SA Journal of Industrial Psychology, 2005, 31 (1), 1-8
SA Tydskrif vir Bedryfsielkunde, 2005, 31 (1), 1-8

CONSTRUCT, ITEM, AND METHOD BIAS OF COGNITIVE AND PERSONALITY TESTS IN SOUTH AFRICA

D MEIRING
South African Police Services
Pretoria
AJR VAN DE VIJVER
Tilburg University, The Netherlands
Potchefstroom Campus, North-West University
S ROTHMANN
Potchefstroom Campus
North-West University
MR BARRICK
Tippie School of Business
University of Iowa, U.S.A.

ABSTRACT

Bias was studied for two cognitive tests and a personality test at three levels: the construct underlying the test ("construct bias"), method-related aspects such as response sets ("method bias"), and the items ("item bias"). The sample consisted of 13 681 participants who had applied for entry-level jobs in the South African Police Service. The cognitive instruments produced very good construct equivalence and low item bias. However, various scales of the personality questionnaire revealed construct bias in various ethnic groups. The item bias in the personality scales was low. Method bias did not have any impact on the (small) size of the cross-cultural differences in the personality scales. In addition, several personality scales revealed low internal consistencies, notably in the black groups.

OPSOMMING

Sydigheid is bestudeer vir twee kognitiewe toetse en 'n persoonlikheidstoets op drie vlakke: die konstruk onderliggend aan die toets ("konstruksydigheid"), metode-verwante aspekte soos responspatrone ("metodesydigheid"), en die items ("itemsydigheid"). Die steekproef het bestaan uit 13 681 deelnemers wat aansoek gedoen het om intreevlak-poste in die Suid-Afrikaanse Polisiediens. Die kognitiewe instrumente het baie goeie konstrukekwivalensie en lae itemsydigheid getoon. Verskeie skale van die persoonlikheidsvraelys het egter konstruksydigheid in verskeie etniese groepe getoon. Die itemsydigheid in die persoonlikheidskale was laag. Metodesydigheid het nie enige uitwerking op die (klein) omvang van die kruiskulturele verskille in die persoonlikheidskale gehad nie. Verder het verskeie persoonlikheidskale lae interne konsekwentheid getoon – veral in die swart groepe.

Psychological testing in South Africa cannot be investigated in isolation without taking the country's political, economic, and social history into account (Claassen, 1997). Psychometric testing in South Africa has mainly followed international trends and at the beginning of the 1900s tests were imported from abroad and applied in all sectors of the community (Foxcroft, 1997). Cross-cultural issues emerged in the 1920s, and in the 1940s and 1950s psychological testing focused on the educability and trainability of black South Africans. In the 1980s certain aspects of fairness, bias, and discriminatory practices received more attention in line with international developments. Separate psychological tests were initially developed for the Afrikaans and English-speaking groups (Claassen, 1997). At a later stage bilingual tests were constructed for English and Afrikaans speakers and separate tests were constructed for speakers of African languages.

Since the first democratic elections, held in 1994, the country has had a new constitution and stronger demands for the cultural appropriateness of psychological tests culminated in the promulgation of the Employment Equity Act 55 of 1998, Section 8 (Government Gazette, 1998, p. 9), which stipulates the following: "Psychological testing and other similar assessments are prohibited unless the test or assessment being used (a) has been scientifically shown to be valid and reliable, (b) can be applied fairly to all employees; and (c) is not biased against any employee or group."

The onus of proof has shifted to psychologists using these instruments, who now have to indicate that they adhere to the regulations of the Employment Equity Act. Given the transformation of the South African society, the integration of schools, universities, the work place, and society in general since 1994, there is an urgent need for measuring instruments that meet the Employment Equity Act requirements and can be used for all the cultural and language groups in South Africa.

The current study examines the extent to which the most important tests in the assessment procedure to recruit new police officials for the South African Police Services (SAPS) – two cognitive tests (a Reading and Comprehension Test and a Spelling Test) and a personality questionnaire (15FQ+) meet the criteria imposed by the Employment Equity Act by examining bias in the instruments employed.

Bias and equivalence

Bias and equivalence are pivotal concepts in the application of psychological tests in a multicultural society such as South Africa. According to Van de Vijver and Tanzer (1997), bias occurs when score differences in the indicators of a particular construct do not correspond with differences in the underlying trait or ability. Equivalence on the other hand refers to score comparability, namely the measurement level at which scores obtained for different cultures can be compared. Consequently, bias refers to the influence of nuisance factors (unwanted but systematic sources of variation) in cross-cultural score comparisons whereas equivalence is the consequence of the nuisance factors concerning the comparability of scores across cultures. Van de Vijver and Tanzer (1997) note that bias has to do with the characteristics of an instrument in a (specific) cross-cultural comparison rather than with its intrinsic properties. The question as to

Requests for copies should be addressed to: S Rothmann, Department of Industrial Psychology, North-West University, (Potchefstroom Campus), Private Bag X6001, Potchefstroom, 2520

whether an instrument is biased cannot be answered in general terms, but can be addressed when an instrument is biased in a specific comparison.

Van de Vijver and Leung (1997a, 1997b) propose a taxonomy of bias consisting of three types, namely construct bias, method bias and item bias. Construct bias occurs when the construct measured is not identical across cultures or when behaviours that characterise the construct are not identical across cultures. This type of bias can stem from several sources; for example the definition of a construct may show an incomplete overlap across cultures. Method bias refers to problems caused by the manner in which a study is conducted (method-related issues). Three types of method bias can be distinguished (Van de Vijver, 2002). First, incomparability of samples on factors other than the target variables can lead to method bias (sample bias). Second, method bias also refers to problems arising from instrument characteristics (instrument bias). Third, method bias arises from administration problems (administration bias). Item bias (also referred to as differential item functioning) refers to the situation in which the (psychological) meaning of one or more items is not identical across cultures and relates to anomalies at the item level, such as poor translation or inapplicability of an item to a specific culture.

Van de Vijver and Tanzer (1997) consider bias as an indication of a source of systematic cross-cultural differences that need to be studied. Bias analysis can offer important clues concerning the causes of cross-cultural differences and can thus be regarded as a phenomenon that requires further explanation.

According to Van de Vijver and Leung (1997a, b), equivalence refers to the implications of bias with regard to the comparability of constructs and test scores. Van de Vijver and Tanzer (1997) treat equivalence from a measurement perspective and make a hierarchical distinction between three types of equivalence. The first level is called construct equivalence. This means that the same construct is measured across all cultural groups studied, irrespective of whether or not the measurement of the construct is based on identical instruments across cultures. It implies the universal validity of the underlying psychological construct. The second level of equivalence is called metric or measurement unit equivalence and is obtained when two metric measures have the same measurement unit but different origins. In the case of measurement unit equivalence no direct score comparisons can be made across cultural groups unless the size of the offset (i.e., the difference in scale origin) is known. The highest level of equivalence is scalar equivalence or full-scale equivalence and this is obtained if two metric measures have the same measurement unit and the same origin.

Bias and equivalence in cognitive and personality tests in South Africa

Cognitive tests. Cross-cultural comparison of cognitive test scores is not new in South Africa (Irvine, 1969). Biesheuvel's (1943, 1954) early work in South Africa focuses on the empirical investigation of potential bias problems associated with cross-cultural assessment. Biesheuvel emphasised the importance of home environment, schooling, nutrition, and other factors in cognitive test performance in a multicultural society. Schepers (1974) reported that urban subjects, when compared with rural examinees, have a slightly greater differentiated intellect, with education playing the biggest role in the differentiation process. Freeman (1984) reported that the cognitive skills needed to deal with the Raven Progressive Matrices are better developed in an urbanised population than in a rural one. Verster and Prinsloo (1988) compared the results of IQ points of different generations and found decreasing differences between the English speaking and Afrikaans speaking adults. Claassen (1997) reported that between 1954 and 1984 the mean difference between English-speakers and Afrikaans-speakers was reduced from ten IQ points to five IQ points. Socioeconomic and educational circumstances change from one generation to another and have an impact on cognitive test scores. This phenomenon contributes to method bias.

In South Africa few studies focused on the construct equivalence of cognitive measures across cultures. Most studies that were carried out concerned comparisons between English speakers and Afrikaans speakers. A high degree of structural equivalence was reported in these studies (Cudeck & Claassen, 1983; Verster, 1974; Vorster 1978). Between 1960 and 1984 it was not necessary for psychology to look at the issue of construct equivalence since tests were developed independently for each of the race groups and no cross-cultural comparisons were made (Claassen, 1997; Owen, 1992). In the 1980s there was growing interest in comparing cultural groups with regard to existing cognitive tests. Claassen (1993) applied the New South African Group Test (NSGT) to Blacks, Coloureds, Indians, and Whites in order to assess the cross-cultural suitability of the test. All the respondents wrote the test in English. The verbal part of the test was problematic for the Black group since English was not their mother tongue. Large mean differences were reported for the cultural groups and the structural equivalence was found to be poor. Owen (1986) investigated structural equivalence and item bias by applying three cognitive tests (Senior Aptitude Test, Mechanical Insight Test and Scholastic Proficiency Battery) to Black, Coloured, Indian, and White students. He reported structural equivalence across these cultural groups and item bias analyses supported the suitability of the measures for all groups. Owen (1989) also examined the structural equivalence and item bias of the Junior Aptitude Test for White, Indian and Black pupils in Standard 7. For the Black pupils the structural equivalence was problematic. Many items in the case of the Indian and Black groups were biased. Results pointed to the strong influence of education and understanding of the English language on structural equivalence and of item bias on cognitive tests.

Personality questionnaires. Cross-cultural personality research has focused extensively on the universality of the five-factor model (FFM) (Cheung et al., 2001; McCrae & Allik, 2002; Paunonen, Zeidner, Enggvik, Oosterveld, & Maliphant 2000; Roland, Parker, & Strumf, 1998) and Eysenck's three-factor model (Barrett, Petrides, Eysenck, & Eysenck, 1998). In South Africa a few studies have been conducted, investigating the FFM across cultural groups. Heuchert, Parker, Strumf, and Myburg (2000) applied the NEO-Personality Inventory-Revised (NEO-PI-R) to college students. The authors found a clear five-factor solution for both Black and White students. An unpublished thesis (Horn, 2000) examined a Xhosa translation of the NEO-PI-R. Horn reported that translation was difficult and that various items could not be translated into Xhosa because of its restricted vocabulary. Taylor (2000) carried out a construct comparability study of the NEO-PI-R for Black and White employees in a work setting. The NEO-PI-R did not work as well for Blacks as it did for Whites. In particular the openness factor could not be extracted in the Black sample. Other studies in South Africa made use of the South African Personality Questionnaire (SAPQ) and the 16 PF (South African 1992 version). There was little support for construct equivalence across the different cultural groups in South Africa. Individuals whose first language was not English experienced problems with the questionnaire, especially because some of the items were difficult to understand. Researchers concluded that these tests were not suitable for use in a multicultural society like South Africa (Abrahams, 1996, 2002; Abrahams & Mauer, 1999a, 1999b; Meiring, 2000; Spence, 1982; Tact 1999; Taylor & Boeyens, 1991).

In summary, cognitive and personality cross-cultural studies had seldom been carried out in South Africa before the 1980s. In line with international trends there has been increasing interest in the topic during the last few decades. Structural equivalence and item bias of cognitive tests were studied while in the case of personality tests the focus was mainly on structural equivalence. These studies mainly adopted the designs and statistical procedures found in the Anglo-Saxon literature (Berry et al., 2000). Studies in South Africa reported race, education, language, and understanding of English as the main reasons impacting on construct and item comparability of cognitive and personality tests. There is a need to continue to research the issues of bias in a contemporary South Africa.

Research aims

The first aim of this study was to examine bias at the level of constructs (structural equivalence) and items (item bias) in two cognitive tests and a personality test that were administered to select entry-level police officials for the South African Police Service (SAPS). In addition, method bias was studied by examining the influence of cognition and social desirability on the 15FQ+.

METHOD

Participants

The sample consisted of 13,681 participants throughout South Africa who applied for entry-level police jobs in the SAPS. Applicants came from all nine provinces. The sample consisted of Blacks (*n* = 11,626), Whites (*n* = 570), Indians (*n* = 662) and Coloureds (*n* = 812). Ninety percent (*n* = 11,317) were male and ten percent (*n* = 2,353) were female. The Black group consisted of the following nine ethnicities: Ndebele (*n* = 259), Sepedi (*n* = 1,777), SeSotho (*n* = 1,285), Setswana (*n* = 2,009), Swati (*n* = 294), Tsonga (*n* = 922), Venda (*n* = 978), Xhosa (*n* = 1,725), and Zulu (*n* = 2,404). The mean age of the sample group was 25 years (*SD* = 2.8). The entry-level requirement for the police is Grade 12, 69% of the sample had a Grade 12 qualification, 13% had a degree or diploma, and 18% had a post-graduate qualification.

Instruments

The test battery consisted of a cognitive section, which included an English reading and comprehension test, an English spelling test and the 15FQ+ Questionnaire.

The reading and comprehension test consisted of four paragraphs that were selected from the basic training modules (Module 1: the Bill of Rights on Police Power, Community Policing; Module 2: Non-Verbal Communication; Module 5: Mental Disorders). Five questions were asked in respect of each paragraph. The test requires the applicant to read the paragraphs and comprehend the material in order to answer the questions. The test consists of 20 items and each item has four response alternatives. A time limit of 20 minutes was allowed for the completion of the test. The spelling test was also developed for the SAPS. Training instructors at the training college were asked to generate a pool of police-relevant words (such as *rape* and *homicide*) which students find difficult to spell when they start their basic training. A pool of words was generated and a spelling test consisting of 40 items was developed. An item consisted of four different spellings of a single word. Applicants had to select the correctly spelled word. A time limit of 12 minutes was given for the completion of the test. The reliability of reading and comprehension and spelling test (internal consistency; Cronbach's alpha) for the different language groups is reported in Table 1. The mean alpha coefficients of the two tests are 0,84 (spelling test) and 0,64 (reading and comprehension), respectively. All these values are acceptable (α > 0,60, Clark & Watson, 1995), and thus indicate an acceptable internal consistency.

TABLE 1
VALUES OF CRONBACH'S ALPHA ACROSS CULTURAL GROUPS PER TEST/SCALE

	Cultural Group											
Test/Scale	**Xhosa**	**Zulu**	**Ndebele**	**Sepedi**	**Sesotho**	**Setswana**	**Swati**	**Tsonga**	**Venda**	**Indian**	**Coloured**	**White**
Cognitive												
Reading and Comprehension Test	0,623	0,634	0,647	0,601	0,564	0,633	0,607	0,618	0,586	0,697	0,685	0,764
Spelling Test	0,841	0,840	0,827	0,854	0,838	0,842	0,834	0,854	0,823	0,837	0,816	0,849
Personality Scales												
Cool Reserved – Outgoing	0,429	0,445	0,396	0,510	0,510	0,457	0,510	0,527	0,474	0,643	0,559	0,629
Intellectance	0,551	0,529	0,452	0,551	0,576	0,518	0,465	0,583	0,501	0,670	0,639	0,615
Affected by Feelings – Emotionally Stable	0,590	0,596	0,552	0,581	0,638	0,652	0,603	0,567	0,627	0,750	0,730	0,753
Accommodating – Dominant	0,286	0,383	0,364	0,326	0,356	0,377	0,349	0,328	0,230	0,655	0,587	0,680
Sober Serious – Enthusiastic	0,546	0,603	0,477	0,569	0,611	0,606	0,621	0,568	0,500	0,688	0,700	0,758
Expedient – Conscientious	0,472	0,501	0,468	0,485	0,465	0,460	0,428	0,450	0,537	0,683	0,537	0,624
Retiring – Socially Bold	0,638	0,629	0,602	0,599	0,629	0,637	0,609	0,553	0,518	0,818	0,746	0,826
Tough Minded – Tender Minded	0,384	0,345	0,406	0,354	0,403	0,448	0,388	0,348	0,279	0,712	0,628	0,755
Trusting – Suspicious	0,353	0,364	0,354	0,351	0,392	0,385	0,415	0,364	0,356	0,682	0,607	0,700
Practical – Abstract	0,088	0,138	0,245	0,091	0,154	0,114	0,182	0,0006	0,118	0,447	0,388	0,461
Forthright – Discreet	0,421	0,453	0,530	0,502	0,480	0,479	0,420	0,491	0,421	0,667	0,564	0,698
Self-assured – Apprehensive	0,355	0,404	0,460	0,434	0,453	0,444	0,460	0,426	0,420	0,267	0,378	0,283
Conventional – Radical	0,231	0,157	0,268	0,199	0,163	0,151	0,003	0,160	0,005	0,478	0,346	0,532
Group – Orientated – Self-Sufficient	0,507	0,560	0,544	0,524	0,549	0,552	0,519	0,496	0,421	0,702	0,665	0,760
Undisciplined – Self-Disciplined	0,375	0,400	0,401	0,436	0,362	0,315	0,392	0,391	0,383	0,382	0,384	0,405
Relaxed – Tense Driven	0,429	0,455	0,396	0,506	0,510	0,457	0,510	0,527	0,474	0,643	0,559	0,629

The 15FQ+ is a normative, trichotomous response, personality test that has been developed by Psytech International as an update of the original 15FQ (Tyler, 2002). Both versions of the 15FQ were designed for use in industrial and organizational settings. The original version of this assessment was first published in 1991 as an alternative to the 16PF series of tests. The original 15FQ was designed to assess 15 of the 16 personality dimensions that were first identified by Cattell and his colleagues in 1946. The 15FQ+ is a complete revision of the original 15FQ, with the authors developing and fielding a completely new item set for the 15FQ+. The authors' stated aim was to produce a relatively short, yet robust measure of Cattell's primary personality factors (Tyler, 2002). It has been known for some time that reasoning ability (or intelligence) cannot be reliably measured by reasoning items included in untimed personality tests, as is the case with Cattell's Factor B. For this reason Factor B was excluded from the 15FQ. However, in the case of the 15FQ+, the authors decided to deal with this problem by redefining Factor B as a "metacognitive personality variable" called intellectance. Validity and reliability have been determined for the 15FQ+ (Tyler, 2002). For this study the reliabilities (Cronbach's alpha) for the different language groups are reported in Table 1. The internal consistencies for some of the factors were very low, notably in the Black language groups. There is a serious problem with the internal consistencies of the following factors: Practical – Abstract (mean alpha = 0,20) and Conventional – Radical (0,22) across all groups. These low values seriously challenge the suitability of the 15FQ+ in this multicultural setting.

Procedure

Applicants were tested in groups of 100 during April 2000. A standardised procedure was followed by previously trained personnel of the Psychological Services of the SAPS in order to apply the test battery. The test session lasted for three hours and also contained a break of 15 minutes. Computer-readable answer sheets were utilised for all the tests.

Statistical Analysis

Construct bias and item bias were addressed in two series of analyses for both the cognitive and personality tests. The first involved scale-level analyses and examined the similarity of the factors underlying the cognitive and personality tests, whereas the second addressed bias at item level of the instruments. Method bias in the personality scales was examined by looking at the influence of cognition and social desirability on the personality scores.

Scale-level analysis (construct bias). A two-step procedure was used to examine construct bias which is based on exploratory factor analysis. In the first step the covariance matrices of all the cultural groups were combined (weighted by sample size) in order to create a single, pooled data matrix (cf. Muthén, 1991, 1994). Factors derived from this pooled covariance matrix define the global solution, with which the factors obtained in the separate cultural groups were compared (after target rotation to the pooled solution). The agreement was evaluated by means of a factor congruence coefficient, Tucker's phi (Chan, Ho, Leung, Cha & Yung, 1999; Van de Vijver & Leung, 1997a, 1997b). Values above 0,90 are taken to point to essential agreement and values above 0,95 to very high agreement. High agreement implies that the factor loadings of the lower and higher level are equal up to a multiplying constant. (The latter is needed to accommodate possible differences in the eigenvalues of factors for the language groups).

Item level analysis (item bias analysis). Item bias analysis was undertaken by using two different procedures. Logistic regression was used for the cognitive instruments (yielding dichotomous scores) and analysis of variance (ANOVA) was used for the personality test (yielding interval-level scores). Both kinds of analyses are based on the same conceptualization of item bias. The assumption is that an item is unbiased if persons from different cultures with an equal standing on the theoretical construct underlying the instrument have the same expected score on the item (Van de Vijver & Leung, 1997a, 1997b).

Logistic regression is a general procedure of analysing differential item functioning (DIF) as it can detect both uniform and non-uniform bias (Mellenbergh, 1982; Van de Vijver & Leung, 1997a, 1997b) in dichotomous items and thus provide a model-based approach for studying DIF (Rogers & Swaminathan, 1990, 1993). The total test score (a proxy for ability level) and culture are the independent variables, while the item score is the dependent variable. The presence of a significant main effect of score level is usually taken as an indication of uniform bias. An item is taken to show non-uniform bias if the interaction between level and culture is significant. In the present study the sample size was large so that conventional tests of significance could not be used. The procedure that was used for the cognitive tests computed the effect size for the items, where the difference between the Nagelkerke R^2 of the first step (in which score level was the sole predictor) and second step (in which culture, dummy coded was added as a predictor) provides an estimate of the effect size of culture (uniform bias). In the third step the interaction of culture and score level is added; the difference between the second and the third estimates the impact of the interaction (non-uniform bias).

In the analysis of variance of the personality items the item score was the dependent variable, while culture and score levels were the independent variables. Analogous to the previous analysis, a significant main effect of the culture group was taken to point to uniform bias, and a significant interaction of score level and culture interaction pointed to non-uniform bias.

Finally, the influence of the presence of biased items on the size of cross-cultural differences was examined. This was done by comparing the cross-cultural differences in the original 15FQ+ questionnaire with those in the 15FQ+ questionnaire from which presumably biased items had been removed.

Method bias analysis. Method bias was studied in respect of the personality questionnaire. From the literature it could be concluded that knowledge of the English language could be an important moderator of responses to the 15FQ+. Similarly, differences in response styles across cultural groups could also be expected to exert some influence. In order to examine their impact, a multivariate analysis of covariance was carried out. Cultural group (12 levels) was the independent variable; the dependent variables were the scale scores of the 15FQ+ while cognitive ability (as a proxy for English language proficiency, which was the testing language) and social desirability were the covariates.

RESULTS

Scale-Level Structural Equivalence

Cognitive tests. Based on a scree test, both cognitive tests showed a unifactorial solution in the pooled data. Table 2 shows the agreement of the factor derived from the pooled data with the factor in the 12 language groups for both cognitive tests. Values of Tucker's phi higher than 0,90 were found in the two tests for all the language groups. This provided a strong indication of the structural equivalence of the cognitive factors underlying the performance of all the different groups distinguished.

Table 2
Agreement of the reading and comprehension test in the pooled solution with the black group divided into nine language groups and the three other race groups (Tucker's PHI)

	Test	
	Reading and Comprehension	Spelling
Xhosa	0,992	0,998
Zulu	0,975	0,990
Ndebele	0,957	0,907
Sepedi	0,990	0,995
Sesotho	0,990	0,994
Setswana	0,995	0,989
Swati	0,975	0,975
Tsonga	0,991	0,991
Venda	0,984	0,984
Indian	0,974	0,965
Coloured	0,992	0,975
White	0,966	0,975

Personality. Scree tests of the factor analyses of the separate scales suggested the extraction of a single factor in each analysis. The agreement of the factors of the 15FQ+ in the pooled solution with factors in the 12 language groups is indicated in Table 3. Various entries in the table showed values well below the threshold level of 0,90. More specifically, a column comparison revealed that for four of the groups there were problems with the structural equivalence of the constructs (Ndebele 50%, Whites 44%, Indians 31%, and Coloureds 25% of the factors). A row comparison showed that in particular two scales, Conventional – Radical and Relaxed – Tense Driven, did not show structural equivalence across six of the groups. Only three scales showed equivalence across all of the language groups: Accommodating – Dominant, Retired – Socially Bold, Group Orientated – Self-Sufficient.

Item-Level Analyses

Cognitive tests. It is clear from Table 4 and 5 that when bias is evaluated in terms of significance, many items revealed significant bias (reading and comprehension 50%, spelling test 68%). Cohen's (1988) criteria according to which the lower threshold for medium-size effects is 0,06 was applied to further examine the size of the item bias (this size was chosen as it can be considered to be significantly large to be practically important). It was found that for the reading and comprehension test only one item out of 20 showed non-uniform bias and for the spelling test item one item out of 40 items showed uniform bias. It can be concluded that many items show statistical bias but the bias effect is so slight as to be negligible from a practical perspective.

Personality. In analyses of variance of the item scores of the 15FQ+ we found that many items showed a significant main effect of culture (uniform bias) or interaction of culture and score level (non-uniform bias). Out of the 200 items, 72 turned out to be biased (36%), which is a large proportion. However, only one item showed a medium effect size. It can be concluded that item bias is not a major disturbance in the 15FQ+ in these language groups.

Influence of bias on size of cross-cultural differences. In order to inspect the impact of item bias on cross-cultural differences in the personality scales, the size of these differences was computed before and after the elimination of biased items. An item was taken to be biased if it had an eta square value of at least 0,02 for the uniform or non-uniform bias component. This low value was used because of the overall low level of the effect sizes. One-way analyses of variance were carried out with language group as independent variable and scale scores (sum scores on the items pertaining to the scale) as dependent variables. In a second step the procedure was repeated, but now all biased items were excluded from the computation of scale

Table 3
Agreement of the 16 factors in the pooled solution with the black group divided into nine sub-language groups and the three other race groups

Factor	Xhosa	Zulu	Ndebele	Sepedi	Sesotho	Setswana	Swati	Tsonga	Venda	Indian	Coloured	White
Reserved – Outgoing	0,989	0,992	0,882	0,998	0,995	0,993	0,980	0,996	0,992	0,986	0,989	0,974
Intellectance	0,997	0,998	0,942	0,989	0,997	0,991	0,959	0,995	0,992	0,952	0,987	0,891
Affected by Feelings – Emotionally Stable	0,997	0,998	0,951	0,991	0,995	0,998	0,994	0,975	0,988	0,980	0,990	0,971
Accommodating – Dominant	0,969	0,985	0,838	0,948	0,963	0,992	0,924	0,983	0,736	0,972	0,962	0,908
Sober Serious – Enthusiastic	0,989	0,989	0,948	0,992	0,994	0,995	0,957	0,983	0,979	0,950	0,968	0,952
Expedient – Conscientious	0,983	0,993	0,859	0,987	0,976	0,991	0,915	0,970	0,988	0,980	0,974	0,956
Retiring – Socially Bold	0,995	0,996	0,927	0,991	0,998	0,994	0,976	0,984	0,984	0,995	0,991	0,990
Tough Minded - Tender Minded	0,983	0,958	0,956	0,948	0,976	0,994	0,937	0,907	0,947	0,814	0,851	0,780
Trusting – Suspicious	0,985	0,993	0,895	0,989	0,992	0,989	0,957	0,971	0,988	0,819	0,737	0,582
Practical – Abstract	0,995	0,997	0,943	0,994	0,994	0,991	0,944	0,992	0,981	0,945	0,910	0,806
Forthright – Discreet	0,966	0,980	0,862	0,982	0,993	0,988	0,948	0,991	0,959	0,962	0,937	0,953
Self-assured – Apprehensive	0,993	0,982	0,975	0,984	0,987	0,987	0,936	0,987	0,976	0,893	0,961	0,877
Conventional - Radical	0,853	0,988	0,705	0,913	0,966	0,970	0,877	0,940	0,962	0,352	0,441	0,400
Group – Orientated – Self-Sufficient	0,988	0,996	0,965	0,989	0,990	0,993	0,979	0,959	0,986	0,980	0,993	0,971
Undisciplined – Self-Disciplined	0,985	0,994	0,615	0,986	0,987	0,978	0,940	0,980	0,954	0,935	0,939	0,928
Relaxed – Tense Driven	0,901	0,969	0,761	0,938	0,851	0,930	0,916	0,929	0,895	0,821	0,825	0,847

scores. The extent of the cross-cultural differences was evaluated as the effect size (eta square) of the culture component. The mean effect size was 0,027 before the removal of biased items and 0,028 after bias removal. It could be concluded that the correction for biased items did not affect the size of the cross-cultural differences observed.

Table 4
Items with bias of effect size and significance for the reading and comprehension for the different language groups

Item	Uniform bias	Non-uniform bias
1	0,010	0,003*
2	0,007*	0,001
3	0,004*	0,002*
4	0,001	0,002
5	0,004*	0,002
6	0,007*	0,491*
7	0,003	0,001
8	0,012	0,002*
9	0,003	0,001
10	0,002*	0,001
11	0,004*	0,001*
12	0,006	0,001
13	0,005*	0,003*
14	0,003*	0,001
15	0,002	0,002
16	0,002*	0,003*
17	0,005	0,001
18	0,004*	0,002*
19	0,004*	0,005*
20	0,002	0,001

*p < 0,05 (item shows significant (non-)uniform bias if followed by an asterisk)

Table 5
Items with bias of effect size and significance of the spelling test for the different language groups

Item	Uniform bias	Non-uniform bias
1	0,031*	0,002
2	0,012*	0,002*
3	0,005*	0,003*
4	0,012	0,002
5	0,009*	0,002*
6	0,030*	0,002*
7	0,007*	0,005*
8	0,007*	0,001
9	0,003*	0,002*
10	0,006*	0,004*
11	0,063*	0,006*
12	0,022	0,004*
13	0,007	0,002
14	0,048*	0,002*
15	0,006*	0,003*
16	0,002*	0,003*
17	0,013*	0,001*
18	0,022	0,002
19	0,008*	0,002*
20	0,010	0,001
21	0,012*	0,005*
22	0,005*	0,008*
23	0,002	0,001
24	0,003	0,002
25	0,015*	0,003*
26	0,019*	0,011*
27	0,038*	0,003*
28	0,007	0,001
29	0,004	0,003*
30	0,005	0,002
31	0,012*	0,002*
32	0,006*	0,002
33	0,008	0,002
34	0,002	0,002*
35	0,006*	0,005*
36	0,058*	0,001*
37	0,005*	0,003*
38	0,005*	0,003*
39	0,015	0,002
40	0,026*	0,001*

*p < 0,05 (item shows significant (non-)uniform bias if followed by an asterisk)

Method Bias in the Personality Questionnaire

In order to evaluate the impact of method bias the effects of cognitive/language ability and social desirability were scrutinized in an analysis of covariance. The size of the cross-cultural differences was computed before and after correction for the covariates (ability and social desirability). The main effect of the cross-cultural difference was 0,026 before correction and 0,025 after correction for covariates. Clearly the results of covariate analysis revealed that cognitive ability and social desirability scores did not have any impact on the size of the cross-cultural differences of the personality questionnaire.

DISCUSSION

This study was the first South African study in which different types of bias were studied: bias at the level of constructs, items, and the method of administration. The sample consisted of 13,681 participants throughout South Africa who had applied for entry-level police jobs in the SAPS. The sample was split into 12 different language groups. A police-specific cognitive test containing subtests of reading/comprehension and spelling test, and a personality questionnaire, the 15FQ+, were administered in this study.

Both cognitive measures showed low levels of construct bias; both revealed factorial invariance in all the language groups. Item bias analyses showed several items revealing significant bias. Instead of the significance of item bias indicators, their effect size was used as the criterion to evaluate the presence of item bias (this was done because of the large sample size). If the presence of a medium or large effect size for the indicators of uniform or non-uniform bias is taken as the criterion of item bias, almost no items showed significant bias. It seems fair to conclude that the extent of item bias is not very consequential in the cognitive measures.

The examination of the construct bias of the personality measures showed less favourable results. Structural equivalence was particularly problematic for the two factor scales (Conventional – Radical, Relaxed – Tense Driven) in four language groups (Whites, Coloureds, Indians, and Ndebele). The item bias analyses did not point to major problems at item level in any personality scale. Not surprisingly, the removal of the biased items did not affect the size of the cross-cultural differences observed. An analysis of the influence of cognitive ability (as a proxy for English language proficiency) and social desirability (as a measure of response style) revealed that the extent of the cross-cultural differences between the language groups was not influenced by these factors, thereby suggesting that the influence of these sources of method bias could be safely ignored in the current data.

The Anglo-Saxon literature, often reporting studies done in the U.S.A., provides support for the structural equivalence of most cognitive tests (Berry et al., 2002). However, for personality questionnaires the equivalence picture is not so clear (Ellis, 1995). In this study high levels of structural equivalence were found for the cognitive tests but in the case of personality test structural equivalence across the different language groups was problematic for the 15FQ+. The current results are fairly consistent with the mainstream literature. The findings with respect to item bias in the cognitive tests are also in line with the mainstream literature (Berk, 1982; Holland & Wainer, 1993): many items were found to be biased, but the size of the bias is small as is its impact on the size of intergroup differences. Similarly, the personality questionnaires showed many biased items, but their size was small and their impact on observed scores obtained in the various language groups very limited. One of the reasons for the small size of the bias may be the educational entry-level requirement, which apparently reduced the educational heterogeneity of the sample considerably. As a consequence, the results may not be generalizable to a broader, more unselected sample of the South African population. Even though the bias was small, the current findings underscore Church's (2001) conclusion that a major challenge for cross-cultural personality studies is that equivalence of constructs and measures will rarely, if ever, be fully met.

A serious problem concerns the low internal consistencies (more in the Black than the other groups). The reliability values of various personality scales are so low that they cannot be adequately used for individual assessment and selection purposes.

The nature of the the construct bias of some of the personality scales was further explored, using an expert group, consisting of Black SAPS psychologists and two African language experts. They were asked to identify aspects of the personality measures that might be a threat to the structural equivalence. Several aspects of the questionnaire were mentioned, such as the level of the words being used and the understanding of the context and interrelationship of words could be problematic, especially for Black groups (e.g., analytical, intellectually, conventional, gullible, genuinely, temperamental, smashing). The use of double meanings in items could cause confusion. The use of idiomatical expression raised concerns (e.g., "both feet firmly on the ground", "head in the clouds"). Qualifying words such as "rarely", "generally", "less", and "on occasion" could also be problematic. Finally, it was pointed out that some of the constructs could be more culture specific. Looking at the history of South Africa for example the construct of Conventional – Radical will have a stronger political connotation for the Black respondents than for other groups. Relaxed – Tense, African respondents can be seen as more relaxed people than others.

Prinsloo and Ebersohn (2002) argue that different response rates to personality items could reflect real differences in underlying traits. In the case of personality traits, which often comprise of highly socialised constructs, it is reasonable to expect that various additional sources contribute to intergroup differences. What role do education and the understanding of English play in the construct bias of the 15FQ+? Abrahams and Mauer (1999b) qualitatively examined the impact of home language on the responses to the items of the 16PF. They concluded that the understanding of items and concepts in English was problematic, especially for Black groups. Prinsloo and Ebersohn (2002) proposed that by testing respondents' English proficiency can help to assess its impact on performance in personality measurement.

Does the present study answer the question of whether the test battery being used by the SAPS to select entry-level applicants can stand the scrutiny of the Employment Equity Act 55 of 1998 and its subsections (Government Gazette, 1998)? The cognitive tests did not show much bias, whereas some personality tests were problematic. Moreover, various personality scales showed unacceptably low internal consistencies. Consequently, the results of the cognitive tests are encouraging, whereas an uninformed application of the personality scales could be problematic. In addition to problems with the structural equivalence, there is the additional problem of low internal consistencies – one more issue that challenges the use of the scales in selection. If personality constructs can be identified that are important to a police official, a selective strategy can be followed and factors that did not show structural equivalence and factors with unacceptably low internal consistencies can be avoided.

The current study did not address all aspects of test usage. More specifically, the predictive validity and predictive bias of the tests were not considered. Even an unbiased instrument may not work equally well for different language groups. The current study did not address the question whether the cognitive and personality scales 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 battery can only be given when data on the predictive bias are available. Although the jury is still out, the prospects for the personality instrument are dim because of its low reliability in notably the Black groups.

REFERENCES

Abrahams, F. (1996). *The cross-cultural comparability of the Sixteen Personality Factor Inventory (16PF).* Unpublished doctoral thesis, University of Pretoria, Pretoria, South Africa.

Abrahams, F. (2002). Fair usage of the 16PF (SA 92) in South Africa: A response to C. H. Prinsloo & I. Ebersohn. *South African Journal of Psychology*, *32*, 58-61.

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

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

Barrett, P. T., Petrides, K. V., Eysenck, S. B. G. & Eysenck, H. J. (1998). The Eysenck Personality Questionnaire: An examination of the factorial similarity of P, E, N, and L across 34 countries. *Personality and Individual Differences*, *25*, 805-819.

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

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

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

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

Chan, W., Ho, R. M., Leung, K., Cha, D. K-S., & Yung, Y-F. (1999). An alternative method for evaluating congruence coefficients with Procrustes rotation: A bootstrap procedure. *Psychological Methods*, *4*, 378-402.

Cheung, F. M., Leung, K., Zhang, J. X., Sun. H. F., Gan, Y. Q., Song, W. Z., & Xie, D. (2001). Indigenous Chinese personality constructs. *Journal of Cross-Cultural Psychology*, *32*, 407-433.

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

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309-219.

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

Claassen, N. C. W. (1993). *Verslag oor die funksionering van die NSAG intermedier G in verskillende bevolkingsgroepe*. Pretoria: Raad vir Geesteswetenskaplike Navorsing.

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

Cudeck, R., & Claasen, N. C. W. (1983). Structural equivalence of intelligence tests for two language groups. *South African Journal of Psychology*, *13*, 1-5.

Ellis, B. B. (1995, August). *Examination of the measurement equivalence of a Spanish version of the 16 PF using Item Response Theory*. Paper presented at the 103rd Annual Convention of the American Psychological Association, New York.

Freeman, M. C. (1984). *The effect of cultural variables on the Goodenough-Harris Drawing Test and the Standard Progressive Matrices*. Unpublished master's dissertation, University of the Witwatersrand, Johannesburg, South Africa.

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

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

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

Holburn, P. T. (1992). *Differential item functioning in mental alertness test*. Unpublished master's dissertation, University of South Africa, Pretoria, South Africa.

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

Horn, B.S. (2000). *A Xhosa translation of the revised NEO Personality Inventory*. Unpublished master's dissertation, University of Port Elizabeth, Port Elizabeth, South Africa.

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

McCrae, R. R., & Allik, J. (2002) (Eds.). *The five-factor model of personality across cultures*. New York: Kluwer Academic/ Plenum Publishers.

Meiring, D. (2000, June). *Revisiting the cross-cultural comparability of the 16 Personality Factor Inventory (16PF) in the South African context*. Paper presented at the Industrial Psychology Conference (incorporating the Psychometrics Conference), Pretoria, South Africa.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*, 105-118.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*, 338-354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376-398.

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

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

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

Paunonen, S. V., Zeidner, M., Enggvik, H., Oosterveld, P., & Maliphant, R. (2000). The nonverbal assessment of personality in five cultures. *Journal of Cross-Cultural Psychology*, *31*, 220-239.

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

Rogers, H. J., & Swaminathan, H. (1990). Detecting differential item functioning using logistic regression procedures. *Journal for Educational Measurement*, *27*, 361-370.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*, 105-117.

Roland, J. P., Parker, W. D., & Strumf, H. (1998). A psychometric examination of French translation of the NEO-PI–R and NEO-FFI. *Journal of Personality Assessment*, *71*, 269-29.

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

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

Tact, H. (1999). *The cross-cultural validity and compatibility of the Sixteen Personality Factor Questionnaire*. Unpublished master's dissertation, University of Pretoria, Pretoria, South Africa.

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

Taylor, T. R., & Boeyens, J. C. A. (1991). *A comparison of black and white responses to the South African Personality Questionnaire*. Pretoria: Human Sciences Research Council.

Tyler, G. (2002). *A review of the 15FQ+ Personality Questionnaire*. Pulloxhill, UK: Psychometrics Limited.

Van de Vijver, F. J. R. (2002). Cross-cultural assessment: Value for money? *Applied Psychology: An International Review*, *51*, 545-566.

Van de Vijver, F. J. R., & Leung, K. (1997a). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (2nd ed.). (pp. 257-300). Boston, MA: Allyn & Bacon.

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

Van de Vijver, F. J. R., & Leung, K. (2001) Personality in cultural context: Methodological Issues. *Journal of Personality*, *69*, 1006-1030.

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

Verster, J. M. (1974). A study of intellectual structure in two groups of South African scientists. *Psychologia Africana*, *15*, 169-190.

Verster, J. M., & Prinsloo, R. J. (1988). The diminishing test performance gap between English speakers and Afrikaans speakers in South Africa. In S. H. Irvine, S. H. Berry, J. W. (Eds.). *Human abilities in cultural context* (pp. 534-560). Cambridge: Cambridge University Press.

Vorster, J. F. (1978). The invariance of the factor structure of the intermediate and senior forms of the NSAGT. *Humanitas*, *3*, 409-417.