Roomaney, R. & Koch, E. (2013). An item and construct bias analysis of two language versions of a verbal analogies scale.
South African Journal of Psychology 43(3) 314–326.
http://dx.doi.org/10.1177/0081246313493377

An item and construct bias analysis of two language versions of a verbal analogies scale

Rizwana Roomaney and Elize Koch

Abstract

The Woodcock Muñoz Language Survey is a test of cognitive academic language proficiency that has been adapted from English into Xhosa by a South African team of researchers. This study was primarily concerned with the Verbal Analogies Scale of the Woodcock Muñoz Language Survey and aimed to extend previous research on the equivalence of the two language versions of the scale. The study employed a monolingual two-group design consisting of 150 mainly English-speaking and 149 mainly Xhosa learners in Grades 6 and 7. The first research objective was to investigate item bias (or differential item functioning items) in the Visual Analogies Scale across the Xhosa and English versions using logistic regression and Mantel–Haenszel statistical techniques. Five items were identified as differential item functioning. The second objective was to evaluate the construct equivalence of the two versions by conducting a factor analysis after removing the differential item functioning items from the scale. Two factors were identified. The first factor displayed significant loadings across both language versions. The second factor was stable for the English version but not for the Xhosa version. Results were supported by calculating a Tucker's phi coefficient for both factors. It was therefore concluded that Factor 1 is structurally equivalent across the two language versions but that Factor 2 was not structurally equivalent. Thus, the detection and removal of differential item functioning items did not result in structural equivalence.

The main focus of this study was an investigation into the equivalence of two language versions (English and Xhosa) of a Verbal Analogies (VA) Scale that is used for language testing in an additive bilingual education programme. This focus developed out of a broader need for tests that can be used in a multilingual South African (SA) society that are valid for use across language groups (Foxcroft, 1997). A requirement for tests that are valid across groups is built into current SA legislation (Employment Equity Act, 1998).

A number of studies on psychological and educational tests in the SA context demonstrated the level and extent of bias that is present in current available tests and supports this requirement and the focus of the study (Abrahams, 1996, 2002; Claassen, 1993; Koch, 2007; Meiring, Van de Vijver, Rothmann, & Barrick, 2005). One way of going about producing tests that are valid for use across groups is to have a test available in more than one language and to produce evidence that the scores of the two (or more) versions have the same meaning. These concepts and the methodology will be explored

further in this article. In conducting this study, the researchers hoped to learn lessons that can be applied in other studies of a similar nature in the SA context.

The study was conducted as part of the research into the Additive Bilingual Education Project (ABLE; see Koch, Landon, Jackson, & Foli, 2009 for a comprehensive discussion of the ABLE project and its aims). In this research, the Woodcock Muñoz Language Survey (WMLS) was adapted from English into Xhosa. The WMLS is a test of language proficiency (Woodcock & Muñoz-Sandoval, 2001) and was one of a battery of tests used to assess the language outcomes of the project. Additive bilingual education programmes are based on the sound pedagogical principle of allowing children to use their strongest language for conceptual development, while simultaneously developing proficiency in the language of power; in South Africa, this language is English. An emphasis is placed on the maintenance of the home language and effective acquisition of additional language(s). In piloting such programmes, evidence therefore needs to be provided that they are supporting the development of high levels of language proficiency in more than one language.

Tests play a crucial role in determining the outcomes of these programmes. However, the lack of appropriate tests, especially in indigenous African languages, led to the adaptation of a language test, the WMLS from English into Xhosa. Adaptation is the process whereby some items are literally translated and others are changed in a manner that enhances its cultural appropriateness (Poortinga & Van de Vijver, 2006).

This study is only concerned with the VA scale of the WMLS. Research on the English version of the VA scale, focusing on its validity across English and Xhosa groups produced evidence of bias and inequivalence (Ismail & Koch, 2012). This finding supported the need for the scale's availability in the home languages of learners. In contrast, previous research on the two language versions of the VA scale on unmatched groups (different ability groups) and using weighted multidimensional scaling (WMDS) provided evidence of construct equivalence, albeit on two dimensions in each language version (Koch, 2009). Only three items revealed differential item functioning (DIF) in this unmatched sample. However, secondary data analysis on the same data set but matched groups (on ability) using exploratory factor analysis (EFA) also revealed two factors on the two language versions but demonstrated construct equivalence on only one of the factors (Arendse, 2009).

This study was therefore an extension of the previous research on the two language versions of the WMLS by investigating the contribution of DIF items to the finding in the Arendse (2009) study, by performing a secondary data analysis on the matched sample groups (which will be described in the 'Method' section). The overall goal of this study was thus to further evaluate the equivalence of the two language versions of the VA scale of the WMLS. This goal consisted of two aims. The first aim was to investigate item bias or DIF in the VA scale across the Xhosa and English versions, using matched sample groups. The second aim was to evaluate the construct equivalence between the Xhosa and English versions of the VA scale, using matched sample groups after the DIF items were removed from the scale.

Related literature

VA tests are useful as they provide a measure of verbal reasoning that is independent of curriculum content (Primrose, Fuller, & Littledyke, 2000). This allows comparisons to be made between pupils from different school backgrounds. However, VA tests draw strongly upon previous exposure to language and are also regarded as being culturally specific, possibly rendering them inherently biased. However, these tests may provide a good measure of levels of cognitive functioning independent of subject content at a particular point in time (Primrose et al., 2000) and to assess the development of verbal reasoning over time. As such, VA tests, if they can be demonstrated to be valid measures of verbal ability, may end up being very important in the evaluation of bilingual and other educational programmes.

VA tests require the test-taker to complete an analogy by choosing one or two missing words (Goswami, 1991; Roccas & Moshinsky, 2003). In order to complete this task, the test-taker needs to understand the meaning of the question's words, determine the relationship between the words, and complete the analogy so that each pair of words have the same relationship (Roccas & Moshinsky, 2003). Analogies typically consist of a stem of two or three words that are to be matched to a correct answer from a number of response alternatives in a multiple choice format (Ulstadius, Carlstedt, & Gustafsson, 2008).

Analogies play a central role in learning and development from an early age. Analogies are often used in classrooms to aid learning in other areas such as decoding and comprehension, by facilitating the understanding of difficult texts in reading comprehension tasks and improving comprehension in science (Goswami, 1991). They are also commonly used in tests designed to predict academic success (Roccas & Moshinsky, 2003).

VA tests can be regarded as multidimensional as they measure both verbal ability and general ability of intelligence. Difficult analogies with rare words are dependent upon both vocabulary (language proficiency and verbal ability) and the cognitive capacity to detect the relationship between words (general ability). In order to measure general ability more purely, researchers use more well-known everyday words as this minimizes the impact of verbal ability. Ulstadius et al. (2008) are of the opinion that measures of general ability can be obtained by excluding foreign and infrequent words as this reduces the influence of vocabulary. This view also indicates that testing in a second language will be influenced by language proficiency. Therefore, in order to have good measures of verbal reasoning, it is important to have tests available in the first language of test-takers and provide a rationale for the adaptation of VA measures into more than one language.

In the adaptation of VA scales from one language to another, it has to be taken into consideration that the difficulty of verbal analogy items is influenced by a number of factors such as word-rarity (Ulstadius et al., 2008), the inclusion of a negative component (i.e., negative wording), and the order of words in the analogy (Roccas & Moshinsky, 2003). These factors should also be taken into account when developing, adapting, and/or translating analogy items. Items containing these factors should be adapted with care in order to minimize their effects. However, the availability of measure in more than one

language necessitates research into equivalence of the different language versions of scales. The theoretical framework that guided the methodology is provided in the next section.

Bias and equivalence

Bias and equivalence are pivotal and related concepts in measurement across groups (Van de Vijver & Tanzer, 2004). Bias refers to systematic errors in measurement or prediction of tests. When a test is biased, scores obtained from the test do not have the same meaning across cultures, and thus are not equivalent (Van de Vijver, 1998).

There are three types of bias. The first is known as construct bias and occurs when constructs being measured are not identical across cultures (Van de Vijver & Tanzer, 2004). A variety of statistical techniques can be employed to evaluate construct bias, such as EFA and multidimensional scaling (see Van de Vijver & Tanzer, 2004 for a thorough discussion). The second type is known as method bias, and this refers to all sources of bias emanating from methodology and procedures, including factors such as instrument differences, sample incomparability, tester and interviewer effects, and the mode of test administration (Van de Vijver & Rothmann, 2004). The final type is item bias; this refers to anomalies at an item level. An item is said to be biased if respondents with the same standing on the underlying construct but from different groups do not obtain the same score on the item. Item bias is also known as DIF (Van de Vijver, 1998). The removal of biased (DIF) items from any test may increase the reliability and validity of scores and their comparability (Hambleton & Kanjee, 1995); however, this is a hypothesis that needs to be tested.

Equivalence refers to the implications of bias with regard to the comparability of test scores and constructs (Meiring et al., 2005). Thus, the presence of bias in tests results in inequivalence and jeopardizes the comparability of test scores. It refers to whether the measurement level at which scores are acquired for diverse cultures or language groups (or on different language versions of a test) can be compared (Van de Vijver & Rothmann, 2004). There are three levels of equivalence, namely, structural/construct equivalence, measurement unit equivalence, and scalar equivalence (Van de Vijver, 1998). There is a hierarchical relationship between the levels of equivalence, with scalar equivalence being the highest level and imperative in the case of the comparison of scores across groups.

The first level of equivalence is known as construct equivalence or structural equivalence and exists when an instrument administered to different cultural (language) groups measures the same construct across groups (Van de Vijver & Rothmann, 2004). The presence of construct bias is an indication of construct inequivalence. In measurement unit equivalence, the different language versions of a measurement instrument have the same scale unit, but different origins. Examples of this are the Celsius and Kelvin scales of measurement. While no direct comparisons can be made across cultures with measurement unit equivalence, differences obtained within each group may still be compared across groups (Van de Vijver & Tanzer, 2004). The presence of item bias affects measurement level equivalence and might change either the origin or the unit of measurement (Van de Vijver, 1998).

Scalar equivalence exists when tests measure the same constructs and have the same origin and unit of measurement across groups (Van de Vijver, 1998). Only tests that demonstrate scalar equivalence can be used for comparisons of different language or cultural groups (Hambleton & Kanjee, 1995). Scalar equivalence cannot be proven directly but can be argued for by investigating all types of bias and then making claims with regard to whether there is evidence of scalar equivalence. The presence of bias always lowers the level of scalar equivalence. The ultimate aim of this research was therefore towards establishing the scalar equivalence of the two language versions of the VA scale.

Research on adapted tests

Previous research on test adaptation has demonstrated that translation and adaptation of tests into other languages often lead to biased items being produced, thereby hindering the scalar equivalence of the different language versions of a test (Lan, 2007; Meiring et al., 2005; Robin,

Table 1. Distribution of gender and grade per language group.

Language		English (*N*)	English (%)	Xhosa (*N*)	Xhosa (%)
Gender	Female	93	62	93	62.4
	Male	57	38	56	37.6
	Total	150	100	149	100
Grade	Grade 6	68	45.3	48	32.2
	Grade 7	82	54.7	101	67.8
	Total	150	100	149	100

Sireci, & Hambleton, 2003). Further adaptation may be used to achieve a higher level of structural equivalence, or a DIF analysis can identify DIF items to be removed which in itself may lead to better equivalence (Lan, 2007; Meiring, Van de Vijver, & Rothmann, 2006; Robin et al., 2003). VA items are particularly difficult to adapt or translate. One adaptation study demonstrated that items in a VA scale produced higher DIF rates compared to items requiring sentence completion, logic, and reading comprehension (Allalouf & Sireci, 1998). VA items are also less likely to retain their meaning when translated. These studies further support the focus of this study.

Method

This quantitative study was situated within the field of measurement theory, more specifically within the field of cross cultural and linguistic measurement. It was comparative research that employed a (mainly) monolingual, matched two-group design that focused on establishing the equivalence of the two language versions (English and Xhosa) of a VA scale (see Koch, 2009). Secondary analysis of data used in the main research study (discussed in the introductory section) was conducted. The two sample groups used in the study were matched on their total scores of the VA scale. Participants were identified as either mainly English speaking or mainly Xhosa speaking and were assigned to these 'monolingual' groups accordingly. Matching is regarded as an important design factor to control for the effect of ability on DIF and construct equivalence results (Sireci & Khaliq, 2002).

Participants

Researchers from the main study employed convenience non-probability sampling in order to select homogenous sample groups (as far as possible) in terms of educational background, various types of schools, grade, and gender (Koch, 2009). After matching the sample group on their total scores for this study, the sample consisted of 149 Xhosa and 150 English learners in Grades 6 and 7 from rural and urban areas in the Eastern Cape. The Xhosa learners did not include learners from ex-model C schools. Table 1 provides a description of participants in terms of gender, grade, and language.

Ethical clearance was obtained from the SA University at which the study was housed, as well as the Eastern Cape Education Department. Researchers contacted the principals of the schools and explained the project to them. The research was conducted in accordance with the ethical procedures of the university; parents completed consent forms, and only learners with consent forms were tested to collect data for the research on the test.

Instrument

The instrument that was being evaluated for equivalence was the VA scale of the English and adapted Xhosa versions of the WMLS. The WMLS is an individually administered test originally developed in the United States. It takes approximately 40 min to administer and can be applied to age groups 3–99 years (Woodcock & Muñoz-Sandoval, 2001). Since the WMLS is a commercially purchased test, the items are not made available due to confidentiality. In the United States, the test is available in English and Spanish.

The VA scale requires test-takers to complete an analogy by providing the missing word (Woodcock & Muñoz-Sandoval, 2001). It contains 35 items. The stimuli are auditory and the response oral, meaning that the test administrator reads the incomplete analogy to the test-taker who then responds orally by completing the analogy. The test-taker sees the sentence during reading. While the vocabulary remains simple throughout the test, the relationships become more complex.

The items are scored dichotomously. Correct answers obtain a score of 1 and incorrect answers a score of 0; answers are summed to obtain a total score. The total score for the scale therefore ranges between 0 and 35. The test is discontinued when the test-taker fails to respond correctly to three consecutive items. Raw scores were used in this study as normative data for SA population are unavailable at this stage of the development and validation of the instrument in South Africa. The reported median reliabilities for the sub-tests of the WMLS range between 0.80 and 0.93 for the original versions of the test (Woodcock & Muñoz-Sandoval, 2001). This was calculated on an American sample using split-half reliability.

Test adaptation process

The process of adaptation in South Africa was taken up by a multi-disciplinary and multilingual team that consisted of bilingual (English and Xhosa) language educators, Xhosa-speaking translators and linguists, language educators, and an English–Afrikaans bilingual psychometric expert with a background in research psychology. Test adaptation took place during two workshops, and the adaptation process was done in accordance

with the 22 guidelines of the International Test Commission for the development or adaptation of tests into more than one language (Koch, 2009).

The main adaptation on the whole test took place during the first 2-day workshop. Data were collected on this version, and the first exploratory analysis of equivalence was conducted across the two language versions of the test. Results indicated that the Xhosa version of the VA scale was problematic as there was no gradation in item difficulty. All items in the Xhosa version of the VA scale were difficult (mean = 0.29, as compared to the English version with a mean item difficulty of 0.41). In addition to this, 6 items were found to be biased (Koch, 2009).

A second 1-day workshop was then conducted in which the adaptation team took the decision to adapt the entire scale from scratch. The approach utilized during the second workshop entailed a further move away from direct translation and placed more emphasis on adaptation (see an explanation of adaptation in the introductory section). The focus was thus shifted to employing the same linguistic and cognitive processes that the analogies measured in English in a way that makes sense in the Xhosa language. An example is changing:

Idyasi iyanxitywa, njengoko i-apile i- . . . (English: the coat is being put on, the same as the apple . . .), *Echanekileyo* (correct answer): *iyatyiwa* (English: is being eaten), to: *Idyasi iyakhululwa, ibhanana i-* . . . (English literally: the coat is taken off, and the banana . . . ?) *Echanekileyo:iyaxotyulwa* (is peeled). (Koch, 2009, p. 76)

Data were then collected on the adapted second version of the WMLS, and this study utilized the VA data collected on this version.

Psychometric properties of the WMLS in South Africa

In a previous SA study on this adapted scale, the Cronbach's alphas for the English and Xhosa versions of the VA scale was found to be .78 and .75 respectively, indicating sufficient reliability for research purposes (Arendse, 2009). The psychometric properties of the WMLS have not yet been fully established for the locally adapted English and Xhosa versions, and this study forms part of the process (Koch, 2009).

Data analysis and results

Two techniques, namely, Mantel–Haenszel and logistic regression, were employed to investigate item bias or DIF across the Xhosa and English versions of the VA scale. Both analyses were run with the Statistical Package for the Social Sciences (SPSS) software. The Mantel–Haenszel procedure compares the likelihood of success on a particular test item between two groups that are matched on the construct of interest (Sireci, Patsula, & Hambleton, 2005); in the case of this study, the learners were matched on their total scores on the VA scale. The Mantel–Haenszel chi-squared statistic tests the null hypothesis that the odds of getting an item correct is the same for both the focal and reference groups across all levels (as categories) in the matching criteria (Kamata & Vaughn, 2004). A significant Mantel–Haenszel chi-square statistic is an indicator of DIF. DIF detection procedures may over-identify DIF in small samples (Robin et al., 2003), as was the case in this study. A stringent significant level was thus set at $p \leq .0001$.

While the Mantel–Haenszel method identifies uniform DIF, it is not effective in identifying non-uniform DIF (Sireci et al., 2005). Uniform DIF refers to a flagged item that affords a consistent advantage to the reference group throughout the distribution on ability, whereas non-uniform DIF refers to the conditional dependence shifts and changes in direction and degree at different points on the ability continuum (Osterlind & Everson, 2009). For this reason, a logistic regression was used to cross-validate the results and to detect both uniform and non-uniform DIF. Logistic regression estimates the relationship between a set of metric or non-metric variables and a single non-metric dependent variable. There is a general lack of assumptions in logistic regression, and this lack of strict assumptions allows its application to be appropriate in many situations (Hair, Black, Babin, & Anderson, 2009).

In the logistic regression DIF procedure, the dependent variable represents the likelihood of responding to an item in an estimably predictable manner such as correct or incorrect and is categorical in nature. The response is conditioned on group membership. Group membership is dummy coded for both focal and reference groups. The matching criterion (total VA score) and interaction term are the other independent variables (Osterlind & Everson, 2009). DIF exists when group membership and/or the interaction term rather than ability (the total score) contribute significantly to the likelihood of a correct response. The following model was utilized for DIF detection

$$P = (u = 1 \mid \theta, g) = \frac{e^{\tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g)}}{e^{\tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g)}} \qquad (1)$$

where the parameters $\tau 0$, $\tau 1$, $\tau 2$, and $\tau 3$ represent the intercept and the weights for the ability, group difference, and the ability and group interaction terms, respectively, θ is ability denoted by the total test score, and g is the group membership, in this case coded as 1 for the reference group (English) and 0 for the focal group (Xhosa). The logistic regression analysis involved a stepwise analysis in which three steps were entered. Step 1 entered the total score of subtest as the conditioning variable. In Step 2, the group membership was added to the analysis, and finally, in Step 3, the interaction of the group membership and the conditioning variable (total score on the subtest) were entered.

In this study, the logistic regression DIF identification consisted of two steps. The first step was to determine whether an item was biased or not (Jodoin & Gierl, 2001). In order to evaluate bias, the significance of the difference (DIF) chi-square distribution at 2 degrees of freedom was assessed using a stringent criterion ($p < .01$). Second, in order to assess the magnitude of DIF, the DIF effect size was obtained by calculating $\Delta R2$ between the models. Only items that displayed large DIF ($\Delta R2 > .07$) were further considered for removal.

Six items on the VA scale were identified as biased using the Mantel–Haenszel procedure. The logistic regression method identified seven items as displaying large DIF. Table 2 indicates which items were flagged using each technique. Due to logistic regression's tendency to over-identify DIF in small samples, such as was the case in study of sample

groups of about 150 (Robin et al., 2003) and Mantel–Haenszel's inability to identify uniform DIF, as previously mentioned, the null hypotheses of no DIF were only rejected for items identified as DIF by both techniques, thus controlling for both Type 1 and Type 2 errors arising from limitations in the two DIF detection methods.

The items that were identified as DIF by both procedures (Items 2, 6, 7, 15, and 18) were removed from the scale for Research Aim 2. EFA was used as the technique to explore this aim. In addition to these items, items that displayed no variance – no learners answered the items correctly in either the English or the Xhosa groups – (Items 27, 30, 31, 32, 33, 34, and 35) or exhibited low factor loadings in the previous EFA study on these data (Items 1, 5, 8, and 23) were also removed prior to the analysis (Arendse, 2009). Thus, 19 items remained for the factor analysis to test the hypothesis that the removal of items that are biased should result in a higher level of equivalence being established.

The second aim was thus to evaluate the construct equivalence between the Xhosa and English versions of the WMLS on the VA scale with the DIF items removed from the scale. This aim was assessed using EFA following Van der Vijver and Tanzer (2004). The Tucker's phi coefficient was used to further investigate the findings (Pienaar & van Wyk, 2006). EFA was conducted using the Comprehensive Exploratory Factor Analysis (CEFA) package (Browne, Cudeck, Tateneni, & Mels, 2008) to control for the fact that the items were dichotomously scored. Reise, Waller, and Comrey (2000) explain that dichotomously scored items in EFA can lead to serious distortions of the correlation matrix, as the phi coefficients of two items measuring the same construct can be significantly underestimated where response proportions differ. Dichotomously scored data also cannot meet the assumption of multivariate normality that is a requirement in estimation methods such as maximum-likelihood estimation. Tetrachoric correlations were thus used in the EFA estimations (see Browne et al., 2008; Reise et al., 2000). The study employed a Common Factor analysis as the method of extraction in order to identify underlying latent factors that the variables shared. An Oblique Geomin rotation was selected for this study, as it produces correlated factors in line with expectations in terms of the construct of VA (Hair et al., 2009; Ulstadius et al., 2008). Factor loadings of the Pattern Matrix table were used to consider the relative contribution of each variable to a factor.

The following criteria were taken into account in deciding about factor stability: items loaded on a factor when it had a factor loading of at least 0.30, individual items should not load on more than one factor (this requirement is often relaxed), and a minimum three items should load on a factor in order for that factor to be stable (Field, 2009; Hair et al., 2009).

Table 2. DIF items using both Mantel–Haenszel and logistic regression.

VA item	Mantel–Haenszel chi-square	Mantel–Haenszel significance	Logistic regression ΔR^2
VA1	0.004	.952	.01
VA2	**25.842**	**.000[a]**	**.16[b]**
VA3	0.399	.527	.01
VA4	0.073	.787	.01
VA5	3.209	.073	.11[b]
VA6	**15.141**	**.000[a]**	**.10[b]**
VA7	**42.890**	**.000[a]**	**.22[b]**
VA8	0.635	.426	.08[b]
VA9	7.949	.005	.03
VA10	2.294	.130	.02
VA11	1.730	.188	.01
VA12	1.246	.264	.01
VA13	8.680	.003	.04
VA14	4.451	.035	.02
VA15	**16.138**	**.000[a]**	**.07[a]**
VA16	15.092	.000[a]	.06
VA17	0.014	.906	.03
VA18	**42.731**	**.000[a]**	**.18[a]**
VA19	0.175	.676	.00
VA20	3.548	.060	.02
VA21	0.000	.993	.00
VA22	1.901	.168	.02
VA23	0.115	.735	.00
VA24	2.351	.125	.04
VA25	0.432	.511	.03
VA25	0.606	.436	.06
VA27	2.250	.134	.17
VA28	0.131	.718	.01
VA29	0.649	.420	.03
VA30	0.000	1.000	.12
VA32	0.500	.480	.13
VA34	0.500	.480	.13

Note. DIF = differential item functioning.
Items in bold were identified as biased using both techniques.
[a]Denotes significant Mantel–Haenszel.
[b]Denotes large DIF.

The EFA was run separately on the two language versions of the test. Due to findings of previous research on this scale and data set (as discussed earlier), two factors were specified from the outset in both analyses, as a one factor solution proved untenable based on the unexplained variance in the previous research (Arendse, 2009). Table 3 shows the results of the EFA.

For the English group, 11 items loaded significantly on the first factor, named Higher Order Verbal Reasoning. A total of 6 items loaded on the second factor that was identified as Direct Verbal Reasoning (Arendse, 2009). A total of 2 items loaded on both factors. Thus, both factors were relatively stable for the English versions.

Table 3. Rotated pattern matrix.

Item	English group		Xhosa group	
	Factor 1	Factor 2	Factor 1	Factor 2
VA3	−0.08	**0.69**	0.02	**0.88**
VA4	−0.11	**0.82**	0.05	**0.69**
VA9	0.07	**0.32**	**0.42**	−0.20
VA10	0.07	**0.66**	**0.50**	0.15
VA11	0.02	**0.61**	**0.42**	0.14
VA12	0.21	**0.37**	**0.47**	−0.12
VA13	**0.60**	0.23	**0.35**	0.17
VA14	**0.54**	0.14	**0.33**	−0.04
VA16	**0.51**	0.10	**0.52**	−0.06
VA17	**0.58**	**0.44**	**0.55**	**−0.51**
VA19	**0.70**	−0.02	**0.82**	−0.04
VA20	**0.93**	−0.13	**0.86**	**0.38**
VA21	**0.69**	−0.05	**0.57**	−0.21
VA22	**0.59**	**0.32**	**0.39**	−0.09
VA24	**0.84**	−0.05	**0.78**	0.01
VA25	**0.98**	−0.06	**0.61**	**−0.37**
VA26	**0.85**	−0.07	**0.97**	0.10
VA28	**0.78**	0.22	**0.72**	0.11
VA29	**0.68**	0.29	**0.68**	−0.15

Bold: significant loadings > .3.

For the Xhosa group, 14 items loaded on Factor 1, thereby indicating that Factor 1 once again emerged as a stable factor as this factor met the criteria specified above. Only two items loaded uniquely on Factor 2 while three items loaded on both factors.

The Tucker's phi coefficient (Tucker's coefficient of agreement) was used to estimate factorial agreement (Pienaar & van Wyk, 2006). The Tucker's phi is an indicator of how similar the pattern of low and high factor loadings are across different groups, and a high value indicates equivalence. Factor loadings indicate the relationship between a factor and test item. Tucker's phi was calculated using a free software program by Marley Watkins called Rc (Watkins, 2002). Tucker's phi values higher than 0.95 are viewed as evidence of factorial similarity, whereas values less than 0.85 may point to non-negligible incongruities (Van de Vijver & Leung, 1997). The Tucker's phi of each of the two factors in the two language versions was obtained.

The Tucker's phi value for the first factor was 0.93, and while this is not ideal, it can be viewed as pointing towards structural equivalence (Van de Vijver & Leung, 1997). The second factor produced a Tucker's phi of 0.42, indicating that the second factor was not structurally equivalent across the two language versions of the scale.

Discussion

The identification of DIF items in this study is similar to that of other studies that provided evidence of DIF in adapted tests and thus inequivalence of the different language versions of a test (Lan, 2007; Meiring et al., 2005; Robin et al., 2003). The result also echoes that of other studies that found it difficult to establish the equivalence

of different language versions of VA scales. VA items often produce higher DIF rates and are less likely to retain their meaning compared to other types of items that test verbal reasoning (Allalouf & Sireci, 1998; Allalouf, Rapp, & Stoller, 2009). The removal of the DIF items from the scale, furthermore, did not lead to a finding of structural equivalence.

The researchers in this study posit two alternative explanations for the failure to establish equivalence in this particular VA scale. First, as previously noted, analogies do not always make sense when translated. Even though the adaptation team used a creative approach to ensure item meaning and the grading of item difficulty during the adaptation process (as discussed under method), these results may nevertheless be a reflection of the true state of affairs. If this is the case, then this study has not provided evidence of scalar equivalence, and the two language versions can therefore not be used for comparison and cannot be added on a common scale for norming.

Alternatively, the presence of two factors in mainly the English version of the scale may be a spurious finding as a result of the technique used. Items may not have clustered in terms of factors, but rather in terms of difficulty; this is a limitation of EFA at item level (Kishton & Wideman, 1994 cited in G. De Bruin, 2004). An investigation into the differences in item difficulty across the two language versions and its impact on the results was not an aim of this study. Preliminary research did not find differences in the mean item difficulties of the two versions per se (English = 0.35; Xhosa= 0.39; Koch, 2009), but further research to assess its impact on the EFA results may need to be conducted. In addition, further research into DIF and construct equivalence, using Rasch modelling is recommended, especially given the results of the WMDS on this scale in Koch (2009). Unidimensionality is also an important assumption that is tested in Rasch modelling. This technique thus holds much promise for further research into this scale.

It can therefore be concluded that the removal of DIF items did not result in producing evidence towards scalar equivalence of the two language versions of the VA scale, using EFA as a technique to explore construct bias. However, given the small number of DIF items using a research design that control for the effect of difference in ability on DIF findings, the previous finding of construct equivalence, and the limitations of using EFA at an item level, the need for further research into the two language versions has been established. The need for tests in more than one language that can be used to improve the valid use of tests across groups has been established. The approach that was followed in the adaptation of the VA scale of this test is an approach that holds much promise for test adaptation practices in the SA context.

Declaration of conflicting interests
None declared.

Funding
This research received the NRF Thuthuka grant for Elize Koch, 2005–2010.

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.

Allalouf, A., Rapp, J., & Stoller, R. (2009). Which item types are better suited to the linking of verbal adapted tests. *International Journal of Testing*, *9*, 92–107.

Allalouf, A., & Sireci, S. G. (1998, April). *Detecting sources of DIF in translated verbal items*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, USA.

Arendse, D. (2009). *Evaluating the structural equivalence of the English and Xhosa versions of the Woodcock Munoz Language Survey on matched sample groups* (Unpublished master's thesis). University of the Western Cape, Cape Town, South Africa.

Browne, M. W., Cudeck, R., Tateneni, K., & Mels, G. (2008). CEFA: Comprehensive exploratory factor analysis. Retrieved from http://faculty.psy.ohio-state.edu/browne/software.php

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

De Bruin, G. (2004). Problems with the factor analysis of items: Solutions based on item response theory and item parcelling. *South African Journal of Industrial Psychology*, *30*(4), 16–26.

Employment Equity Act. (1998). Number 55. Department of Labour.

Field, A. (2009). *Discovering Statistics using SPSS*. 3rd ed. London: Sage Publications.

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

Goswami, U. (1991). Analogical reasoning: What develops? A review of research and theory. *Child Development*, *62*, 1–22.

Hair, J. F., Black, B., Babin, B., & Anderson, R. E. (2009). *Multivariate data analysis*. Upper Saddle River, NJ: Pearson Publishing.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment*, *11*, 147–157.

Ismail, G., & Koch, E. (2012). Investigating item and construct bias in an English verbal analogies scale. *South African Linguistics and Applied Language Studies*, *30*, 325–338.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*, 329–349.

Kamata, A., & Vaughn, B. K. (2004). An introduction to differential item functioning analysis. *Learning Disabilities: A Contemporary Journal*, *2*(2), 49–69.

Koch, E. (2007). The monolingual testing of competence: Acceptable practice or unfair exclusion. In P. Cuvelier, T. Du Plessis, M. Meeuwis, & L. Teck (Eds.),

Multilingualism and exclusion: Policy, practice and prospects (pp. 79–103). Pretoria, South Africa: Van Schaik Publishers.

Koch, E. (2009). The case for bilingual language tests: A study of test adaptation and analysis. *South African Linguistics and Applied Language Studies*, *27*, 301–317.

Koch, E., Landon, J., Jackson, M. J., & Foli, C. (2009). First brushstrokes: Initial comparative results on the additive bilingual education project (ABLE). *Southern African Linguistics and Applied Language Studies*, *27*, 109–127.

Lan, O. S. (2007). Comparing two language versions of science achievement tests using differential item functioning. *Jurnal Pendidik dan Pendidikan*, *22*, 45–59.

Meiring, D., Rothmann, S., & Van de Vijver, F. J. R. (2006). Bias in an adapted version of the 15FQ+ in South Africa. *South African Journal of Psychology*, *36*(2), 340–356.

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.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Thousand Oaks, CA: SAGE.

Pienaar, J., & van Wyk, D. (2006). Teacher burnout: Construct equivalence and the role of union membership. *South African Journal of Education*, *26*, 541–551.

Poortinga, Y. H., & Van de Vijver, F. J. R. (2006). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting Educational and Psychological Tests for Cross-cultural Assessment*. London, England: Routledge Academic.

Primrose, A. F., Fuller, M., & Littledyke, M. (2000). Verbal reasoning test scores and their stability over time. *Educational Research*, *42*, 167–174.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment*, *12*, 287–297.

Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing*, *3*, 1–20.

Roccas, S., & Moshinsky, A. (2003). Factors affecting the difficulty of verbal analogies. *Applied Measurement in Education*, *16*, 99–113.

Sireci, S. G., & Khaliq, S. N. (2002). *Comparing the psychometric properties of monolingual and dual language test forms* (Center for Educational Assessment Research No. 458). Amherst: School of Education, University of Massachusetts Amherst.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting Educational and Psychological Tests for Cross-cultural Assessment* (pp. 93–106). Mahwah, NJ: Lawrence Erlbaum Associates.

Ulstadius, E., Carlstedt, B., & Gustafsson, J. (2008). The multidimensionality of verbal analogy items. *International Journal of Testing*, *8*, 166–179.

Van de Vijver, F. J. R. (1998). Towards a theory of bias and equivalence. *ZUMA-Nachrichten Spezial*, 41–65.

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

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

Van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée*, *54*, 119–135.

Watkins, M. W. (2002). *Coefficient of congruence* (Rc) [Computer software]. State College, PA: Author.

Woodcock, R. W., & Muñoz-Sandoval, A. F. (2001). *Woodcock-Muñoz Language Survey: Normative update*. Itasca: Riverside Publishing Company.