement unit equivalence is the next level of equivalence and occurs when two or more measures have the same measurement unit, but might have different origins. An example cited most often is the measurement of temperature in which the Celsius or Kelvin scales are used, because the origins of these two scales differ by 273 degrees, but one degree on the Celsius scale has the same meaning as one degree on the Kelvin scale. Direct score comparisons can be made only when the differences between the origins on the scales are known, a rare occurrence in psychological research (Van de Vijver & Leung, 1997).

Full scale equivalence

Also referred to as or scalar equivalence, full scale equivalence is found at the top of the hierarchy. It occurs when measures have the same measurement unit and the same origin. This level of equivalence allows direct comparisons across population subgroups or across cultures. It is important to note that full scale equivalence can be attained only when measurement is entirely bias free, that is, when there is no construct, method or item bias. For psychological variables, full scale equivalence cannot be proven directly. It has to be assessed indirectly by means of the available methods for studying bias. When the research question deals with the constructs measured in the comparison groups, construct equivalence is all that is required and this level of equivalence will not be affected by method or item bias (Van de Vijver & Leung, 1997). However, when the aim is to directly compare the means obtained in the groups or directly compare scores of individuals belonging to the various groups, such as for personnel decision making, full scale equivalence must be present. In the current study, the focus is on construct equivalence as a first step in the assessment of bias in applications of a particular measuring instrument.

In South Africa, with its multicultural society, it has long been recognised that testing poses special problems for test developers and users (Foxcroft, 2004; Van de Vijver & Rothmann, 2004; Wallis, 2004). There clearly is a dearth of evidence that indicates that tests being used across population groups are free from bias, because far too few studies have been published that investigated the possible presence of test bias in general or construct bias in particular. Of particular relevance here is that a test that does not measure what it proposes to measure across subgroups invalidates all inferences drawn from the test results (Wallis, 2004).

Personality testing and the OPQ32

There has been a substantial increase in the use of personality and related tests when hiring for a broad spectrum of jobs (Clevenger, Pereira, Weichmann, Schmitt & Harvey, 2001; Ones & Anderson, 2002; Saville & Willson, 1991). Recent surveys have indicated unequivocally that the use of personality tests is becoming increasingly popular among employers for personnel selection decisions (Ones & Anderson, 2002).

Personality tests are also used widely in South Africa, but establishing comparability across groups is vital in a country where people from a variety of cultural or demographic groups compete for job opportunities (Bedell, Van Eeden & Van Staden, 1999; Meiring *et al*., 2005). Yet few attempts have been made to test the comparability of results for different cultural groups (Van de Vijver & Rothmann, 2004). Van de Vijver and Rothmann (2004) concluded that much more research is needed on the equivalence and bias of assessment tools before psychology as a profession can live up to the demands implied in the Employment Equity Act.

A variety of factors can cause group differences in test scores, such as race, culture, socio-economic status, education, language and cognitive style (Meiring *et al*., 2005). Many tests that are used across South African population groups are, at present, administered in English only. Apart from other possible cultural nuisance variables, there is evidence that the level of proficiency in the English language affects performance in cognitive and personality tests (Abrahams & Mauer, 1999a; Claassen, 1993; Foxcroft & Aston, 2006; Koch, 2007; Owen, 1989, Van Eeden & Van Tonder, 1995; Van Eeden & Visser, 1992). Evidence of construct bias when tests have been administered in English only has been found by Meiring *et al*. (2005), Abrahams and Mauer (1999b) and Koch (2007). Construct bias also resulted when (mostly) Black students had to complete Schepers's Locus of Control Inventory in a second language, whereas (mostly) White students could complete the questionnaire in their mother tongue (Berg, Buys, Schaap & Olckers, 2004; Schaap, Buys & Olckers, 2003). The same data set was used for both studies.

Nevertheless, there also are examples of research in South Africa that reported construct equivalence across groups where tests were administered in English only. Schaap and Basson (2003) found evidence that the constructs measured by the PIB/SpEEx Motivation Index, namely internal locus of control and external locus of control, were equivalent for Black, Asian and White entry-level job applicants. Vorster, Olckers, Buys and Schaap (2005) investigated the equivalence of the

SA Journal of Industrial Psychology

Article #748

structural model of the Job Diagnostic Survey and reported that the model held for Black and White groups. Only 22% of the respondents completed the questionnaire in their mother tongue (English) and it should be noted that approximately 69% of the White group did not complete the questionnaire in their mother tongue because a large percentage were Afrikaans speaking. It is not evident that the same results would have been obtained had the sample been split into first- and second-language groups.

In another study, Coetzer and Rothman (2007) found evidence that supported the hypothesised dimensionality of the constructs burnout (as measured by the Maslach Burnout Inventory – General Survey) and work engagement (as measured by the Utrecht Work Engagement Scale) for two groups, one with English as home language and the other with Afrikaans, or an African language, as home language. They also found that construct equivalence existed for the two groups when certain items were deleted from the data. It should be noted that the majority (76%) of the Afrikaans/African group consisted of Afrikaans speakers, with the implication that the results did not provide convincing evidence for African-language speakers.

A number of studies have focused on the comparability of personality measures for different population groups in South Africa, with mixed results. For instance, Taylor and Boeyens (1990) found some support for construct equivalence for the South African Personality Questionnaire (SAPQ), but it was evident that the instrument suffered from item bias in their application of comparing the scores of White and Black respondents. Research by Van Eeden, Taylor and Du Toit (1996) and Abrahams (1997) indicated that two versions of the Sixteen Personality Factor Questionnaire (the 16PF5 and the 16PF, SA92) may not be suitable for individuals who do not have English as first language. Van Eeden and Prinsloo (1997) reported some degree of construct equivalence for the 16PF, SA92, but cautioned that there were differences between the factor loadings of the second-order factors for Black and White people. Heuchert, Parker, Stumpf and Myburgh (2000) investigated the construct equivalence of the NEO Personality Inventory – Revised (NEO PI-R) and found a clear five-factor structure for Black and White students that conformed to the five-factor model (FFM) of personality. In another context, Taylor (2000) found that the openness factor could not be extracted for black employees, whereas the factor structure found for White employees was in line with expectations regarding the FFM.

The most extensive South African bias study to date was conducted by Meiring *et al*. (2005) using a sample of 13 681 applicants from 12 different cultural groups for entry-level jobs in the South African Police Service. One of the measuring instruments included in the study was the 15FQ+ Personality Questionnaire, which was developed for use in industrial and organisational settings. The alpha coefficients for some of the factors were exceptionally low, particularly for the Black language groups. Furthermore, exploratory factor analysis with target rotation to a pooled solution of 15FQ+ factors yielded poor agreement with the factors of the Ndebele, White, Indian and Coloured groups, thereby indicating structural or construct inequivalence. In addition, significant item bias was found for many items, although a medium effect size was obtained for one item only. Meiring *et al*. (2005) also found that neither removal of the biased items nor cognitive/English-language ability or social desirability affected the magnitude of the cross-cultural differences observed.

The present study is yet another attempt to investigate the structural equivalence across population groups of a personality questionnaire in a South African context. In this instance we focused on a personality questionnaire that is currently being used extensively in South African organisations, namely the Occupational Personality Questionnaire (version OPQ32n), because no research results on this issue have been published regarding the OPQ32n. The main aim with the development of the OPQ32 was to provide an instrument that would give a comprehensive, detailed description of personality likely to be relevant in occupational contexts for the selection, development and counselling of predominantly managerial-level staff.

The OPQ32 is based on an occupational model of personality that describes 32 dimensions or scales of individuals' preferred or typical styles of behaviour at work. In addition, it includes a Social Desirability scale. The model consists of three domains, (1) relationships with people, (2) thinking style and (3) feeling and emotions. The three domains are joined by a potential fourth – the dynamism domain – which relates to sources of energy (OPQ32 Technical Manual, 2006). There are two questionnaires for measuring personality using the above model, namely the OPQ32n (normative) and OPQ32i (ipsative).

With regard to the comparability of the OPQ for different population groups, it was found, in a study conducted in the United Kingdom (UK), that the questionnaire's internal consistency reliabilities for a combined sample of Black and Asian respondents was lower than that for a White sample (OPQ32 Technical Manual, 2006). Furthermore, it was found that, for a sample from the general population, an analysis of background information showed that there was a higher proportion of the ethnic minority sample with poor education than in the White group, possibly resulting in less accurate responses. The mean reliability for the Black and Asian sample was equal to 0.70.

When the OPQ32 mean scale scores of White and minority ethnic groups were compared in the UK, only nine of the mean differences reached statistical significance. The largest of these differences was on 'achieving', with a medium effect size (d = 0.43). These results were ascribed to the occupational relevance of the OPQ32 content, together with the straightforward way in which items are phrased. This means that people from different demographic backgrounds were able to relate to the questionnaire in a similar manner (OPQ32 Technical Manual, 2006). It was earlier argued that such results may also not necessarily be obtained in relation to the various South African population groups.

It is clear that language of administration, race and culture may be among the main factors impacting on the construct comparability of personality tests and that these factors are particularly salient in contemporary South Africa.

The objective of the present study, therefore, was to investigate the structural invariance of the OPQ32n for two South African population groups. It was also decided to examine differences in OPQ32n scale scores between Black and White demographic groups and to establish whether these were likely to arise from a lack of construct equivalence between the two groups.

RESEARCH DESIGN

Research approach

Data were collected by means of a questionnaire and the results were analysed using quantitative statistical methods.

Research method

Research participants

The data were collected from various South African companies using the OPQ32 normative version (OPQ32n) for the selection and development of their personnel. The original population on record at SHL South Africa consisted of 1579 respondents, of whom 248 were Black, 29 were Coloured, 37 were Indian and 476 were White. Sixteen respondents indicated another population group, whereas 773 candidates did not indicate which population group they belonged to. The latter group of

TABLE 1
Biographical information for the White and Black groups (N = 724)

	Black respondents		White respondents		Total Sample
Age*					
n	221		397		618
Mean	32.13		30.99		31.4
SD	8.00		8.66		8.44
	n	%	*n*	%	
Gender					
Female	90	36.3	198	41.6	
Male	158	63.7	278	58.4	
Total	248	100.0	476	100.0	
Education					
Matric	43	17.3	86	18.1	
Post-matric certificate	12	4.8	23	4.8	
Degree	85	34.3	133	27.9	
Postgraduate	103	41.5	224	47.1	
Missing	5	2.0	10	2.1	

*Age were obtained from the respondent's South African identity documents

773 respondents was excluded from the sample because they could not be included in comparisons between the population groups. Due to the small sizes of the Coloured and Indian groups, we decided to compare the Black (*n* = 248) and White (*n* = 476) groups only. These two groups constituted the majority of the original sample and it was considered prudent to omit the influence of smaller demographic groups.

Biographical information for the sample of 724 respondents appears in Table 1. Their ages ranged from 19 to 65, with a mean age of 31.40 (SD = 8.44). There were 288 women (39.78%) and 436 men (60.22%) in the sample. All the respondents reported an educational level of matric or higher; in fact, 545 of them (76.87%) held a first or higher degree. There were 15 missing values with regard to educational level. The comparability of the Black and White groups is important for a study on measurement equivalence. Table 1 provides information regarding the biographical characteristics of the two groups. It appeared that the groups did not differ significantly on any of the variables.

Measuring instrument

All the respondents had completed the normative version of the OPQ32 at the request of their respective organisations and the questionnaires were scored by SHL South Africa. This OPQ32 version was chosen because it is often used in developmental and counselling applications in the industry and in practice. Furthermore, the results for the ipsative version, where forced choices have to be made, are not suitable for factor analysis (Baron, 1996; Dunlap & Cornell, 1994; Johnson, Wood & Blinkhorn, 1988; Kerlinger & Lee, 2000; Visser & Du Toit, 2004). As recommended by SHL, the ipsative version is the version of the OPQ32 used most frequently, particularly for selection, because of the hypothesis that socially desirable responding will bias individual responses.

On some of the 32 scales, a high score is indicative of a positive outcome for the scale, whereas a low score indicates a favourable description within parameters of the work context on other scales. A specific personality style is not, in itself, good or bad, but appropriate or inappropriate depending on the circumstances.

In South Africa, SHL makes norms available for a total population of South Africans, but not for separate population groups. The internal consistency reliabilities for the scales ranged from 0.65 to 0.87 (median = 0.79) for a general population sample (*N* = 2028) in the UK (OPQ32 Technical Manual, 2006). The alpha coefficient for the Social Desirability scale was equal to 0.63. In South Africa, a reliability study on a composite sample of 1181 employees and students resulted in alpha coefficients ranging from 0.69 to 0.88. The sample included 19.64% Black people, 2.71% Asian people, 2.29% Coloured people and 33.02% White people (42.34% of the respondents did not indicate their ethnic origin) (SHL South Africa, 2002).

Test-retest reliability was established in the UK using a sample of 107 undergraduates at various higher education institutions (OPQ32 Technical Manual, 2006). After one month, the reliabilities ranged from 0.64 to 0.91, with a median of 0.79. No test–retest studies have been done in South Africa.

In terms of construct validity, in the UK it was found that the scale intercorrelations for the OPQ32n ranged from -0.51 to 0.56, with two-thirds of the correlations falling between -0.2 and 0.2 (OPQ32 Technical Manual, 2006). This suggests a relatively high degree of independence for most of the scales, despite the large number of narrow scales included. Seventy-seven per cent of the OPQ32n scale pairs shared less than 10% common variance, but there were some pairs of scales that were highly related. The OPQ32n was also subjected to exploratory factor analysis, and principal components extraction followed by varimax rotation gave the clearest results. Six factors were extracted in four different data sets (two from the United Kingdom and one each from the United States and South Africa), explaining 51% – 53% of the total variance in the respective data sets. In interpreting these factors, comparisons were made with the 'Big Five' model of personality of McCrae and Costa (1987). Five of the derived clusters of dimensions, namely extraversion, agreeableness, conscientiousness, neuroticism and openness to experience, clearly represented typical Big Five descriptions (OPQ32 Technical Manual, 2006). The sixth dimension was not consistent across the samples, but in the South African sample it related to adaptability. In another South African study, Visser and Du Toit (2004) obtained a six-factor solution that included the Big Five factors plus a factor labelled as Interpersonal Relationship Harmony, which was likened to the concept of *ubuntu*.

The criterion validity of the OPQ32 has been verified in many studies in the UK and elsewhere (OPQ32 Technical Manual, 2006). In these studies, OPQ32 results were correlated with indicators of performance of various kinds, generally managers' ratings of competence. With total sample sizes exceeding 6000 for an earlier version of the OPQ and 2500 for the OPQ32, they provide a robust body of evidence to support the occupational use of the OPQ32 questionnaire because the patterns of relationships found in the studies provided strong support for the criterion validity of the OPQ32 (OPQ32 Technical Manual, 2006).

The Social Desirability scale of the OPQ32n measures the extent to which a person is more/less self-critical in responses and more/less concerned with making a good impression. Socially desirable responding has been shown to be more prevalent among black than white populations in UK and USA standardisation samples (OPQ32 Technical Manual, 2006). However, in a South African study, Visser (2002) found that Black and White groups did not differ statistically significantly with regard to their scores on the Social Desirability scale of the OPQ5.2 Concept Model, an earlier version of the currently used instrument. On the basis of these conflicting results, it was decided to test for structural invariance with and without the effect of Social Desirability partialled out.

Research procedure

The administration of the OPQ32n was done in a number of South African companies, using the paper-and-pencil version, or completing the questionnaire online. Psychometrists or trained OPQ32 staff administered the questionnaires. All responses were captured on a SHL South Africa database. The data required for the analyses for the current study were extracted from this database.

Statistical analysis

In this section, the rationale for, and procedure of, the structural equation modelling used in the current study is explained. The OPQ32n is focused on measuring multiple narrow traits that are important for certain domains of interest (such as job competencies), rather than broad personality factors. Factor analytic research in which the Big Five personality factors are extracted from the OPQ typically explain only approximately 50% of the variance (OPQ32 Technical Manual, 2006; Visser & Du Toit, 2004). Conceptually, only 25 out of 32 traits of the OPQ are related to the Big Five (OPQ32 Technical Manual, 2006). This implies that a higher-order factor model fits the data poorly. There is no merit in comparing a structure across cultures that does not fit in the reference group in the first place, because a good fit for one group has to be achieved first before doing multi-group analyses (Byrne, 2006). Therefore, it is meaningless to follow a commonly used procedure for establishing factorial invariance by comparing the factor structures between the groups in this study. Instead, the question that we wished to answer was whether the 32 narrow constructs are equivalent in two groups through comparing their nomological networks (theoretical concepts and their relationships with the other constructs). Such networks are represented in the trait correlation matrix (Clark & Watson, 1995; Cronbach & Meehl, 1955). Comparing correlation matrices directly provides a suitable global test of equivalence in cases in which there is no predefined factor structure to be fitted to the measured constructs (Bentler, 2005). Furthermore, the trait correlation matrix provides all the necessary information for factor analysis and, therefore, serves as a necessary condition for the equivalence of higher-order factor structures.

Since the statistical theory is based on covariance matrices, a special set-up procedure is required to model the correlation matrices correctly in EQS. Bentler described the global test of equivalence for correlation matrices as follows:

> *Let* Y_1 *be the vector of observed variables in the first group, and let the population covariance be* Σ_1*. The population correlation matrix is* P_1 *and* $\Sigma_1 = D_1P_1D_1$*, where* D_1 *is the diagonal matrix of standard deviations of the variables. Thus if we present* Y_1 *as* $Y_1 = D_1X_1$*, then it is apparent that the covariance matrix of* X_1 *is* P_1.
>
> (2005, p. 152)

The logic of the EQS model is, therefore, to present the observed variables Y_1 (unstandardised OPQ scale scores) through dummy factors X_1 ($Y_1 = D_1X_1$), with factor loadings equal to the scales' observed standard deviations (D_1) and no measurement error. Dummy factors X_1 are standardised so that their variances are 1 and their covariance matrix is therefore a correlation matrix. There are as many such equations as variables and the factor loadings are estimated freely. Each of the dummy factors' variances has to be fixed to 1 in the model for their covariance matrix to become the correlation matrix. The covariances of the factors are free parameters and, corresponding to the off-diagonal elements of P_1, they are correlations between the observed variables.

The same type of set-up applies to the second group, where $\Sigma_2 = D_2P_2D_2$, and consists of the following two steps:

- To evaluate the first hypothesis, that $P_1 = P_2$ (i.e. that the correlation matrices are equal), cross-group constraints have to be made on the covariances of the dummy factors, but since their variances are constrained to unity, these are in fact equality of correlations. In this first step, the diagonal matrices D_1 and D_2 (i.e. the factor loadings representing the observed scale scores' standard deviations), are not constrained to be equal across groups. Only trait correlations are constrained.
- The second hypothesis is that the scales' standard deviations are also equal across samples and, thus, not just the correlations, but also the covariance matrices, are equal ($\Sigma_1 = \Sigma_2$). This is a stronger hypothesis and requires constraining the factor loadings representing the scales' standard deviations (D_1 and D_2) to be equal across the groups, in addition to the constraints set in the previous step.

These two steps were also repeated with the effect of the Social Desirability scale partialled out by computing partial correlations for every intercorrelation between the 32 OPQ32n scales. The comparisons were performed using the structural equation modelling software EQS Version 6.1 for Windows (Bentler, 1985–2005; Byrne, 2006).

RESULTS

The first step of the analyses entailed computing the means, standard deviations and internal consistency reliabilities of the various OPQ32n scales for the Black and White groups separately and for the total group. Internal consistency was assessed in two ways, namely by computing coefficient alphas and mean inter-item correlations (Clark & Watson, 1995). Subsequently, the magnitude of the differences between the means of the black and white groups on the various OPQ32n scales was assessed. The *d* statistic, which is calculated by standardising the raw effect size as expressed in the measurement unit of the variables by dividing it by the pooled standard deviation of the two groups, was used for this purpose (Cohen, 1988). This statistic therefore expresses score distances in units of variability and is an estimation of the effect size index. The results of these calculations are presented in Table 2.

For the Black sample, the Cronbach alpha values ranged from 0.57 for Conscientiousness and 0.59 for Variety Seeking to 0.85 for Rule Following and Worrying. It is evident that there were only two alpha values marginally lower than 0.60, which is regarded by some as a lower limit for acceptability for internal consistency reliabilities for personality scales in basic and applied research (Clark & Watson, 1995). However, Nunnally (1978) has advocated that 0.70 be regarded as the lower limit during the early stages of research. In total, only eight scales yielded alphas lower than 0.70 for the Black group. For the White group the lowest alpha of 0.71 was obtained for Independent Minded, whereas the highest value of 0.91 was obtained for Tough Minded. For the total group, the alphas ranged from 0.72 for Independent Minded to 0.88 for Worrying and Rule Following. The mean alpha for the Black group on the 32 OPQ scales was 0.74, whereas the mean alpha for the White group was 0.84. For Social Desirability the alphas were 0.66 for the Black group, 0.66 for the White group and 0.68 for the total group. The mean inter-item correlations per scale for the Black group varied between 0.15 and 0.48, whereas those for the White group varied from 0.31 to 0.62.

TABLE 2
Means, standard deviations, alpha coefficients and *d* statistics for the Black and White groups

Scale	Black respondents (*n* = 248)			White respondents (*n* = 476)			Total (*N* = 724)			*d*
	M	SD	*α* (mean r)	M	SD	*α* (mean r)	M	SD	*α*	
Persuasive	23.22	4.92	0.82 (0.43)	21.58	5.4	0.86 (0.50)	22.14	5.29	0.85	0.31*
Controlling	24.31	3.96	0.72 (0.30)	24.39	4.18	0.85 (0.49)	24.36	4.1	0.81	-0.02
Outspoken	25.11	4.3	0.67 (0.21)	22.94	4.9	0.80 (0.33)	23.68	4.81	0.76	0.45*
Independent Minded	21.1	5.09	0.73 (0.25)	21.38	4.38	0.71 (0.24)	21.28	4.64	0.72	-0.06
Outgoing	22.04	5.19	0.80 (0.41)	21.53	5.58	0.90 (0.60)	21.7	5.45	0.87	0.09
Affiliative	23.98	5.27	0.78 (0.31)	23.82	5.59	0.86 (0.45)	23.87	5.48	0.83	0.03
Socially Confident	23.74	3.96	0.68 (0.27)	22.68	5.06	0.88 (0.55)	23.04	4.74	0.83	0.22*
Modest	16.67	4.89	0.84 (0.47)	18.69	4.78	0.87 (0.53)	18	4.91	0.87	-0.41*
Democratic	27.19	4.28	0.72 (0.26)	25.57	4.6	0.79 (0.32)	26.13	4.55	0.77	0.35*
Caring	27.09	4.09	0.65 (0.22)	26.03	4.87	0.81 (0.34)	26.4	4.64	0.77	0.23*
Data Rational	25.26	4.64	0.82 (0.43)	22.43	5.19	0.86 (0.50)	23.4	5.18	0.86	0.54*
Evaluative	26.61	3.64	0.60 (0.16)	26.1	4.26	0.78 (0.31)	26.27	4.06	0.73	0.13
Behavioural	26.5	4.51	0.75 (0.28)	27.57	5.07	0.88 (0.49)	27.2	4.91	0.84	-0.22*
Conventional	16.82	4.66	0.70 (0.23)	18.47	5.71	0.86 (0.34)	17.9	5.43	0.82	-0.30*
Conceptual	24.52	4.49	0.70 (0.23)	24.28	4.75	0.78 (0.31)	24.36	4.66	0.76	0.05
Innovative	24.61	4.01	0.76 (0.35)	23.62	5.15	0.90 (0.60)	23.96	4.81	0.86	0.21*
Variety Seeking	26.1	4.16	0.59 (0.15)	25.19	5.47	0.83 (0.38)	25.5	5.08	0.77	0.18
Adaptable	22.36	4.78	0.79 (0.39)	21.11	4.67	0.79 (0.40)	21.54	4.74	0.79	0.26*
Forward Thinking	25.92	3.73	0.72 (0.31)	24.98	4.37	0.88 (0.55)	25.31	4.18	0.83	0.22*
Detail Conscious	26.32	5.02	0.77 (0.30)	25.54	5.46	0.81 (0.36)	25.81	5.32	0.8	0.15
Conscientious	26.72	3.16	0.57 (0.19)	26.24	3.8	0.79 (0.39)	26.41	3.6	0.73	0.13
Rule Following	21.98	5.27	0.85 (0.48)	20.64	5.39	0.90 (0.59)	21.1	5.38	0.88	0.25*
Relaxed	23.16	4.1	0.72 (0.30)	22.09	5.21	0.90 (0.59)	22.46	4.88	0.85	0.22*
Worrying	18.88	5.12	0.85 (0.48)	20.84	5.14	0.90 (0.60)	20.17	5.21	0.88	-0.38*
Tough Minded	21.99	4.76	0.77 (0.36)	19.71	5.6	0.91 (0.62)	20.49	5.43	0.87	0.42*
Optimistic	29.26	4.2	0.70 (0.24)	27.97	5.44	0.89 (0.51)	28.41	5.08	0.84	0.25*
Trusting	21.81	5.41	0.79 (0.32)	20.85	5.77	0.87 (0.45)	21.18	5.67	0.84	0.17
Emotionally Controlled	20.5	4.5	0.75 (0.34)	20.52	4.95	0.86 (0.52)	20.51	4.8	0.83	0
Vigorous	28.21	4.31	0.79 (0.32)	27.84	4.52	0.83 (0.39)	27.96	4.45	0.81	0.08
Competitive	18.9	5.54	0.84 (0.48)	17.38	4.74	0.84 (0.48)	17.9	5.08	0.85	0.30*
Achieving	28.68	4.32	0.71 (0.25)	28.42	4.69	0.83 (0.40)	28.51	4.57	0.79	0.06
Decisive	17.27	3.88	0.68 (0.24)	19.95	4.31	0.79 (0.38)	19.03	4.35	0.78	-0.61*
Mean *α*			**0.74**			**0.84**			**0.82**	
Social Desirability	**20.31**	**4.74**	**0.66**	**18.04**	**4.29**	**0.66**	**18.81**	**4.57**	**0.68**	**0.50***

*Statistically significant difference between means at the 0.01 level.

The intercorrelations between the 32 OPQ32n scales were computed for each group separately. These two 32 × 32 tables are too large to reproduce here, but are available to interested readers upon request. In addition, the mean intercorrelation coefficient between the 32 OPQ32n scales was computed for each group separately, using absolute values and excluding the main diagonal from the averaging. For the Black group the mean intercorrelation was equal to 0.185 (SD = 0.12) and a strongly similar result was obtained for the White group, namely 0.184 (SD = 0.13).

The *d* statistics for comparing the means of the Black and White groups on the OPQ32n scales varied from negligible (*d* = 0.00) to values representing moderate effect sizes (*d* = 0.54 for Data Rational, *d* = -0.61 for Decisive and *d* = 0.50 for Social Desirability). Apart from these three moderate effect sizes and three more scales approaching the value of 0.50 (*d* = -0.41 for Modest, *d* = 0.42 for Tough Minded, *d* = 0.45 for Outspoken), altogether 14 of the 32 scales yielded small effect sizes, with the remainder being smaller than 0.20.

The second step of the analyses dealt with conducting a global test of the equality of the covariance matrices of the Black (*n* = 248) and White (*n* = 476) groups to investigate the structural equivalence of the OPQ32n for the two groups. The procedure followed was explained in the statistical analysis section. The null hypothesis that $\Sigma_{\text{black}} = \Sigma_{\text{white}}$, where Σ_g is the population variance-covariance matrix, was tested. Because the exact equality of covariance matrices is hard to verify in large samples (Bentler, 1985–2005), the null hypothesis that the correlation matrices of the Black and White groups are equal was also tested as an initial step. This hypothesis implies that the correlation matrices of the measured variables are the same, although the covariance matrices may differ between the groups due to variables not having equal variances. This was achieved by fixing the variances of the dummy factors in the model at one, so that the covariances of the variables were then equal to the correlations.

Firstly, we fitted a model with 32 latent variables, each represented by a single indicator (the observed scale score). In this model, the error variances were fixed to be zero, the factor variances were all fixed to unity, the factor loadings that represented the standard deviations of the observed scores were free, as were the covariances between the factors. This model was then fitted in a multigroup analysis, with all covariances constrained to be equal, thereby testing the equality of the factor correlations.

Thereafter we followed the same procedure, but the 32 factor loadings that had been previously unconstrained between the samples, were subsequently constrained to be equal, thus producing a stronger hypothesis. Again, in a multigroup analysis the 32 factor loadings were constrained to be equal between the two samples. This model tested the equality of the covariance matrices, because the equality of the variances of the observed variables was also tested.

The comparisons were performed using the structural equation modelling software EQS Version 6.1 for Windows (Bentler, 1985–2005; Byrne, 2006). Structural equivalence was therefore tested by establishing whether the patterns of scale intercorrelations (and/or covariances) were equivalent.

In summary, four separate analyses of covariance structures using maximum likelihood estimation were conducted and, in

SA Journal of Industrial Psychology

Article #748

TABLE 3
Analysis of covariance structures goodness-of-fit statistics for the Black and White groups

Model	Chi-square	*df*	*p*	χ^2/df	RMSEA (90% conf. interval)	CFI	SRMR
Comparison of correlation matrices	791.17	496	0.0000	1.595	0.041 (0.035–0.046)	0.961	0.066
Comparison of correlation matrices – Robust method	689.50	496	0.0000	1.390	0.033 (0.027–0.039)	0.970	
Comparison of covariance matrices	972.57	528	0.0000	1.842	0.048 (0.043–0.053)	0.942	0.083
Comparison of covariance matrices – Robust method	904.77	528	0.0000	1.713	0.044 (0.039–0.049)	0.941	
Comparison of correlation matrices, but with effect of social desirability removed	795.84	496	0.0000	1.604	0.041 (0.036–0.046)	0.959	0.067
Comparison of covariance matrices, but with effect of social desirability removed	961.270	528	0.0000	1.821	0.048 (0.043–0.052)	0.943	0.084

the hypothesised models, the latent variables were allowed to correlate with one another. Being the larger sample, the data of the White group were used to represent the hypothesised model. The steps undertaken were as follows: Firstly, the groups were compared with regard to their correlation matrices only and, secondly, they were compared with regard to their covariance matrices on the 32 scales. These analyses were then repeated with the effect of the Social Desirability scale partialled out by computing partial correlations for every intercorrelation between the 32 OPQ32n scales.

Before carrying out these analyses, the data were inspected to establish whether the assumption of the multivariate normality of the data, on which the maximum likelihood method is based, held true for the two samples. Violation of this assumption may render the model chi-square test invalid, such that alternative estimation methods may have to be employed (Byrne, 2006). The EQS structural equation modelling software was the first to introduce a correction for the chi-square statistic developed by Satorra and Bentler (1988) as the so-called 'robust' alternative to conventional maximum likelihood estimation. The robust option should be used whenever distributional assumptions are violated. It provides output statistics, such as the Satorra-Bentler scaled model chi-square, and robust versions of some other fit statistics.

We found clear evidence of deviation from multivariate normality in the data, because the sample statistics for both samples yielded several significant non-zero univariate kurtoses. In addition, the normalised estimate of Mardia's coefficient was equal to 27.89 for the White group and 26.63 for the Black group. Both values are substantially larger than 5, the cut-off beyond which data should be regarded as non-normal (Bentler, 2005). Consequently, the robust method, which requires raw data for its computation, was the desired option in the current study and was carried out on the initial data set. However, the input data for the steps in which social desirability was partialled out consisted of partial correlations only, meaning that robust statistics could not be computed in these instances. Where possible, we report robust results, but we also report the conventional maximum likelihood results for comparative purposes.

The model fit indices that were used were the model or likelihood-ratio chi-square, normed chi-square (χ^2/df), root mean square error of approximation (RMSEA), including its 90% confidence intervals, comparative fit index (CFI) and standardised root mean square residual (SRMR). The results of the structural equation modelling indicate that the null hypotheses of identical covariance matrices for the four separate analyses cannot be rejected, because all of the fit indices, with the exception of the significant model chi-squares, indicated good fit or closely approached well-fitting models. The goodness-of-fit indices are reported in Table 3.

In every case, the statistically significant model chi-square values were the only goodness-of-fit values that consistently did not meet the accepted levels indicative of good model fit. The implication of significant model chi-square values is that hypotheses of identical correlation matrices should be rejected for the four separate analyses. However, this result is obtained often in research and may usually be ascribed to the large size of the sample and/or the lack of model fit (Byrne, 2006; Tabachnik & Fidell, 2001). In the present study, the sample sizes for both groups were substantially larger than sample sizes of 100 to 200, which are regarded as the likely *N* for obtaining non-significant chi-square statistics (Hair, Anderson, Tatham & Black, 1998). Because large samples are required for obtaining precise parameter estimates in the analysis of covariance structures, model chi-square values are regarded as unrealistic criteria on which to base decisions regarding model fit. Nevertheless, the possibility that the model did not fit, as well as the large samples, remain as viable explanations for the results.

The remainder of the goodness-of-fit indices represent a positive picture of very good fitting models. The chi-square/degrees of freedom ratios were smaller than 2 (the limit recommended by Hair *et al*., 1998) in every instance. Also, according to the limits recommended by Carmines and McIver (1981), these comparisons were indicative of good fit. The RMSEA values had magnitudes representing good fit, because they were smaller than 0.05 (Browne & Cudeck, 1993; Hu & Bentler, 1999). In every instance, the 90% confidence intervals were narrow, so that none of the upper values exceeded 0.06 (Hu & Bentler, 1999). In addition, the CFI values were higher than the value of 0.95 that was recommended by Bentler (1990) and Hu and Bentler (1999) as indicative of good model fit, with the exception of the CFIs for the comparison of covariances, where the values of 0.942, 0.941 and 0.943 fell just below the cut-off for well-fitting models. Surprisingly, the use of the robust option generally yielded only marginally better results than the conventional maximum likelihood method, with the largest improvement being for the correlation matrices, where the robust RMSEA value of 0.033 represented an improvement of 0.008 on 0.041.

Finally, the SRMR magnitudes were less than 0.08, which is suggested as the maximum level for acceptance of good fit in the case of comparisons of the correlations (Hu & Bentler, 1999). Once again, the comparison of the covariances marginally missed the criterion for acceptance. The SRMR represents the mean across all standardised residuals, or the mean discrepancy between the correlation matrices of the two groups. From these results, it appears that there is substantial support for the goal of the current study, namely to demonstrate a satisfactory level of structural invariance when Black and White groups are being compared, because support was found for the invariance of the correlation matrices. With regard to the comparison of the covariance matrices, less well-fitting results were obtained in the case of two of the model fit indices. Furthermore, the goodness-of-fit indices remained approximately constant when the correlation matrices were being compared with and without removing the effect of social desirability.

DISCUSSION

The results of this study indicate that the internal consistency reliabilities of the OPQ32n scales are acceptable for the two different groups for basic and applied research, although the mean alpha for the Black group was substantially lower than that for the White group. When compared to findings in the UK, it is evident that the Black group obtained somewhat lower alphas than the lowest and highest alphas reported in the UK, but that the White group obtained substantially higher alpha values than the sample from the general population in the UK (OPQ32 Technical Manual, 2006). The Social Desirability scale yielded an alpha value of 0.66 for the Black group and 0.66 for the White group, which is higher than the value of 0.63 reported in the UK study. Overall, the alpha coefficients for the various scales were acceptable for the total sample, because the lowest value (0.68) was obtained for Social Desirability and the highest for Rule Following and Worrying (0.88). The reliability results obtained here were similar to those of another South African study (SHL South Africa, 2002).

Clark and Watson (1995) cautioned against over-reliance on coefficient alpha to assess the extent of the internal consistency of a measuring instrument. They regarded indexes of internal consistency, such as alpha, as ambiguous because their magnitudes rely on the number of test items plus the mean intercorrelation between the items. Coefficient alpha, as an index of internal consistency, is rendered more or less useless, because the number of items is entirely irrelevant. As a solution, Clark and Watson (1995) recommended that the mean inter-item correlation per scale be used as the measure of internal consistency. They suggested that mean inter-item correlations should fall in the range 0.15–0.50, depending on the nature of the constructs. In the current study, the mean inter-item correlations for the Black group fell into the recommended range, whereas when the scale alphas were around 0.90 for the White group, the corresponding correlations approximated 0.60. Clark and Watson (1995) pointed out that correlations that are too high are indicative of measuring instruments of too narrow constructs, often at the expense of validity.

The reliability findings reported in the context of the current study are markedly higher than those reported by Meiring *et al*. (2005). A major obstacle in their study regarding bias in a personality questionnaire that was developed specifically for use in the workplace, the 15FQ+, was that the alphas for the Black language groups in particular were very weak, in some cases as low as 0.20. In their research, the magnitudes of the obtained reliability coefficients were so low that they probably affected the obtained research findings.

With regard to comparisons between the means of the Black and White groups on the OPQ scales, several statistically significant differences were obtained, but most of these differences were small in magnitude. The Black respondents described themselves as more Persuasive, Outspoken, Socially Confident, Democratic, Caring, Data Rational, Innovative, Adaptable, Forward Thinking, Rule Following, Relaxed, Tough Minded, Optimistic, Trusting and Competitive than the White respondents, whereas the White respondents described themselves as more Modest, Behavioural, Conventional, Worrying and Decisive than the Black respondents. A medium-effect size threshold was reached only in the case of Data Rational, Decisive and Social Desirability. When interpreting these results, one should bear in mind that the Black respondents were more inclined to provide socially desirable responses than the White respondents (d = 0.50).

One would have expected larger differences between Black and White South African groups on the OPQ scales than between White and minority groups in the UK due to possible cultural distance, but this was generally not the case except for the reported medium effect sizes. In the case of Social Desirability, a smaller effect size (d = 0.32) was obtained than in the current study, indicating that the minority group members were more inclined to provide socially desirable responses than members of the White group (OPQ Technical Manual, 2006). In interpreting the obtained differences between groups, readers are reminded that no assumption of full scale equivalence may be made in these studies. All too often, social science research is published without due acknowledgement of the limitations that the untested assumption of full scale equivalence pose.

The structural equation modelling indicated a highly satisfactory degree of structural invariance when the groups were compared with regard to their factor correlation matrices on the 32 scales. South African Black and White respondents therefore were comparable as far as their correlations between the 32 scales were concerned. For the present study, the score patterns obtained by the Black and White groups therefore can be considered structurally equivalent, in the sense that the OPQ32n questionnaire in this particular application of a comparison between Black and White groups was not biased in terms of yielding different correlation matrices for the two groups. Although the results obtained in the present study appear more favourable than those reported by Meiring *et al*. (2005) regarding the 15FQ+, it is important to remember that direct comparisons cannot be made unless comparable methodology and samples have been used. Somewhat less positive results regarding two of the fit indices were obtained when the covariances were compared, indicating that some of the variances between the groups differed. The latter result was expected, given the explanation by Bentler (2005) that exact equality of all Σ_g is hard to verify in large samples.

Furthermore, the analyses indicated that there was structural invariance with the effect of the Social Desirability scale partialled out. Removing the effect of social desirability did not affect the structural equivalence of the two groups substantially, because when the correlations were computed on scores with the effect of social desirability controlled, the fit indices remained largely unchanged. This may indicate that the possible systematic effect of social desirability on the scale scores is similar in the two groups, despite the fact that the groups differed with regard to their means on this variable in the present study and in others (OPQ Technical Manual, 2006). This result is also plausible, because the Social Desirability scale did not correlate substantially with the other scales. The impact of the Social Desirability scale on the research findings can thus be regarded as negligible. Similar conclusions were reached by Meiring *et al*. (2005) when they investigated whether method bias existed as a result of differences in response styles across cultural groups. They found that social desirability scores did not affect the magnitude of differences between twelve South African language groups with regard to the 15FQ+ personality questionnaire. These results also support those found by Ones and Viswesvaran (1998), who reported that social desirability functions neither as a mediator nor as a suppressor variable in personality measurement.

It is important to note that, in the present study, the so-called global test of equal correlation/covariance matrices was conducted as originally advocated by Jöreskog (1971). Byrne (2006) had indicated that this test may lead to contradictory or inconsistent results due to the fact that there is no baseline model that permits an orderly sequence of analytic steps for testing sets of parameters in a series of increasingly restrictive hypotheses. One has to bear in mind that it is not yet finally established what the preferred method for invariance testing should be, because Byrne (2006) admitted that several issues need to be resolved and backed up with sound analytic findings. The global test is regarded as an 'overly restrictive test of the data' and 'substantially more stringent than is the case for tests of invariance related to sets of parameters in the model' (Byrne, 2006, p. 175). We used the global test because it was not our goal to determine the number of underlying factors of the OPQn for each of the groups, nor whether the OPQ items reflected 32 personality factors. Construct *validation* was therefore not the goal of the study, because we assumed that the test measures 32 personality factors. The global test conducted here provided a test of the invariance of the factor correlation and covariance matrices of the OPQ32n. This test indicates whether relationships between multiple constructs (measuring a wide domain) are similar in the groups and, by implication, that the factor structures and convergent or discriminant validity for the groups will be similar.

A limitation of the study may be the relative homogeneity of the sample with regard to education, which implies that generalisability to broader South African samples cannot be assumed. Furthermore, the present study was limited to Black and White people. Future research on the OPQ32n should also endeavour to include samples of Coloured and Indian people. These groups were omitted here due to small sample sizes.

In spite of the positive findings with regard to structural equivalence and the social desirability response style, it should be borne in mind that no assumptions regarding full scale equivalence can be made on the basis of the present findings. In fact, Church (2001) and Van de Vijver and Leung (2000) concluded that personality studies across cultures present researchers with major challenges, because all three levels of equivalence will rarely, if ever, be fully met. Yet these aspects regarding the OPQn should be investigated in more depth. Studies that do not address all aspects of bias do not justify unambiguous interpretations of observed cross-cultural differences. It is recommended that future studies with the OPQn also investigate item bias, method bias and predictive bias across South African population groups. Although there appear to be insurmountable difficulties to overcome when scores obtained in different cultures have to be compared (so-called level-oriented studies), Van de Vijver and Tanzer (1997) suggested that the careful design of empirical studies, with due regard for existing literature findings, will help to pinpoint the types of bias to expect and cater for. The present study focused on aspects of structural equivalence only. This means that the OPQn passed the first hurdle in this particular context, but that further investigation is necessary to provide evidence that the questionnaire is suitable for use in personnel decisions across the various South African population groups.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the assistance provided by Tina Joubert (SHL, South Africa) and Anna Brown (SHL, UK) in conducting the SPSS and EQS analyses for this study. Explanatory comments by Anna Brown regarding the methodology were invaluable to the authors.

REFERENCES

Abrahams, F. (1997). *Problems associated with the continued use of existing psychological tests in a new South Africa: A study using the Sixteen Personality Factor Questionnaire (16PF) as an example.* Paper presented at the 3rd Annual Congress of the Psychological Association of South Africa, 10–12 September 1997, Durban, South Africa.

Abrahams, F., & Mauer, K.F. (1999a). Qualitative and statistical impacts of home language on responses to the items of the Sixteen Personality Factor Questionnaire (16PF) in South Africa. *South African Journal of Psychology*, *29*(2), 76–86.

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

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

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

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, *69*, 49-56.

Bedell, B., Van Eeden, R., & Van Staden, F. (1999). Culture as moderator variable in psychological test performance: Issues and trends in South Africa. *Journal of Industrial Psychology*, *25*(3), 1–7.

Bentler, P.M. (1985–2005). *EQS 6.1 for Windows.* Encino: Multivariate Software.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246.

Bentler, P.M. (2005). *EQS 6 structural equations program manual.* Encino: Multivariate Software.

Berg, A., Buys, M., Schaap, P., & Olckers, C. (2004). Comparability of the construct validity of Schepers' locus of control inventory for first and second language respondents. *South African Journal of Industrial Psychology*, *30*(3), 87–96.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park: Sage.

Byrne, B.M. (2006). *Structural equation modelling with EQS: Basic concepts, applications and programming.* (2nd edn.). New York: Routledge.

Carmines, E., & McIver, J. (1981). Analyzing models with unobserved variables: Analysis of covariance structures. In G. Bohrnstedt & E. Borgatta (Eds.), *Social measurement: Current issues* (pp. 65–115). Beverley Hills: Sage.

Cascio, W.F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Upper Saddle River: Pearson Prentice Hall.

Church, A.T. (2001). Personality measurement in cross-cultural perspective. *Journal of Personality*, *69*, 979–1006.

Claassen, N.C.W. (1993). *Verslag oor die funksionering van die NSAG Intermediêr G in verskillende bevolkingsgroepe* [Report on the functioning of the NSAG Intermediate G in different population groups]. Pretoria: Human Sciences Research Council.

Clark, L.A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309–319.

Clevenger, J., Pereira, G.M., Weichmann, D., Schmitt, N., & Harvey, V.S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology*, *86*(3), 410–417.

Coetzer, W.J., & Rothmann, S. (2007). A psychometric evaluation of measures of affective well-being in an insurance company. *South African Journal of Industrial Psychology*, *33*(2), 7–15.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale: Erlbaum.

Cole, N.S., & Moss, P.A. (1989). Bias in test use. In R.L. Linn (Ed.), *Educational measurement.* (3rd edn.) (pp. 201–219). New York: Macmillan.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

Dunlap, W.P., & Cornwell, J.M. (1994). Factor analysis of ipsative measures. *Multivariate Behavioral Research*, *29*(1), 115–126.

Foxcroft, C. (2004). Planning a psychological test in the multicultural South African environment. *South African Journal of Industrial Psychology*, *30*(4), 8–15.

Foxcroft, C., & Aston, S. (2006). Critically examining language bias in the South African adaptation of the WAIS-III. *South African Journal of Industrial Psychology*, *32*(4), 97–102.

Geisinger, K.F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, *6*, 304–312.

Gregory, R.J. (2007). *Psychological testing.* (5th edn.). Boston: Pearson Education Group.

Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C. (1998). *Multivariate data analysis.* (5th edn.). London: Prentice-Hall.

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

Hu, L., & Bentler, P.M. (1999). Cut-off criteria for fit indexes in covariance structural analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

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*(2), 26–33.

Johnson, C.E., Wood, R., & Blinkhorn, S.F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology*, *61*,153–162.

Jöreskog, K.G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409–426.

Kerlinger, F.N., & Lee, H.B. (2000). *Foundations of behavioral research.* (4th edn.). New York: Harcourt College.

Koch, E. (2007). Die evaluering van 'n eentalige toelatingstoets wat vir toelating tot hoër onderwys in 'n veeltalige konteks gebruik word [The evaluation of a single language entrance exam for admission to higher education used in a multilingual context]. *South African Journal of Industrial Psychology*, *33*(1), 90–101.

Mauer, K.F. (2002). *Psychological test use in South Africa*. Retrieved April 10, 2003, from http://www.pai.org.za/Psychological%20test%20use%20in%20South%20Africa.pdf

McCrae, R.R., & Costa, P.T., Jr. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*, 81–90.

Meiring, D., Van de Vijver, A.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.

Nunnally, J.C. (1978). *Psychometric theory.* (2nd edn.). New York: McGraw-Hill.

Ones, D.S., & Anderson, N. (2002). Gender and ethnic differences on personality scales in selection: Some British data. *Journal of Occupational and Organizational Psychology*, *75*, 255–276.

Ones, D.S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance*, *11*, 245–269.

OPQ32 Technical Manual. (2006). Thames Ditton: SHL Group plc.

Owen, K. (1989). *Test and item bias: The suitability of the Junior Aptitude Test as a common test battery for white, Indian and black pupils in Standard 7.* Pretoria: Human Sciences Research Council.

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

Republic of South Africa (1998). Employment Equity Act No. 55 of 1998. *Government Gazette*, *400*, 19370. Pretoria: Government printer.

Satorra, A., & Bentler, P.M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. In *American Statistical Association 1988 Proceedings of the Business and Economic Sections*, (pp. 308–313). Alexandria: American Statistical Association.

Saville, P., & Willson, E. (1991). The reliability and validity of normative and ipsative approaches in the measurement of personality. *Journal of Occupational Psychology*, *64*, 219–238.

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.

Schaap, P., Buys, M.A., & Olckers, C. (2003). The construct validity of Schepers' Locus of Control Inventory for black and white tertiary students. *South African Journal of Industrial Psychology*, *29*(1), 32–43.

SHL South Africa (2002). Reliability study, *Study No R038, 06 March 2002.*

Society for Industrial and Organisational Psychology of South Africa (2005). *Guidelines for the validation and use of assessment procedures for the workplace*. Pretoria: Society for Industrial and Organisational Psychology of South Africa.

Tabachnik, B.G., & Fidell, L.S. (2001). *Using multivariate statistics.* (4th edn.). Boston: Allyn & Bacon.

Taylor, I.A. (2000). *The construct comparability of the NEO PI-R Questionnaire for black and white employees.* Unpublished doctoral dissertation, University of the Free State, Bloemfontein, South Africa.

Taylor, T.R., & Boeyens, J. (1990). *A comparison of black and white responses to the South African Personality Questionnaire* (NIPR Report PERS-440). Pretoria: Human Sciences Research Council.

Van de Vijver, A.J.R., & Rothmann, S. (2004). Assessment in multicultural groups: The South African case. *South African Journal of Industrial Psychology*, *30*(4), 1–7.

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

Van de Vijver, F., & Leung, K. (2000). Personality in cultural context: Methodological issues. *Journal of Personality*, *69*, 1006–1030.

Van de Vijver, F., & Poortinga, Y.H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, *13*, 21–29.

Van de Vijver, F., & Tanzer, N.K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *47*(4), 263–279.

Van Eeden, R., & Prinsloo, C.H. (1997). Using the South African version of the 16PF in a multicultural context. *South African Journal of Psychology*, *27*, 151–159.

Van Eeden, R., & Van Tonder, M. (1995). *The validity of the Senior South African Individual Scale – Revised (SSAIS-R) for children whose mother tongue is an African language: Model C schools.* Pretoria: Human Sciences Research Council.

Van Eeden, R., & Visser, D. (1992). The validity of the Senior South African Individual Scale – Revised (SSAIS-R) for different population groups. *South African Journal of Psychology*, *22*, 163–171.

Van Eeden, R., Taylor, T.R., & Du Toit, R. (1996). *Adaptation and standardization of the Sixteen Personality Factor Questionnaire Fifth Edition (16PF5) in South Africa: A feasibility study.* Pretoria: Human Sciences Research Council.

Visser, D. (2002). *The relation between population group and social desirability response style among South African telecommunications employees*. Paper presented at the XVIth International Association for Cross-Cultural Psychology Congress, 15–19 July 2002, Yogyakarta, Indonesia.

Visser, D., & Du Toit, J.M. (2004). Using the Occupational Personality Questionnaire (OPQ) for measuring broad traits. *South African Journal of Industrial Psychology*, *30*(4), 65–77.

Vorster, M., Olckers, C., Buys, M.A., & Schaap, P. (2005). The construct equivalence of the job diagnostic survey for diverse South African cultural groups. *South African Journal of Industrial Psychology*, *31*(1), 31–37.

Wallis, T. (2004). Psychological tests do not measure what they claim to measure: A re-evaluation of the concept of construct validity. *South African Journal of Psychology*, *34*(1), 101–121.