Southern African Linguistics and Applied Language Studies
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rall20

Investigating item and construct bias in an English verbal analogies scale

Ghouwa Ismail [a] & Elize Koch [b]

[a] Medical Research Council, University of South Africa Safety and Peace Promotion Research Unit, PO Box 19070, Tygerberg, 7505, South Africa

[b] Research Associate, Education Faculty Nelson Mandela Metropolitan University, PO Box 77000, Port Elizabeth, 6031, South Africa

Version of record first published: 19 Dec 2012.

To cite this article: Ghouwa Ismail & Elize Koch (2012): Investigating item and construct bias in an English verbal analogies scale, Southern African Linguistics and Applied Language Studies, 30:3, 325-338

To link to this article: http://dx.doi.org/10.2989/16073614.2012.739407

Southern African Linguistics and Applied Language Studies 2012, 30(3): 325–338

SOUTHERN AFRICAN LINGUISTICS
AND APPLIED LANGUAGE STUDIES
ISSN 1607-3614 EISSN 1727-9461
http://dx.doi.org/10.2989/16073614.2012.739407

Investigating item and construct bias in an English verbal analogies scale

Ghouwa Ismail[1]* and Elize Koch[2]

[1]Medical Research Council – University of South Africa Safety and Peace Promotion Research Unit, PO Box 19070*,* Tygerberg, 7505, South Africa

[2] Research Associate, Education Faculty Nelson Mandela Metropolitan University, PO Box 77000, Port Elizabeth 6031, South Africa

**Corresponding author, e-mail: ghouwa.ismail@mrc.ac.za*

Abstract: For this study the researcher was interested in investigating the scalar equivalence of the adapted English version of the verbal analogies (VA) subscale of the Woodcock Muñoz Language Survey (WMLS) across English first-language speakers and Xhosa first-language speakers. This was achieved by utilising differential item functioning (DIF) and construct bias statistical techniques. The Mantel-Haenszel DIF detection method was employed to detect DIF, while construct equivalence was examined by means of exploratory factor analysis (EFA) utilising an *a priori* two-factor structure. The Tucker's phi coefficient was used to assess the congruence of the construct across the two language groups. The sample consisted of 192 English and 193 Xhosa first-language learners, who were selected from ex-Model C and previously disadvantaged schools in the Port Elizabeth and Grahamstown region. The main findings of this study indicated that the adapted English version of the VA scale displayed DIF items across the two language groups. Moreover, construct equivalence could only be established for one factor across the two language groups, as the second factor displayed non-negligible incongruities even after the removal of DIF items.

Introduction

Analogy is a powerful means by which children acquire knowledge, and a developmental skill that mediates the progression of children's cognitive abilities, more specifically verbal reasoning ability. Theorists have attempted to explain and understand verbal reasoning in terms of verbal analogies. Therefore, verbal analogies tests are often used to measure verbal reasoning ability in intelligence tests.

Internationally there has been a trend of increasing acceptability coupled with positive perception when it comes to using psychological tests and testing (Foxcroft *et al*., 2004). Test results provide a wealth of information in a short period of time and can be used to form the basis for comparisons of groups, or the evaluation of a test-taker (Paterson & Uys, 2005). However, while tests are important tools, frequently used in assessment, they can also act as a disabling factor if they are inappropriately applied in diverse groups (Paterson & Uys, 2005). This is also referred to as cross-cultural and cross-linguistic testing. A fundamental question in cross-cultural and cross-linguistic testing is whether or not test scores have the same meaning across groups. One issue that could affect the meaning of test scores is bias. Bias affects equivalence, and therefore comparability and validity. 'Equivalence' refers to the measurement level at which scores can be compared across language or cultural groups, while 'bias' refers to nuisance factors differentially affecting test scores across different groups. The concepts of 'bias' and 'equivalence' will be discussed in more depth in the ensuing section. Standardised measures must be equivalent across different language groups before the scores of diverse groups can be compared; in other words, without equivalent measures, observed scores from the different language groups are not directly comparable (Van der Vijver, 1998). All tests that are used across diverse groups, accordingly, have to be evaluated for equivalence before comparative studies can proceed.

The present study falls under the umbrella of a larger study consisting of numerous phases concerning the adaptation and validation of the Woodcock Muñoz Language Survey (WMLS)

within the South African context (Koch, 2009). One of the study's objectives is the evaluation of the construct validity of different language versions of the test across two groups, English and Xhosa first-language speakers, within the South African context. Specifically, this article presents the results of a study investigating the scalar equivalence of the English version of the verbal analogies (VA) subscale of the (WMLS) across English and Xhosa first-language speakers. This scale is utilised as an English monolingual scale to assess the verbal reasoning of not only English first-language learners but also Xhosa first-language learners. The scalar equivalence of the scale across the two groups therefore needs to be established. Scalar equivalence cannot be investigated directly but is empirically and inductively investigated by means of construct (also called structural) and item bias (Van der Vijver, 1998).

For the purpose of this article a distinction is made between monolingual testing and cross-lingual or cross-cultural testing. 'Cross-lingual and cross-cultural testing' refers to tests that have been adapted or translated for use across diverse groups, while 'monolingual testing' refers to tests that are available in only one language but are administered across diverse language groups (Koch, 2005). The English version of the VA scale is an example of a monolingual test.

The promising results on the English version of the VA scale in previous research (Haupt, 2009) where only a few items were found to be biased across the two groups, necessitated cross-validating the results on the scale by using a different DIF technique. DIF research is notorious for type 1 error in the case of small samples such as was the case in this research (Sireci & Khaliq, 2002). Supporting the findings of the previous research using a different DIF technique will therefore strengthen the internal validity of the study. The previous research also found good internal consistency across the English and Xhosa groups with a Cronbach's Alpha of 0.83 and 0.86 respectively (Haupt, 2009). A logistic regression differential item functioning (DIF) analysis found only four items (out of 35 items) displaying DIF on this scale, with two items having large DIF and two items having moderate DIF.

The purpose of this article is to report on the scalar equivalence of the adapted English version of the VA scale of the WMLS across two language groups, namely an English first-language group and a Xhosa first-language group. Scalar equivalence can only be demonstrated when the same construct is measured, with no construct and item bias. Thus scalar equivalence will indirectly be investigated by assessing construct and item bias as evidence towards it.

The specific research aims were to:

(i) Evaluate the differential item functioning (DIF) of the English version of the VA scale across English and Xhosa first-language groups.
(ii) Assess the construct bias of the English version of the VA scale across English and Xhosa first-language groups, initially with all the items included and subsequently with the DIF items removed.

Verbal reasoning

Verbal reasoning, defined as an ability to reason via analogy, according to Spearman (1927) and Sternberg (1977), is fundamental in human intelligence and numerous forms of analogy tests are thus often used for measuring general ability. In this study the terms 'verbal reasoning' and 'verbal analogy' will therefore be used interchangeably. The ability to reason by analogy is generally considered a core component in the development of human cognition. It provides an important foundation for learning and classification, as well as for thought and explanation. Verbal reasoning is viewed as encompassing higher-order reasoning skills which promote the ability to transfer knowledge to new situations, perform successfully on novel problems, and learn by integrating a variety of information from diverse contexts (Goswami & Brown, 1990). Holyoak and Thagard (1995) postulated that the act of formulating an analogy necessitates perceiving one thing as if it were another, and thus the perceiver is required to make a kind of 'mental leap' between domains. This coincides with the generally accepted view of most researchers that verbal reasoning involves reasoning pertaining to relations, in particular with regard to relational similarity, in order that a correlation is ascertained between one set of relations and another (Goswami, 1991; Tagalakis & Keane, 2006). In other words, an individual recognises the relational similarity, for example that a dog is more related to a cat than to a camel. According to Goswami (1991) and Cummins (1992),

this definition allows verbal reasoning to encompass problem-solving by using the solution to a known problem to solve a structurally similar problem. Thus, being able to identify these abstract similarities is the underlying attribute of verbal reasoning.

According to Primrose and colleagues (2000), verbal reasoning tests are inherently biased as they are dependent on prior exposure to language. What they assert is that the acquired knowledge and skill measured in verbal reasoning tests are those associated with language and its everyday use (Primrose *et al*., 2000). Research indicates that verbal reasoning test scores could only be used in assessing the learner's current level of achievement as they were demonstrated to fluctuate over time and are thus not stable indicators in measures of future academic potential (Primrose *et al*., 2000). It was concluded that verbal reasoning assessments provide good measures of the levels of cognitive functioning at a particular point in time (Primrose *et al*., 2000).

Monolingual tests

There is new awareness about the limitations of monolingual tests and their use in multilingual and multicultural contexts. Researchers in various countries therefore conducted research on various issues surrounding the use of monolingual tests.

Allalouf and Abramzon (2008) assert that the use of a monolingual test across two language groups is problematic, because a single test form cannot assess proficiency where there is a large variation in the nature of language ability between the two groups. What this implies is that when a particular construct is being explored, problems may arise as to whether the same underlying construct is being measured in each language group.

Language can cause complications on three levels: (i) the language in which the test is constructed; (ii) the difficulty level of the test language, in particular if the test is administered in the test-taker's second or third language (Van de Vijver & Rothmann, 2004); and (iii) the language competence of the test-taker (Paterson & Uys, 2005). Additionally, Huysamen (2002) contends that cultural contexts should also be considered when using a monolingual test as this may account for the poor standing of test-takers on the construct measured, and not owing to poor performance on their part. Huysamen (2002) refers to this as 'construct-irrelevance', which occurs when a construct being measured may be relevant to one group and not to another. He further asserts that irrelevant variance may not be restricted to language proficiency only, but could extend to cultural differences that the test is not designed to measure (Huysamen, 2002). Thus, it is important to determine whether the performance on the test reflects the test-taker's ability, and not the level of competence in the test language (Foxcroft *et al*., 2004). In other words, one has to ensure that the same construct is measured across groups of different languages.

Following global trends, researchers and clinicians in South Africa tends to use monolingual tests to measure individuals on a particular trait. The dilemma in using these tests is threefold: (i) the tests have not been developed or adapted for use in a multicultural and multilingual context; (ii) some of the tests (e.g. the Bender and the Beery VMI) have been imported from overseas and full-scale national normative studies have never been carried out in an attempt to provide practitioners with appropriate norms; and (iii) a number of the tests developed in South Africa are outdated as they were only developed for specific groups of South Africans e.g. SSAIS-R or JSAIS (Foxcroft *et al*., 2004), leaving the rest of the population discriminated against. In addition, monolingual measures are oblivious to the fact that bilingual individuals may prefer using different words depending on the setting, interlocutor, and context (Iglesias, 2001) as well as their cultural experiences (Peña, 2001).

In a study on a Hebrew Proficiency Test (HPT), Arabic and Russian first-language (L1) participants were examined on differences in performance on second-language (L2) test items. Results revealed that vocabulary and grammar items usually favoured the Arabic speakers because of the similarity between Arabic and Hebrew and because of the presence of cognates in the test. Thus, the HPT functioned differently across the two groups (Allalouf & Abramzon, 2008).

When tests are based on theories that are sensitive to cultural context and environmental influences, construct equivalence is less likely to be observed (Rossier, 2004). Cultural context in particular becomes a problem when performing personality assessments, as constructs have

different meanings and are experienced differently across cultures (Paterson & Uys, 2005). Personality tests in particular require high levels of language proficiency (Vijver & Rothmann, 2004). In this case the cross-cultural equivalence in the scenario of testing individuals from different language groups is problematic.

A South African study conducted by Abrahams and Mauer (1999) investigated the impact of home language on response to items of the Sixteen Personality Factor Questionnaire (16PF). They found that anomalies existed as far as the comparability of items across groups were concerned, thus impacting the cross-cultural use of this measure. In another recent South African study, Koch (2007) evaluated a reading comprehension test for equivalence across three language groups, namely, English, Afrikaans and African-language speakers. It was concluded that the scores of the English L1 and L2 students on the reading comprehension test could not be used to make equivalent statements regarding the construct measured across the groups, since the test displayed unacceptable levels of item bias as well as construct bias.

Further research was conducted by Koch and Dornbrack (2008) evaluating bias in the South African context, particularly with regard to monolingual admissions English language criteria. The study revealed that the criteria for admission will prejudice the African-language and Afrikaans speaking students. Thus, the evaluation of students' performance in a single language as representative of their academic literacy in the language of teaching and learning can be viewed as biased and problematic (Koch & Dornbrack, 2008).

A test that is biased in one context may not be biased in another (Van de Vijver & Tanzer, 2004) and, as a result, needs be evaluated in the context of their usage. Research regarding bias and equivalence of tests in South Africa is still in its infancy. The Equity Act 55 of 1998 demonstrates a zero-tolerance approach stipulating the prohibition of psychological testing and other assessment measures, unless scientific validity and reliability, fairness, and non-bias against participants can be validated (Van de Vijver & Rothman, 2004). Van de Vijver and Rothmann (2004) contend that much more research is required on bias and equivalence of assessment tools used in a South African context before tests and testing can live up to the demands implied in this Act. This serves as a further motivation for conducting the research in this study.

A theoretical framework of equivalence and bias

The concept of bias and the attainment of equivalence are of fundamental significance in cross-linguistic research and testing in multilingual and multicultural contexts (Van de Vijver & Leung, 1997). These concepts are associated with the validity of a measure and are intrinsic in the characteristics of an instrument in cross-linguistic comparison (Van de Vijver, 1998).

Though bias and equivalence is interrelated (Van de Vijver, 1998), they provide different perspectives on the same question, namely the extent to which scores have the same meaning across groups (in the case of monolingual tests), or different languages versions of a test.

For test scores to be comparable it has to be demonstrated that the test is not biased. Equivalence is always challenged when bias, at any level, occurs and thus to maintain the utmost level of equivalence, the adapted measure and its subsequent application must be as free from bias as possible (Van de Vijver, 1998). Sources of bias in multilingual or cultural assessment can be distinguished into three categories, namely: construct, method, and item bias (Van de Vijver & Rothman, 2004). Construct bias occurs when the construct (of this study, 'verbal reasoning'), is not identical across groups (Van de Vijver & Rothman, 2004). Method bias consists of sample bias, administration bias, and instrument bias and refers to all sources of bias emanating from a methodological-procedural aspect which includes factors such as sample incomparability, instrument differences, tester and interviewer effects, and the mode of administration (Van de Vijver, 1998; Van de Vijver & Rothman, 2004). Item bias, also known as differential item functioning (DIF), refers to anomalies of an instrument at an item level (Van de Vijver, 1998).

DIF is a statistical analysis procedure and has been utilised widely to evaluate adapted tests – that is, tests that are available in two or more language versions (Gierl & Khaliq, 2001). DIF identifies items that function differently across two different groups, and is based on the underlying assumption that examinees with similar ability should perform similarly on an item (Sireci & Allalouf,

2003). DIF occurs when an item is significantly more difficult for one group than for another when ability is held constant (Allalouf *et al.*, 1999; Sireci & Allalouf, 2003). Van de Vijver and Leung (1997) regard DIF as 'dangerous' for equivalence and maintain that DIF results in compromised scalar equivalence; thus the comparability of test scores across groups.

According to Hambleton and Kanjee (1995), for any comparison between different language groups to be valid, the test utilised must demonstrate equivalence. There is a hierarchical order in the types of equivalence and they can be divided into three categories, namely: construct equivalence, measurement unit equivalence and, at the top of the hierarchy, scalar equivalence. At the lowest level of the hierarchy is construct (also called structural) equivalence which occurs when the instrument measures the same construct across different language groups (Van de Vijver, 1998; Van de Vijver & Rothman, 2004). Measurement unit equivalence occurs when instruments have the same units of measurement across language groups but the origin differs, such as the Kelvin and Celsius scales in temperature measurement (Van de Vijver & Rothman, 2004). Scalar equivalence assumes that identical interval or ratio scales apply to measures in the language groups compared and cannot be assessed directly (Van de Vijver, 1998; Van de Vijver & Rothman, 2004). Thus, scalar equivalence can only be demonstrated when the same construct is measured, with no item and measurement bias. This is the only type of equivalence that allows the researcher to make valid conclusions when averages are compared across language groups, for example, by utilising t-tests and analysis of variance (Van de Vijver, 1998; Van de Vijver & Rothman, 2004).

Research methodology

Design

The researchers evaluated the scalar equivalence (by means of investigating construct and item bias) of the VA scale across two language groups, Xhosa first-language speaking and English first-language speaking groups. Comparative and correlational statistical techniques were used to conduct comparisons between the two language groups on the English version of the VA scale. A differential research design was utilised (Gravetter & Forzano, 2008).

Participants

Since the researcher was using SDA, the participants of the larger study were retained for the present study (see Ismail, 2010). The participants consisted of 198 English first-language learners and 197 Xhosa first-language learners, who were tested on the English version of the WMLS during the second half of 2006 and the second half of 2007. The English and Xhosa first-language speakers were selected from ex-model C and previously disadvantaged schools in the Port Elizabeth and Grahamstown regions. The sampling procedure used in the main study consisted of convenience purposive sampling.

Ethics

All research procedures and data collection were done strictly in accordance with the ethical regulations of the University of Port Elizabeth. The current researcher received permission from the main researcher to use and re-analyse the data collected for the main study (see Ismail, 2010).

Test

The WMLS consists of four sets of individually administered scales designed to measure a broad sampling of proficiency in four critical areas of oral language, listening, reading, and writing. The four subscales are: Picture Vocabulary, Verbal Analogies (forming the oral language cluster), Letter Word Identification, and Dictation (forming the reading-writing cluster). These scales are primarily measures of language skills predictive of success in situations characterised by cognitive academic language proficiency (CALP) requirements (Woodcock & Muñoz-Sandol, 2001). In other words, the instrument provides an overall measure of language competence as well as CALP.

The WMLS was standardised on populations in the USA, central America, South America and Spain. The median reliabilities were found to range from 0.80 to 0.93 for the scales and 0.88 to

0.96 for the clusters. The median reliabilities for the VA scale were found to be 0.81 (Woodcock & Muñoz-Sandol, 2001).

This study utilised one of the scales of the English version of the WMLS, namely the VA Scale. This 35-item scale is used to measure listening and speaking skills, either individually or collectively, and purports to assess an individual's ability to complete oral analogies, which necessitates verbal comprehension and verbal reasoning, such as 'A bird flies; a fish swims'. The vocabulary remains simple throughout, but the relationships become increasingly complex.

Though research is currently in progress (Koch, 2009), the WMLS has not yet been normed for the South African population. Therefore, a complete psychometric properties dossier of the test for the South African context is not yet available. The research in progress indicates that both the adapted English version and the Xhosa version of the WMLS demonstrate promising results on two of the scales of the test, namely the VA and the Letter-Word Identification (LWI) (Arendse, 2009; Haupt, 2009). According to results on the English version of the VA Scale, in particular, good internal consistency was displayed across the English first-language and Xhosa first-language groups, with a Cronbach's Alpha of 0.83 and 0.86 respectively (Haupt, 2009). Furthermore, a logistic regression differential item functioning (DIF) analysis across English and Xhosa first-language groups on the English version of the scale indicated that only six items (1, 5, 8, 9, 14 and 18) displayed DIF on this scale, two items having large DIF, two items having moderate DIF and two items displaying negligible DIF (Haupt, 2009).

Data Analysis

Procedure one: Item bias (also called DIF), which relates to research aim (i), was explored by means of the Mantel-Haenszel DIF detection method using the Mantel-Haenszel chi-square statistic. The Mantel-Haenszel DIF technique is a commonly used procedure to detect bias in dichotomously scored data (Sireci & Allalouf, 2003). In the current study one would expect the English and Xhosa first-language groups who have the same total test score to perform in an equivalent manner on each VA item. The items with significant Mantel-Haenszel chi-squared statistics were identified as biased, and thus the null hypothesis on these items were rejected. The MH chi-square was computed using the crosstabs procedure in the statistical software SPSS package. The significance of the chi-square was assessed using a stringent *p*-value of 0.0001 ($\rho < 0.0001$). Items that met this criterion were flagged as displaying DIF. Furthermore, a 'constant odds ratio' was used to provide an estimate on the magnitude of the DIF (Sireci & Allalouf, 2003).

This DIF effect size estimate ranges from zero to infinity with an expectation of 1 under the null hypothesis of no DIF (Dorans & Holland, 1993). Thus, a value of 1 implies that there is no differential item performance between the two groups, larger values imply that the item favours the reference group, and values smaller than 1 indicates possible bias against the focal group. The DIF effect size estimate is usually rescaled onto the delta metric to make it more interpretable. However, the effect size was not used in this study as a criterion for detecting DIF items. The current study used a stringent significance value of 0.0001 ($p = 0.0001$) in order to detect DIF items.

This transformed effect size (MH D-DIF) is calculated as:

$$MH\_D\text{–}DIF = -2.35\ln [\alpha\ M\ H]$$

A MH D-DIF value of 1.0 is equivalent to a difference in proportion corrected to about 10%. Rules of thumb exist for classifying these effect sizes into small, medium, and large DIF (Dorans & Holland, 1993). According to Kamata and Vaughn (2004), a MH D-DIF displaying an absolute value greater than 1.5 and significantly greater than 1.0 (at $\alpha = 0.05$) is regarded as a category C item and thus is flagged for large DIF. Any item with a MH D-DIF value less than 1.0 or not significantly greater than zero (at $\alpha = 0.05$), is a category A item and is considered negligible for DIF, while category C items display intermediate DIF with absolute values significantly greater than 1.0 and less than 1.5 or not significantly greater than 1.0. These categories were used to classify the DIF items as large, medium or negligent DIF.

Procedure two: Exploratory factor analysis of dichotomous items at an item-level was utilised to evaluate construct equivalence across the two groups as indicated in research aim (ii), using tetrachoric correlations to extract the factors (Kubinger, 2003).

The Tucker's phi coefficient was used to assess the congruence of the construct(s) across the two language groups. The Tucker's phi coefficient is commonly used to evaluate the similarity of factors across different groups (Zumbo *et al*., 2003).

Tucker's phi values higher than 0.95 are viewed as evidence of factorial similarity, whereas values less than 0.85 may indicate non-negligible incongruities (Van de Vijver & Leung, 1997). The aforementioned is regarded as a rule of thumb and thus requires no hypothesis. There are, however, some theorists who have used a more relaxed Tucker's phi value of 0.90 or 0.80 as an indication of factorial similarity (Van der Oord *et al*., 2005).

A scatter plot was used to assess the similarity of the factor patterns by means of cross-plotting the factor pattern coefficients of the two groups and drawing an identity line through the plotted points. Ideally the points on the plot should fall close to the identity line (De Bruin, 2009).

The current study utilised a Common Factor analysis in order to ascertain whether the variables shared underlying latent factors. Since an *a priori* factor structure was employed based on previous research on the Xhosa version of the scale (see Arendse, 2009), the use of a scree-plot and its eigenvalues to determine how many factors to retain, was excluded. A two factor solution was therefore specified from the outset. Exploratory Factor analysis identifies latent subsets of characteristics or factors that underlie a specific domain (Schaap & Vermeulen, 2008). According to de Wet (2005, as cited in Taliep, 2012) the use of exploratory factor analysis as opposed to confirmatory factor analysis, is appropriate when the aim of the study is to examine the factor structure of a questionnaire in a population or language group. There are no inferential statistics (i.e. testing of hypothesis and making decisions regarding the acceptance or the rejection of hypothesis on the basis of probability). Thus, it is an approach that quantifies or measures the similarity of the factor loadings across groups (i.e. the two language groups) by rotating the two factor solutions to be most advantageously similar, and then computing some sort of similarity index.

An oblique rotation was decided on for this study, as it produces correlated factors facilitating easy interpretation (Hair *et al*., 2010) and one is likely to discover a relationship between factors (Cummins, 2000). No target rotation was applied prior to comparing the factors. In order to consider the relative contribution of each item to a factor, the Pattern Matrix table was examined using a strict critical value of 0.40 (Hair *et al*., 2010) to evaluate the factor loadings on the two factors. Items that loaded on more than one factor were regarded as poor items, as at least three items should load on a factor in order for it to be considered a stable factor.

The factor analysis was run separately for the English and Xhosa first-language groups, and the results were compared. The first phase of the factor analysis required the selection of a two-factor solution using the data of the English first-language speaking group first. A two-factor solution was specified based on a previous study conducted by Arendse (2009) across two language versions of the VA scale of the WMLS, namely an English version and a Xhosa version. This study revealed a stable structure for a two-factor solution across both language versions.

The other steps that were followed in this analysis will be described in the results section. Subsequently, the analysis of the data for the Xhosa first-language group was specified to include the same items, as well as using a two-factor solution.

Results

Item bias

These results link to research aim (i), which evaluates the differential item functioning of the scale across the two language groups. Using a strict significance level of 0.0001 (p = 0.0001) to detect DIF items, the Mantel-Haenszel DIF procedure identified three items (8, 9, 18), all displaying large DIF as indicated in Table 1 below.

Items 8 and 9 identified in Table 1 indicate that the English first-language group is favoured on two of the three items. This is in corroboration with the findings of Haupt's study (2009) where

similar results were found using a logistic regression, indicating an overlap in the direction of bias in the two methods used. Item 18 on the other hand favoured the Xhosa first-language group.

Construct equivalence

The following results speak to aim (ii), which assesses the construct equivalence of the VA scale across the two language groups.

Results of the factors with the DIF items included: Table 2 indicates the loadings on factor 1 (higher-order reasoning) and factor 2 (concrete reasoning). The factors were named based on the content of the individual items of the VA scale (following Arendse, 2009). The two factors are distinguished by their high factor loadings and the sufficient number of items loading on a particular factor and the loadings are as to be expected with the easier items loading on factor 2 (concrete reasoning) and the more difficult items loading on factor 1 (higher-order reasoning) for the English language group.

Table 1: Verbal analogies

Item	MH chi-square	df	Significance	Estimate	Direction	Group	MH D-DIF
VA8	16.044	1	0.000	3.596	Reference	English	−3.00
VA9	26.417	1	0.000	5.094	Reference	English	−3.82
VA18	15.095	1	0.000	0.292	Focal	Xhosa	2.89

Table 2: The pattern matrix loadings for the English and Xhosa first-language group

	English first-language group		Xhosa first-language group	
Item	**Higher-order reasoning**	**Concrete reasoning**	**Higher-order reasoning**	**Concrete reasoning**
2	−0.32	**0.58**	**−0.61**	0.36
3	0.09	**0.47**	**−0.40**	**0.57**
4	0.00	**0.79**	0.12	**0.47**
5	−0.34	**0.79**	−0.36	**0.73**
7	0.15	**0.68**	0.08	**0.84**
10	0.28	**0.53**	−0.01	**0.72**
11	0.28	**0.61**	−0.02	**0.90**
12	0.32	**0.46**	0.35	**0.50**
8	**0.42**	0.29	0.24	**0.65**
13	**0.63**	0.26	0.27	**0.77**
14	**0.46**	0.22	0.24	**0.71**
15	**0.44**	0.17	0.37	0.22
16	**0.58**	0.17	0.02	**0.65**
17	**0.64**	0.20	0.04	**0.70**
18	**0.53**	0.11	−0.31	**0.87**
19	**0.81**	−0.07	0.33	**0.58**
20	**0.93**	−0.14	**0.46**	**0.47**
21	**0.82**	−0.22	**0.78**	0.38
22	**0.57**	0.32	0.31	**0.41**
23	**0.94**	−0.01	**0.94**	0.01
24	**0.85**	−0.01	**0.86**	0.16
25	**0.84**	0.07	**0.81**	0.13
26	**0.78**	−0.13	**0.90**	−0.05
27	**0.58**	0.11	**0.69**	0.12
28	**0.82**	0.25	0.74	**0.41**
29	**0.56**	0.18	**0.50**	**0.46**
31	**0.63**	−0.31	**0.57**	0.16
32	**0.62**	0.13	**0.43**	**0.43**
33	**0.68**	−0.28	**1.01**	−0.22

Factor stability is primarily dependent on the sample size and the number of items per factor. In other words, there should be a minimum of at least five observations per item and the factor should have a minimum of three items loading on it (Hair *et al*., 2010). Since the sample size was previously established and there were no items that loaded on both factors simultaneously, as well as three or more items loading on each factor, these factors appear to be stable factors for the English first-language group. High loadings are evident in both the first and second factor for the English language group.

An examination of the factor loadings in the Xhosa language group indicate that there are problematic items with items 3, 20, 28, 29 and 32 simultaneously loading on both factors while item 15 did not load on either factor. The remaining loadings were split with items 2 (factor 1 – higher-order reasoning as opposed to concrete reasoning) and 8, 13, 14, 16, 17, 18, 19 and 22 (factor 2 – concrete reasoning as opposed to higher-order reasoning) loading on different factors than was the case with the English first-language group.

The Tucker's phi coefficient prior to the DIF items being removed indicated non-negligible incongruities (Van de Vijver & Poortinga, 1994) on both factor 1 and factor 2 with values of 0.74 and 0.79 respectively.

Results of the factors with the DIF items excluded: The results indicate distinct loadings on factor 1 and factor 2 for the English language group, similar to results found without the DIF items being removed (Table 3). Loadings are in line with expectations with the easier items loading on factor 2 (concrete reasoning) and the more difficult items loading on factor 1 (higher-order

Table 3: The pattern matrix loadings for the English and Xhosa first-language group with the DIF items removed

Item	English first-language group		Xhosa first-language group	
	Higher-order reasoning	Concrete reasoning	Higher-order reasoning	Concrete reasoning
2	−0.34	**0.58**	−0.37	**0.52**
3	0.10	**0.46**	−0.01	**0.62**
4	0.00	**0.80**	**0.44**	0.29
5	−0.35	**0.82**	0.13	**0.60**
7	0.15	**0.67**	0.63	**0.46**
10	0.29	**0.55**	**0.47**	**0.41**
11	0.27	**0.58**	**0.57**	**0.58**
12	0.32	**0.46**	**0.67**	0.12
13	**0.63**	0.25	**0.77**	0.35
14	**0.47**	0.22	**0.70**	0.30
15	**0.43**	0.14	**0.50**	0.03
16	**0.57**	0.14	**0.45**	**0.43**
17	**0.65**	0.19	**0.50**	**0.51**
19	**0.81**	−0.10	**0.71**	0.30
20	**0.93**	−0.17	0.77	0.10
21	**0.82**	−0.25	**1.01**	−0.10
22	**0.57**	0.30	**0.57**	0.11
23	**0.95**	−0.02	**0.93**	**−0.41**
24	**0.86**	−0.01	**0.94**	−0.29
25	**0.85**	0.06	**0.86**	−0.31
26	**0.79**	−0.14	**0.83**	**0.44**
27	**0.58**	0.09	**0.74**	−0.20
28	**0.81**	0.24	**1.00**	−0.03
29	**0.56**	0.19	**0.79**	0.04
31	**0.66**	−0.31	**0.65**	−0.09
32	**0.61**	0.13	**0.70**	0.09
33	**0.68**	−0.28	0.84	−0.66

reasoning).

The pattern of loading in the Xhosa first-language groups, however, changed when the DIF items were removed. More items, namely 7, 10, 11, 16, 17, 23, 26 and 33, simultaneously loaded on both factors. Only two items (4 and 12 – higher-order reasoning as opposed to concrete reasoning) loaded on a different factor compared to the English group.

Fourteen items (13, 14, 15, 19, 20, 21, 22, 24, 25, 27, 28, 29, 31 and 32) loaded on the same factor, namely factor 1 (higher-order reasoning) as in the English first-language group. The results for factor 2 (concrete reasoning) for the Xhosa first-language group demonstrated that only three items, namely 2, 3, and 5, loaded on this factor compared to the English group.

After the exclusion of the DIF items the Tucker's phi value for the first factor improved to 0.95 and can be regarded as confirming that an identical construct was being measured across the two groups. A value of 0.75 on factor 2 (concrete reasoning) still indicates non-negligible incongruities (Van de Vijver & Poortinga, 1994).

Based on these findings it is evident that only the first factor can be accepted as structurally equivalent, as was also indicated in Arendse's study (2009) across the two language versions of the test, while the second factor continued to display a value not indicative of structural equivalence. However, the fact that so many items in the Xhosa first language cross-loaded on the two factors (they were included in the calculation of the Tucker's phi for the first factor) remains problematic for the factor congruence of factor 1 (higher-order reasoning).

Figure 1(a) below illustrates that the item loadings are fairly closely aligned around the identity line across the two language groups for factor 1 after the DIF items had been removed. This alludes to an indication that factor 1 with the DIF items removed is structurally equivalent, which corroborates the results of the Tucker's phi illustrating a value of 0.95.

Figure 1 (b) below continues to illustrate items that are not closely aligned even after the removal of the DIF items and thus confirms the results of the Tucker's phi (0.75), indicating that the structural equivalence of factor 2 (concrete reasoning) across the English first-language group and the Xhosa first-language group remains problematic even with the removal of the DIF items.

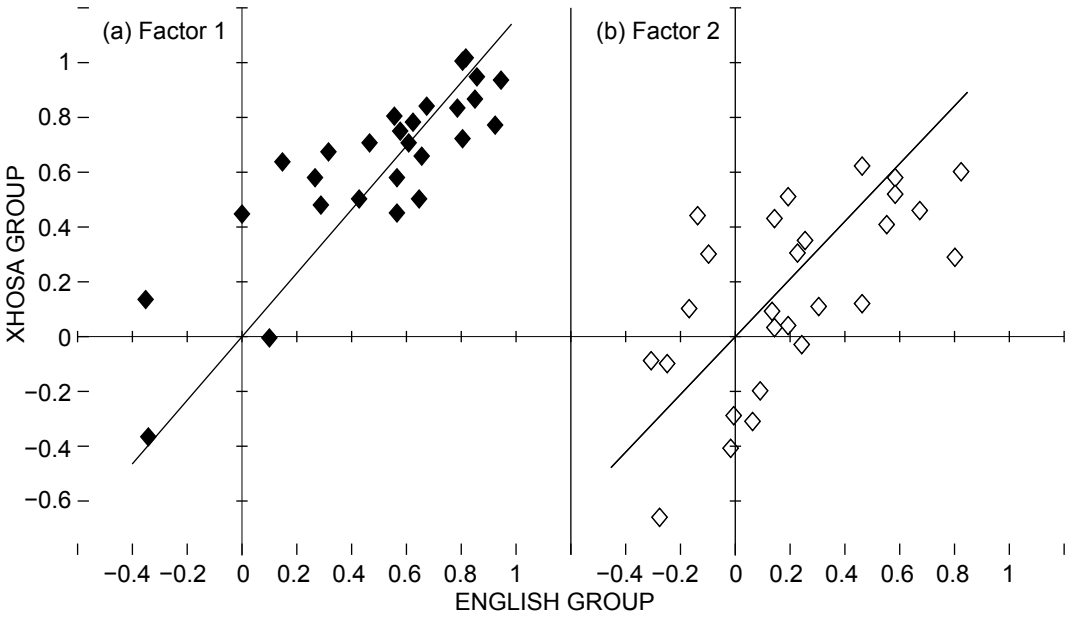

Figure 1: A scatter plot of the factor pattern coefficients for the VA subscale for factor 1 and 2 across the English and Xhosa first-language groups with the DIF items removed

Discussion and conclusion

The Mantel-Haenszel DIF procedure results indicated that the adapted English version of the VA scale displayed differential item functioning (DIF) or bias across the two language groups. Therefore the null hypothesis was rejected for three items on this scale. These three items identified as having DIF were found to corroborate Haupt's study (2009) where similar results were obtained using a different DIF detection technique, namely, logistic regression. Furthermore, the result of this study displayed an overlap in the direction of bias with the two DIF methods used.

Results indicated that items 8 and 9, identified as having DIF, favoured the English first-language group, while item 18 favoured the Xhosa first-language group. What was interesting in these results was that items 8 and 9 that favoured the English first-language group were among the easier items on the VA scale that required concrete reasoning, while item 18 that favoured the Xhosa first-language group, required higher-order reasoning (Arendse, 2009). Item 18 could possibly have favoured the Xhosa first-language group because reasoning on this specific item is based on relational similarity. Relational similarity involves the underlying relations in tasks recognising communalities between different domains in higher-order thinking (Halford, 1996). Halford (1996) regards this knowledge as central to mechanisms that are basic to human reasoning, such as analogy and planning. When relational similarity is used, lower-order thinking is absent or abandoned, which could be an alternative explanation for their lack of performance on items 8 and 9.

However, whatever the reasons for the findings are, DIF points to inequivalence (Van de Vijver & Leung, 1997). A conventional approach in dealing with DIF is to deal with it as a distortion at an item level that should be removed (Van de Vijver & Tanzer, 2004). Therefore, DIF analysis is used in order to identify and remove biased items, using the unbiased items for comparison across groups; in other words, it is assumed that after the removal of DIF items, the scores of the two groups would be comparable. In this study, this assumption was tested.

The results observed from the factor analysis of the English first-language group prior to the DIF items being removed, revealed that two factors were distinguishable by their high factor loadings, with the easier items loading on factor 2 (concrete reasoning) and the more difficult items loading on factor 1 (higher-order reasoning). The Xhosa first-language group, on the other hand, displayed problematic items with certain items cross-loading on factors and other items loading on a different factor in comparison to the English first-language group. These two factors as demonstrated by the Tucker's phi, displayed non-negligible incongruities (Van de Vijver & Poortinga, 1994). These results support the research conducted by Arendse (2009) across the two language versions of the VA scale.

The next step was to evaluate construct equivalence after the DIF items had been removed. The results of the factor analysis again indicated distinct loadings on factor 1 and factor 2 for the English first-language group. Loadings were once again in line with expectations, with the easier items loading on factor 2 (concrete reasoning) and the more difficult items loading on factor 1 (higher-order reasoning). The same pattern of loadings as previously continued for the Xhosa first-language group, though. Items that could be regarded as belonging to the concrete reasoning factor continued to load on the higher-order reasoning factor in the Xhosa first-language group. Thus, some of the easier items (the more 'direct' items) loaded on the more 'indirect' items even after the removal of the DIF items. In other words, for the Xhosa first-language group, because English is not their first language, concrete analogy items became higher-order reasoning analogy items. Even though the Tucker's phi value improved, providing construct equivalence for factor 1, the same could not be said for factor 2 (concrete reasoning).

According to Singer-Freeman and Goswami (2001), analogies become increasingly more difficult if the learner is not familiar with the domain knowledge. The question thus arises, how do Xhosa first-language learners access the appropriate domain knowledge if they lack the language proficiency to understand English instruction in the first place? Poor performance on these items could thus be due to a lack of domain knowledge and not due to inadequate verbal reasoning skills.

The detection and removal of the DIF items for factor 2 did not achieve the desired outcome and, as a result, construct equivalence was not established. Since construct equivalence was not displayed even after the DIF items were removed, differential item functioning is not enough of an explanation for the construct inequivalence found in factor 2. Even though we did not identify a large number of

DIF items, evidence still indicates that two different constructs were being measured.

When a test is biased towards a group, the scores for the group consistently underestimate or overestimate their true values, and could result in a vicious cycle of groups experiencing persistent social prejudice and stereotyping. Thus, until scalar equivalence is established on this scale, it cannot be utilised with confidence as a monolingual language measure for use across different language groups in the South African context.

Recommendations

Since spurious results are a weakness of factor analysis when conducted at an item level (De Bruin, 2004), a Rasch modelling technique is recommended to cross-validate the factor analysis results in order to identify the latent construct with confidence and prevent the identification of spurious factors, in order that full scalar equivalence can be obtained.

Acknowledgements – The authors would like to express their sincere gratitude to Kharnita Mohammed for her invaluable contribution in assisting with the editing of this article.

References

Abrahams F & Mauer KF. 1999. Quantitative and statistical impacts of home language on responses to the items of the Sixteen Personality Questionnaire (16PF) in South Africa. *South African Journal of Psychology* **21**: 76–86.

Allalouf A & Abramzon A. 2008. Constructing better second language assessments based on differential item functioning analysis. *Language Assessment Quarterly* **5**(2): 120–141.

Allalouf A, Hambleton RK & Sireci SG. 1999. Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement* **36**(3): 185–198.

Arendse D. 2009. Evaluating the structural equivalence of the English and isiXhosa versions of the Woodcock Munoz Language Survey on matched sample groups. MA thesis, University of the Western Cape.

Cummins DD. 1992. Role of analogical reasoning in the induction of problem categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **18**(5): 1103–1124.

Cummins J. 2000. *Language, power and pedagogy: Bilingual children in the crossfire.* Clevedon: Multilingual Matters.

De Bruin D. 2009. *Factor analysis. Industrial Psychology Programme.* Department of Human Resource Management. Johannesburg: University of Johannesburg. Unpublished class notes. (Industrial Psychology Programme).

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

Dorans NJ & Holland PW. 1993. DIF detection and description: Mantel-Haenszel and standardization. In Holland PW & Wainer H (ed.) *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum, pp 35–66.

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

Gierl MJ & Khaliq SN. 2001. Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement* **38**: 164–187.

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

Goswami U & Brown AL. 1990. Higher-order structure and relational reasoning: Contrasting analogical and thematic relations. *Cognition* **36***:* 207–226.

Gravetter FJ & Forzano LB. 2008. *Research methods for behavioral sciences*. 3rd edition. Belmont, CA: Thomas Wadsworth.

Hair JF, Anderson RE, Babin B & Black B. 2010. *Multivariate data analysis.* Upper Saddle River, NJ: Pearson.

Halford GS. 1996. *Relational knowledge in higher cognitive processes*. Unpublished paper delivered at the Biennial Meeting of the International Society for the Study of Behavioral Development, Quebec City. August, 12–16.

Hambleton RK & Kanjee A. 1995. Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptation. *European Journal for Psychological Assessment* **11**(3): 147–157.

Haupt GR. 2009. The evaluation of the group differences and item bias of the English version of a standardized test of academic language proficiency for use across English and Xhosa first language speakers. MA thesis, University of the Western Cape.

Holyoak KJ & Thagard P. 1995. *Mental leaps*. Cambridge, MA: MIT Press.

Huysamen GK. 2002. The relevance of new APA standards for educational and psychological testing for employment testing in South Africa. *South African Journal of Psychology* **32**: 26–33.

Iglesias A. 2001. What test should I use? *Seminars in Speech and Language* **22**(1): 3–16.

Ismail G. 2010. Towards establishing the equivalence of the English version of the verbal analogies scale of the Woodcock Muñoz Language Survey (WMLS), across English and Xhosa first language speakers. MA thesis, University of the Western Cape.

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

Koch SE. 2005. Evaluating the equivalence, across language groups, of a reading comprehension test used for admission purposes. PhD thesis, University of Port Elizabeth.

Koch E. 2007. The monolingual testing of competence: Acceptable practice or unfair exclusion. In Cuvelier P, Du Plessis T, Meeuwis M & Teck L (eds) *Multilingual and exclusion. Policy, practice and prospects.* Pretoria: Van Schaik, pp 79–103.

Koch E & Dornbrack J. 2008. The use of language criteria for admission to higher education in South Africa: Issues of bias and fairness investigated. *Southern African Linguistics and Applied Language Studies* **26**(3): 333–350.

Kubinger KD. 2003. On artificial results due to using factor analysis for dichotomous variables. *Psychology Science* **45**(1) : 106–110.

Oakland T. 2004. Use of educational and psychological tests internationally. *Applied Psychology: An International Review* **53**(2): 157–172.

Paterson H & Uys K. 2005. Critical issues in psychological test use in the South African workplace. *SA Journal of Industrial Psychology* **31**(3): 12–22.

Peña ED. 2001. Assessment of semantic knowledge: Use of feedback and clinical interviewing. *Seminars in Speech and Language* **22**(1): 51–94.

Primrose AF, Fuller M & Littledyke M. 2000. Verbal reasoning test scores and their stability over time. *Educational Research* **42**(2): 167–174.

Rossier J. 2004. An analysis of the cross-cultural equivalence of some frequently used personality inventories. In *International perspectives on career development.* Symposium conducted at a joint meeting of the International Association for Educational and Vocational Guidance and the National Career Development Association, San Francisco.

Singer-Freeman KE & Goswami U. 2001. Does half a pizza equal half a box of chocolates? Proportional matching in an analogy task. *Cognitive Development* **16**: 811–829.

Sireci SG & Allalouf A. 2003. Appraising item equivalence across multiple languages and cultures. *Language Testing* **20**(2): 148–166.

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

Sireci SG, Patsula L & Hambleton RK. 2005. Statistical methods for identifying flaws in the test adaptation process. In Hambleton RK, Merenda PF & Spielberger CD (eds) *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, New Jersey: Lawrence Erlbaum Associates Inc., pp 93–116.

Spearman C. 1927. *The abilities of man*. New York: Macmillan.

Sternberg RJ. 1977. Component processes in analogical reasoning. *Psychological Review* **84**: 353–378.

Taliep N. 2010. Evaluating the construct validity of the KIDSCREEN-52 quality of life questionnaire within a South African context utilising exploratory factor analysis: Initial validation. MA thesis, University of the Western Cape.

Tagalakis G & Keane M. 2006. Familiarity and relational preference in the understanding of noun–noun compounds. *Memory & Cognition* **34**: 1285–1297.

Van der Oord S, Van der Meulen EM, Prins PJM, Oosterlaan J, Buitelaar JK & Emmelkamp PMG. 2005. A psychometric evaluation of the social skills rating system in children with attention deficit hyperactivity disorder. *Behaviour Research and Therapy* **43**: 733–746.

Van de Vijver FJR. 1998. Towards a theory of bias and equivalence. In Harkness JA (ed.) *Cross-cultural survey equivalence*. ZUMA-Nachrichten Spezial, nr. 3 Manheim: ZUMA (Zentrum Für UmFragten, Methoden und Analysen). pp 41–62.

Van de Vijver FJR & Leung K. 1997. *Methods and data analysis for cross-cultural research.* Thousand Oaks: Sage.

Van de Vijver FJR & Poortinga YH. 2002. Structural equivalence in multilevel data. *Journal of Cross-Cultural Psychology* **33**(2): 141–156.

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

Van de Vijver F & Tanzer NK. 2004. Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology* **54**: 119–135.

Whetton C. 1985. Verbal reasoning tests. In Hussen T & Postleth-Waite N (ed.) *The international encyclopedia of education*. Oxford: Pergamon, pp 5431–5433.

Woodcock RW & Muñoz-Sandol AF. 2001. *Woodcock-Muñoz Language Survey: Normative update*. Itasca: Riverside Publishing Company.

Zumbo D, Sireci G & Hambleton K. 2003. *Re-visiting exploratory methods for construct comparability: Is there something to be gained from the ways of old?* Chicago: National Council of Measurement in Education.