Southern African Linguistics and Applied Language Studies 2012, 30(1): 65–76
Printed in South Africa — All rights reserved

SOUTHERN AFRICAN LINGUISTICS
AND APPLIED LANGUAGE STUDIES
ISSN 1607-3614 EISSN 1727-9461
http://dx.doi.org/ 10.2989/16073614.2012.693715

The argument for evaluating monolingual language tests for equivalence across language groups

Genevieve Haupt[1]* and Elize Koch[2]
[1]*HSRC, 38 Heron Cove, Gie Road, Table View 7441, Cape Town, South Africa*
[2]*University of the Western Cape, PO Box 15243, Emerald Hill, 6011*
** Corresponding author, e-mail: ghaupt@hsrc.ac.za*

Abstract: The demonstration of scalar equivalence in language proficiency tests (which can be viewed as monolingual language tests) has often been deemed as unnecessary as it is argued that the biases associated with the language of a test (used across multilingual language groups) will not occur. However it is increasingly acknowledged that scalar equivalence is as important in monolingual language tests as it in multilingual language tests. This paper will provide empirical support for the argument that the meaning of tests scores across groups (scalar equivalence) is as important in monolingual language proficiency testing as it is in any other cross-linguistic testing. The authors will present research conducted on the equivalence of an adapted English version of a standardised academic language proficiency test (Woodcock-Muñoz Language Survey, WMLS, 2001), with its intended use being across English-first-language speakers and isiXhosa-first-language speakers. More specifically, the focus will be on an item bias analysis across the English- and isiXhosa-first-language speakers for all the sub-tests of the adapted English version of the WMLS.

Preamble

Cross-linguistic testing refers to the testing of individuals from different language groups. When testing different groups, tests can be administered in various languages (multilingual tests), or be available only in one language and used across language groups (monolingual tests). When using cross-linguistic tests, various researchers have confirmed the importance of ensuring that the instruments should demonstrate the highest form of equivalence, namely scalar equivalence (Van de Vijver & Leung, 1997). In scalar equivalence, the primary concern is that the scores obtained on tests must have the same meaning across groups (Van de Vijver & Leung, 1997). Scalar equivalence is evaluated by assessing bias within tests.

According to Van de Vijver and Leung (1997), it is generally accepted when using multilingual tests that evidence of scalar equivalence must be provided. However, this importance has not been recognised when using tests of proficiency in a specific language, for example, proficiency in English. The argument often put forward is that, when an instrument is used for the purpose of assessing language proficiency, the possible biasing effect of language on the scores of some groups should not be an issue, as language is what the test is attempting to measure (Huysamen, 2002). Nonetheless, some researchers have indicated in recent years that scalar equivalence is a central issue even in language proficiency tests that are used across languages groups (such as Koch, 2005, 2007, 2009).

The purpose of this paper is to demonstrate empirically that the meaning of tests scores across groups (scalar equivalence) is as important in monolingual language proficiency testing as it is in any other cross-linguistic testing. In order to demonstrate this, the authors will present research conducted on the equivalence of an adapted English version of a standardised academic language proficiency test (Woodcock-Muñoz Language Survey, WMLS, 2001), with its intended use for English-first-language speakers and isiXhosa-first-language speakers. This research was conducted as a sub-study of a larger study where the WMLS was adapted into South African English and isiXhosa to evaluate a bilingual education project (Koch, 2009). The validity of both language versions of the test for the South African context is currently being evaluated and this paper reports

Southern African Linguistics and Applied Language Studies is co-published by NISC (Pty) Ltd and Routledge, Taylor & Francis Group

on the research conducted on the English version of the test. More specifically, the focus will be on an item bias analysis across the English- and isiXhosa-first-language speakers for all the sub-tests of the adapted English version of the WMLS. Equivalence and bias will be explained comprehensively in the next section, followed by a review of some of the literature on cross-linguistic testing.

Equivalence and bias

As Van de Vijver and his co-researchers (Van de Vijver & Leung, 1997; Van de Vijver & Poortinga, 1997; Van de Vijver & Tanzer, 1998) did significant theoretical and empirical work in the area of measurement equivalence in cross cultural measurement, this article draws mainly from the work of these theorists. Their views are also supported by the adoption of similar positions on bias in the International Testing Commission's Guidelines for the Adaptation of Tests (ITC, 2001). These researchers regard the attainment of equivalent measures as perhaps the central issue in cross-cultural and/or cross-linguistic comparative research (Van de Vijver & Leung, 1997). They argue that individuals with the same or similar standing on a construct, such as learners with high science ability but belonging to different groups, such as English- and isiXhosa-speaking groups, should obtain the same or similar scores on a test for a measure to be equivalent. Should this not be the case, the items are said to be biased and the two versions of the measure are non-equivalent. Furthermore, should the basis of comparison not be equivalent across different measures (in the case of multilingual tests) or different groups (in the case of monolingual tests), valid comparisons across these measures and/or groups cannot be made, as the test scores are not directly comparable.

There are various levels of equivalence, namely construct (structural) equivalence, measurement unit equivalence and scalar equivalence. Construct equivalence exists when the same construct is being measured across the various groups. In order to ensure that construct equivalence is achieved, the nomological networks of the instrument should be investigated in each culture (Van de Vijver & Leung, 1997: 8). The next level of equivalence is measurement unit equivalence. This level of equivalence indicates that a test that has been adapted for use in a specific culture should continue to have the same units of measurement as the original version of the test (Van de Vijver & Rothman, 2003). For example, the two forms (the original and the adapted version) of a test must continue to yield judgment that follows an interval scale and, in addition, they must be on the same interval scale (Van de Vijver & Rothman, 2003). The last and highest form of equivalence is scalar equivalence, which can be achieved when the instrument has the same measurement unit, the same construct is being measured, and the measurement scale has the same origin (Van de Vijver & Leung, 1997).

There are three types of bias that play a vital role in cross-cultural and/or cross-linguistic comparisons, namely construct bias, method bias and item bias. Van de Vijver and his colleagues have made a clear theoretical link between equivalence and bias. This theoretical framework asserts that, should bias be found in a measure, equivalence cannot exist (Van de Vijver & Leung, 1997; Van de Vijver & Poortinga, 1997). Construct bias occurs when the construct being measured is not identical across various groups (see Van de Vijver and Leung's (1997) studies on conceptions of intelligence). The second type of bias, method bias, occurs as a result of characteristics of the instrument and/or its administration (Van de Vijver & Leung, 1997). According to Van de Vijver and Leung (1997: 13), differential response styles across groups, such as agreement and extremity ratings, can represent method bias.

The last type of bias, item bias refers to nuisance factors at an item level. According to Van de Vijver and Leung (1997), the term item bias has largely been replaced by the term differential item functioning (DIF). However, Van de Vijver and Leung (1997: 18) prefer the term 'item bias' as opposed to 'DIF'. They warn that the issue of item bias is a measurement problem, and that biased items, if not handled properly, will affect the measurement equivalence of a test, in other words, score comparability. As a result of this theoretical departure point, they are of the opinion that biased items should not be retained in a test, even when deemed 'construct relevant', when the scores of different groups are used for comparison (Van de Vijver & Leung, 1987: 85). These recommendations will be elaborated on in the recommendations section of the article. Item bias occurs when people of the same group (language, gender, ethnicity, and so on) with the same latent trait (the same ability or skill, for example) have a different probability of giving a certain response on a measure (Van de Vijver & Leung, 1997).

With reference to item bias, a distinction is made between uniform and non-uniform DIF. Uniform bias refers to influences of bias on sources that are more or less the same for all score levels, and non-uniform bias refers to the influences that are not identical for all score levels (Van de Vijver & Leung, 1997). The introduction of uniform bias will lead to the loss of scalar equivalence. When a constant is added to all scores in one group but not in the other, as is the case in uniform item bias, the differences in scores between the groups no longer have a natural or common origin, therefore uniform bias will not affect measurement unit equivalence (Van de Vijver & Leung, 1997). On the other hand, should non-uniform item bias arise, it will destroy equivalence to a significant extent because the measurement units in the two groups are no longer the same (Van de Vijver & Leung, 1997). Therefore, when several items show this kind of bias, cross-cultural and/or cross-linguistic score comparisons are likely to generate incorrect results (Van de Vijver & Leung, 1997).

Cross-linguistic testing

As stated previously, cross-linguistic testing refers to the testing of individuals from different language groups where the tests are available in various languages (multilingual) or in one language only and used across groups (monolingual).

Multilingual testing

South Africa is a multilingual society and to this end various researchers have illustrated that it is pertinent that tests be available in more than one language. Some of the reasons contributing to this awareness of the need for multilingual test development include developing 'culture-reduced' or 'culture-common' tests, as another issue facing test developers is that most of the available measures were developed in the United States of America or the United Kingdom. As a result, the measures that have been developed internationally have a predisposition to be more suitable for westernised, English-speaking individuals (Foxcroft & Roodt, 2009). According to Foxcroft and Roodt (2009), the focus of psychological testing during the 1980s and 1990s was thus on cross-cultural test adaptation. Test adaptation refers to the process of making a test more applicable to a specific context while using the same language (Foxcroft & Roodt, 2009). The focus on test adaptation had come about largely due to the need for tests to be more culturally appropriate, and because many measures were largely available in only one language (Foxcroft & Roodt, 2009). The International Test Commission (ITC) accordingly released their Guidelines for Adapting Educational and Psychological Tests in 2001 (ITC, 2001). These guidelines have been at the forefront of all cross-cultural test adaptation around the world and assist in advocating against assessment practices in which test-takers are tested in languages they are not proficient in, or in which a translator is sometimes used to translate the test. In addition, various methodologies and statistical techniques have been developed to aid in establishing whether different language versions of a test (multilingual tests), or tests that are only available in one language and used across groups (monolingual tests), are equivalent.

Monolingual testing and language testing

Monolingual testing refers to the use of assessment measures that are available only in one language across two or more language groups (Koch, 2007). Many countries, including South Africa, use monolingual tests to measure individuals on a particular trait. An example of a type of test used in the South African context is admissions tests for the purpose of selection for higher education (Koch, 2007). When a test is used for admission purposes, it is often based on achievement and aptitude measures, and not on an assessment of an individual's proficiency in a specific language. It is important to note that, unless monolingual tests are used to assess proficiency in a specific language, cross-cultural research would indicate that the tests should be made available in more of the languages of the given population. Huysamen (2002) asserts that, if a test is not intended as a measure of language proficiency and is used with testees who are not proficient in the language of the test, it is likely that construct irrelevance will occur. However, Koch (2007) argues, and demonstrates empirically, that even if tests are used to assess an individual's proficiency in a specific language (such as reading in English), construct irrelevance can still occur due to influences related to construct irrelevant factors. For example, first- and second-language speakers may have different reading processes and the construct (English reading comprehension) that is measured may be more relevant to a first-language speaker than to a second-language speaker. Hence it becomes imperative that all tests, both multilingual and monolingual, including monolingual language tests, are evaluated for equivalence (Koch, 2007).

Rationale and research aims

The significance of this particular study is two-fold. Firstly, the study evaluates the scalar equivalence of the English version of the WMLS to be used across English-first-language group and isiXhosa-first-language group (through this rigorous evaluation process of both the larger study and the sub-study). Secondly, the study serves to create awareness that the same importance should be given to the evaluation of the equivalence, across language groups, of a monolingual language test as is given to multilingual language tests that are intended to be used across multilingual or bilingual language groups. This study will demonstrate that there is a need for studies of this nature on monolingual language tests in general, and thereby will support previous studies on this topic (such as Koch, 2007, 2009) and contribute to the empirical evidence supporting theoretical work on the issue of equivalence in monolingual tests and language tests.

The overall aim of the study therefore was to evaluate the equivalence of the adapted English version of the WMLS (2001) for use across English- and isiXhosa-first-language speakers. In order to achieve the overall aim, group differences on item characteristics and psychometric properties, and item bias, were evaluated. The researcher thus outlines two specific research objectives:

(i) to evaluate the group differences between the English-first-language and isiXhosa-first-language groups on the adapted English version of the WMLS in terms of:
(a) mean scores
(b) reliability
(c) mean item characteristics
(ii) to evaluate item bias across the various sub-tests of the adapted English version of the WMLS across the English- and isiXhosa-first-language groups.

Methodology

The study can be regarded as a comparative quantitative study with secondary data. The data was analysed by means of descriptive and inferential statistics. The study is situated in a larger study, for which the WMLS was adapted into South African English and into isiXhosa to evaluate a bilingual education project (Koch, 2009). The purpose of the larger study is to evaluate the psychometric properties of the adapted English and isiXhosa versions of the tests, and the data was collected as part of this bigger project.

Sample

The sampling technique used was convenience-purposive sampling. This technique of sampling allowed the researcher to ensure homogeneity in terms of equal numbers of males and females, as well as equality of educational background. The sample comprised 198 English first-language speakers, of which 98 were males and 99 were females. Of these, 82 were Grade 6 learners and 110 were Grade 7 learners. The sample also included 197 isiXhosa first-language speakers, of which 76 were males and 116 were females, and 98 were Grade 6 and 99 were Grade 7 learners.

Measurement tool

The measurement tool of interest is the adapted English version of the Woodcock-Muñoz Language Survey (WMLS). This measure is an individually administered test that takes approximately 40 minutes to administer and consists of four sub-tests, namely Picture Vocabulary, Verbal Analogies, Letter-word Recognition and Dictation (Woodcock & Muñoz-Sandoval, 2001). The content of each sub-test was selected to represent important skills needed for evaluating language proficiency in a diverse population, covering a broad range of development (from age two to adulthood). The adaptation of the test from English into isiXhosa used the guidelines of the ITC (2001) as framework. The adaptation of the English version was restricted to the inclusion of a few additional, correct options in some of the sub-scales, for example 'Nike' as a correct response in addition to 'running shoe' in the Vocabulary sub-scale. The adaptation into isiXhosa will not be discussed in detail in this article, and the interested reader is referred to Koch (2009). Table 1 is a summary of each sub-test in terms of content, prompts and responses.

Psychometric properties of the WMLS

The median reliabilities range from .80 to .93 for the tests and from .88 to .96 for the clusters (Woodcock & Muñoz-Sandoval, 2001). The validity of the WMLS (the American version) was evaluated on content and concurrent, as well as construct, validity (Woodcock & Muñoz-Sandoval, 2001) for an English-speaking American population. Correlations have been done among the WMLS Normative Update tests and clusters at selected age levels, and most of these tests showed inter-correlations at a moderate level of 0.4 or above (Woodcock & Muñoz-Sandoval, 2001).

The abovementioned reliability and validity are based on the original American versions of the WMLS. No psychometric properties are available for the adapted South African versions (English and isiXhosa versions) of the WMLS. This study is a contribution to the research on the English version for the South African context.

Data analysis

The overall aim of the study was to evaluate the equivalence of the adapted English version of the WMLS across two language groups, namely English- and isiXhosa-first-language speakers using the Statistical Program for the Social Sciences (SPSS), version 16. The analysis will be discussed per research aim.

Evaluating group differences. Research objective 1a was had been analysed by means of descriptive statistics, using mean and standard deviations per sub-test for each group. The Hotelling's T^2 statistic and the post hoc *t*-tests were used to determine whether the differences displayed were significant (if any were detected). The null hypothesis was that there are no group differences between the English- and isiXhosa-first-language groups with regard to their mean scores on the adapted English version of the WMLS.

Table 1: Summary of each sub-test in terms of content, prompts and responses

Sub-test	Linguistic and Curriculum areas	Stimilu	Test Requirement	Response
Picture Vocabulary (PV)	-Oral expression -Language development -Expressive vocabulary	Visual (pictures)	Identifying objects	Oral (word) Total = 57
Verbal Analogies (VA)	-Receptive-expression -Vocabulary	Auditory (phrases)	Stating a word to complete an analogy	Oral (word) Total = 35
Letter-word Recognition (LWR)	-Reading -Reading-decoding	Visual (text)	Identifying printed letters and words	Oral (letter name, word) Total = 57
Dictation (Dict)	-Spelling, writing -Language development -English usage	Auditory (words)	Spelling orally presented Pre-writing and writing skills	Motor (writing) Total = 56

Research objective 1b entailed evaluating group differences by determining the reliability of each sub-test for each language group. This was done firstly by calculating the Cronbach's alpha for each language group on each sub-test. In addition, the equality of reliability was calculated using the statistic (1-alpha1)/(1-alpha2), following the approach proposed by Van de Vijver and Leung (1997). When calculating the equality of reliability, the critical value of 1.26 was used. The difference follows an F distribution, with N1-1 and N2-1 degrees of freedom. The null hypothesis was that there are no significant group differences with regard to the Cronbach's alpha for each of the sub-tests between the English- and isiXhosa-first-language groups on the adapted English version of the WMLS.

In terms of research objective 1c, the item characteristics of the various sub-tests for both language groups were compared descriptively. The mean item difficulty and the mean item discrimination of each group were calculated per sub-scale. Item difficulty refers to the proportion of individuals who answer an item correctly (Foxcroft & Roodt, 2001). Therefore, the higher the proportion of correct responses, the easier an item, and the lower the proportion of correct responses the more difficult an item. According to Foxcroft and Roodt (2009), item difficulties that have a mean of around 0.5 and values ranging between 0.30 and 0.70 give a reasonable indication of the differences between examinees. Item difficulty and item discrimination are closely related; that is, the difficulty level of an item restricts the discriminatory power of an item (Foxcroft & Roodt, 2001). Item-total correlations or item discrimination refers to the potential of a good item to discriminate well between the low and high achievers on a particular test. According to Foxcroft and Roodt (2001), in general, item-total correlations of about 0.20 are considered to be the minimum acceptable discrimination value for item selection purposes. No null hypothesis was tested for this specific research objective.

The second research objective was to evaluate item bias[1] by means of the statistical procedure of logistic regression (see Formula 1).

$$P = (u = 1 \left| \theta, g) = \frac{e^{\tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g)}}{e^{\tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g)}} \right.$$

Formula 1: Logistic regression

According to Swaminathon and Rogers (1993), logistic regression calculates the probability of a correct response to an item on a test. Three steps are entered when using logistic regression (stepwise analysis): Step 1, the total score on the sub-test as the conditioning variable; Step 2, the group membership is added to the analysis; Step 3, the interaction between the group membership and the conditioning variable (total score on the sub-test). Model fit is assessed by evaluating the differences in chi-square (DIFF Chi Square) between Step 1, Step 2, and Step 3 at two degrees of freedom. Due to the repeated nature of the analyses, and to control for increased Type 1 error, the more stringent criterion of 0.01 was used. The critical value (at two degrees of freedom) of 9.55 was used. The next step was to evaluate the effect size using R^2 differences (Nagelkerke) between Step 1 and Step 3. The effect size was categorised into three dimensions:

(i) Negligible DIF: $R^2\Delta < 0.35$
(ii) Moderate DIF: $0.035 < R^2\Delta \leq 0.060$
(iii) Large DIF: $R^2\Delta > 0.060$.

The final step with regard to logistic regression is to determine whether uniform and non-uniform DIF exists. The determination of uniform and non-uniform bias was done by evaluating differences between R^2 from Step 1 to Step 2 (which indicates uniform DIF) and from Step 2 to Step 3 (which indicates non-uniform DIF). Once uniform or non-uniform DIF had been identified, the next and final step was to determine the direction of these items. Uniform items where the beta value (ß) is less than zero (ß < 0) are said to favour the reference group (the reference group in this study is the English-first-language group), while uniform items, where the beta value (ß) is greater than zero (ß > 0), are said to favour the focal group (in this study the focal group is the isiXhosa-first-language group). A non-uniform DIF (in terms of the total score) favours the high ability reference group and the low ability focal group when ß < 0, and the high ability focal group and low ability reference group when ß > 0. The null hypothesis was: the probability of scoring 1 on item (i) for all the sub-tests will be the function of ability only, in other words, there is no item bias across English- and isiXhosa-first-language learners on the sub-tests of the WMLS (it is important to note that the null hypothesis of no-DIF was rejected only for moderate and large DIF effect sizes).

Results

Evaluating group differences in terms of mean scores

Figure 1 indicates that the overall performance of the isiXhosa-first-language group was lower than the English-first-language group (with standard deviations indicated in parentheses).

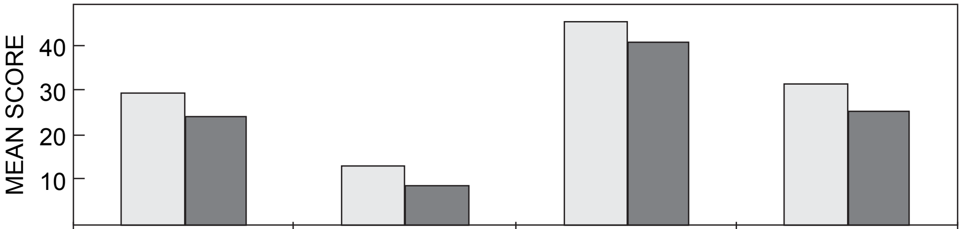
This representation clearly indicates that the isiXhosa-first-language group displayed an overall lower mean score than the English-first-language group. In addition, the standard deviations for both groups were relatively small, indicating that the scores were clustered around the mean score; however, the isiXhosa-first-language group displayed a higher standard deviation on the Letter-word Identification and Dictation sub-tests than the English-first-language group.

The Hotelling's T^2 test [T^2(case-wise MD) = 0.554; F(0.55) = 53.217^a, $p < 0.00$] indicated that the overall difference between the two groups was significant ($p < 0.05$). The post hoc *t*-test demonstrates that the differences displayed between the two language groups are significant on all of the sub-tests: Picture Vocabulary $t = 205.91$ ($p < 0.00$), Verbal Analogies $t = 106.53$ ($p < 0.00$), Letter-word Identification $t = 44.75$ ($p < 0.00$) and Dictation $t = 105.08$ ($p < 0.00$). The null hypothesis was thus rejected.

Evaluating group differences in terms of reliability

According to Anastasi and Urbina (1997), standardised measures should have reliabilities in the range of 0.80s to 0.90s in order to classify them as satisfactory.

The trend, as shown in Table 2, was that the isiXhosa-first-language group displayed an overall higher reliability than the English-first-language group on the various sub-tests of the WMLS. Furthermore, the reliability coefficients indicated good internal consistency levels for both language groups (above 0.80). However, this was not the case for the English-first-language group on the Picture Vocabulary sub-test (0.73). The results of the equality of reliability indicated that there were significant differences between the English-first-language group and isiXhosa-first-language group



	PV	VA	LWI	DICT
□ English	30.11 (3.31)	13.66 (4.56)	46.46 (5.80)	32.31 (4.64)
■ isiXhosa	24.67 (4.12)	8.94 (4.47)	41.69 (8.06)	26.07 (7.10)

Figure 1: Overall performance mean score for English- and isiXhosa-first-language groups

in terms of their reliability on the Picture Vocabulary (1.46, < 0.05) and Dictation (2.05, < 0.05) sub-tests. There were no significant differences between the two language groups in terms of reliability on the remaining sub-tests (Verbal Analogies and Letter-word Identification sub-tests). Therefore the null hypothesis tested for this objective was rejected for the Picture Vocabulary and Dictation sub-tests. In both cases, the isiXhosa-first-language group displayed higher alpha levels than the English-first-language group. The null hypothesis is not rejected for the Verbal Analogies and Letter-word Identification sub-tests.

Evaluating group differences in terms of mean item characteristics

Table 3 indicates the mean item difficulty values and standard deviations (SD) – presented in parentheses – for the two groups per sub-test, as well as the mean item-total correlations.

The trend is that the items of the various sub-tests were more difficult for the isiXhosa-first-language group than for the English-first-language group.

The standard deviation values for item difficulty displayed on Picture Vocabulary and Letter-word Identification were smaller for the English-first-language group than for the isiXhosa-first-language group (standard deviation values are the same for both English and isiXhosa-first-language groups for the Dictation sub-test). On the remaining sub-test (i.e. the Verbal Analogies sub-test), the isiXhosa-first-language group displayed a lower variance than the English-first-language group.

The mean item-total correlations (item discrimination) for Picture Vocabulary and Verbal Analogies were similar for both language groups; however, the mean item-total correlations were different between the language groups with regard to Letter-word Identification and Dictation. The mean item-total correlations for Picture Vocabulary for both groups were lower than 0.20 and therefore below the acceptable range. With regard to the English-first-language group, the mean item-total correlation of the Dictation sub-test indicates that the items of this sub-test generally do not discriminate well between the high and low achievers in this language group (mean item-total correlation, 0.19). However, the isiXhosa-first-language group displayed an acceptable mean item-total correlation (0.30).

Item bias (differential item functioning, DIF)

The Picture Vocabulary sub-test displayed eight items with DIF, of which seven items displayed large DIF (range of $R^2\Delta = 0.07$ and 0.38) and one item displayed moderate DIF ($R^2\Delta = 0.04$). Three of the large DIF items displayed uniform DIF, two of which favoured the isiXhosa-first-language group, and the remaining item favoured the English-first-language group. The remaining four items with large DIF displayed non-uniform DIF, of which three items favoured the high ability English-first-

Table 2: Reliability of the English and isiXhosa first-language-speaking group for each sub-test

Language Group	Picture Vocabulary	Cronbach's Alpha Verbal Analogies	Letter-word Identification	Dictation
English	0.73	0.83	0.89	0.82
IsiXhosa	0.81	0.86	0.92	0.91

Table 3: Mean item characteristics across the two language groups for each sub-test

Language Groups	Picture Vocabulary		Verbal Analogies		Letter-word Identification		Dictation	
	Item Diff. (SD)	Item Disc.	Item Diff. (SD)	Item Disc.	Item Diff. (SD)	Item Disc.	Item Diff. (SD)	Item Disc.
English	0.52 (0.14)	0.13	0.39 (0.31)	0.32	0.82 (0.20)	0.26	0.58 (0.20)	0.19
IsiXhosa	0.43 (0.25)	0.15	0.25 (0.26)	0.34	0.73 (0.26)	0.35	0.46 (0.20)	0.30

language group and the low ability isiXhosa-first-language group, while the remaining item favoured the high ability isiXhosa-first-language group and the low ability English-first-language group.

With regard to the Verbal Analogies sub-test, four items displayed DIF, with one item having a large DIF ($R^2\Delta = 0.08$) and three items having moderate DIF (range of $R^2\Delta = 0.04$ and 0.06). The large item displayed uniform DIF, which favoured the English-first-language group. The remaining three moderate items displayed non-uniform DIF, which favoured the high ability isiXhosa-first-language group and the low ability English-first-language group.

The Letter-word Identification sub-test displayed 20 DIF items, 13 of which were large DIF (range of $R^2\Delta = 0.07$ and 0.19) items and seven of which were moderate DIF items (range of $R^2\Delta = 0.04$ and 0.06). Eleven of the large DIF items displayed uniform DIF, five of which favoured the isiXhosa-first-language group and the remaining six items which favoured the English-first-language group. The two remaining large DIF items displayed non-uniform DIF, which favoured the high ability English-first-language group and the low ability isiXhosa-first-language group. With regard to the seven moderate DIF items, five displayed uniform DIF, of which three favoured the English-first-language group and two favoured the isiXhosa-first-language group, while the remaining two moderate items displayed non-uniform DIF, which favoured the high ability English-first-language group and the low ability isiXhosa-first-language group.

The final sub-test, Dictation, displayed four DIF items of which two displayed large DIF ($R^2\Delta = 0.08$; 0.08) and two displayed moderate DIF ($R^2\Delta = 0.05$; 0.05). The two large items displayed uniform DIF, which favoured the English-first-language group. One of the moderate items displayed uniform DIF, which favoured the English-first-language group, and the remaining item displayed non-uniform DIF, which favoured the high ability English-first-language group and the low ability isiXhosa-first-language group.

Table 4 is a summary table representing the results of the logistic regression analysis. This table presents the large and moderate DIF items found, whether they are uniform or non-uniform, as well as which groups these items favour.

From the above summary table it is clear that item bias (or DIF) exists for all of the sub-tests. In the light of these findings, the null hypothesis is rejected for all of the sub-tests.

Discussion

With regard to the group differences in terms of reliability statistics on the various sub-tests, it was illustrated that the various sub-tests were easier for the English-first-language group than they were for the isiXhosa-first-language group. This is not in itself a problematic finding, as it may be because of real differences in proficiency. However, there also were significant differences in reliability across the groups. Various reasons can be assigned to these differences, such as ‘the nature of the group’ (Anastasi & Urbina, 1997), variability in the scores of the sample (Anastasi & Urbina, 1997), ‘floor and ceiling effects’ (Allen & Yen, 1979; Van der Vijver & Leung, 1997) and more. With regard to the results obtained in this study, the most likely contributing factor to the differences in reliability is the variability in the scores of the sample groups. For example, the lower reliability coefficient of the English-first-language group than the isiXhosa first language speaking group in relation to the Picture Vocabulary sub-test most probably have been the result of the lower standard deviation, and thus variability of the scores, found in the English-first-language group.

In terms of item characteristics it was found that the English-first-language group displayed a higher mean item difficulty overall compared to the isiXhosa-first-language group. Further analysis determined that a number of these differences were due to item bias, as it was found with the DIF analysis that all of the sub-tests displayed DIF.

According to Van de Vijver and Leung (1997), item bias can be produced by various sources. When investigating the possible reasons for DIF on the Picture Vocabulary sub-test, for example, it was found that some of the images or pictures may have been less well known in the isiXhosa-first-language group. An example of this could be Item 29 (an item that displayed large uniform DIF that favoured the English-first-language group), which contains an image of an igloo. On the other hand, some uniform items (Items 15 and 30, images of sneakers and a theatre respectively) favoured the isiXhosa-first-language group. These finding are hard to explain, as is often the case in DIF studies.

Table 4: Summary results of Logistic Regression method

Sub-scale	Effect Size	Type of DIF	Direction of DIF favour	Number of Items	Items
Picture Vocabulary	Large DIF	Uniform	English	1	29
			IsiXhosa	2	15, 30
		Non-uniform	HA English LA isiXhosa	3	26, 32, 34
			HA isiXhosa LA English	1	23
	Moderate DIF	Uniform	English	1	27
			IsiXhosa	0	
Verbal Analogies	Large DIF	Uniform	English	2	8, 9
			IsiXhosa	0	
		Non-uniform	HA English LA isiXhosa	0	
			HA isiXhosa LA English	2	5, 18
Letter-word Identification	Large DIF	Uniform	English	7	22, 25, 31, 40, 42, 50, 54
			IsiXhosa	5	38, 39, 49, 50, 55
		Non-uniform	HA English LA isiXhosa	2	48, 53
			HA isiXhosa LA English	0	
	Moderate DIF	Uniform	English	3	28, 30, 37
			IsiXhosa	2	45, 51
		Non-uniform	HA English LA isiXhosa	2	46, 52
			HA isiXhosa LA English	0	
Dictation	Large DIF	Uniform	English	2	10, 22
			IsiXhosa	0	
	Moderate DIF	Uniform	English	1	19
			IsiXhosa	0	
		Non-uniform	HA English LA isiXhosa	1	32
			HA isiXhosa LA English	0	

In the Letter-word Identification sub-test, several items displayed DIF for both language groups. With regard to the DIF items that favoured the English-first-language group, it could be argued that many of the words used in this sub-test (where each word needed to be pronounced aloud) were more familiar to the English-first-language group, while the phonology of this language and the phonics of its written system are also more familiar to this group. An example of one of the items is Item 31, which was the word ‘island’. It is clear, though, that this was not the case with all the difficult words in the test; this lack of familiarity with words in the isiXhosa group was limited to some words. With regard to the final sub-test, Dictation, all of the large uniform DIF items favoured the English-first-language group. An example of this is Item 10, which required that the testee write down the letter 'Y' as a capital letter. Again, it is difficult to provide the reason for this finding.

Conclusion

It is evident that it is not going to be possible to provide clear reasons for the finding of DIF in most of these sub-tests. The most important aspect of these findings is the fact that there is extensive evidence of bias in all of these sub-tests. In terms of the theoretical position adopted by the authors in this paper, and as supported by seminal work in the area of bias and equivalence (see the section where this is discussed), we accept this to be evidence of score incomparability between the two language groups in relation to an English monolingual language test. Both the origin and the unit of measurement of the sub-tests were affected by the presence of the biased items as both uniform and non-uniform DIF were found.

Van de Vijver and Leung (1997) have identified various ways to deal with DIF or item bias. Firstly, these authors state that the presence of biased items can be viewed as an indicator that an instrument is inadequate for cross-cultural comparison, or that the instrument should not be used across cultural groups. Secondly, item bias may provide important clues about cross-cultural differences (Van de Vijver & Leung, 1997). In other words, unbiased items define the culture commonalities of a construct, while biased items denote cultural idiosyncrasy (Van de Vijver & Leung, 1997). Therefore, as a result of the finding of bias, a more comprehensive picture may be developed on the basis of the universal and culture-specific elements of a construct. The final and most common way to deal with bias is to treat it as a disturbance at the item level that should be removed (Van de Vijver & Leung, 1997). In other words, only unbiased items constitute a solid basis for cross-cultural comparison, and item bias analysis can be used to identify and remove biased items. Van de Vijver and Leung (1997) also emphasise the importance of extending research on equivalence to the level of construct bias to explore the extent to which the problems with tests can be linked to test-wide bias (thus construct bias). Sireci and Khaliq (2002) recommend cross-validating the findings of item bias before bias is accepted, as Type 1 error in DIF findings in small samples may be an issue. Group differences (on mean scores) can thus be the result of the fact that different constructs are being measured, in addition to the fact that the scales may have different origins and measurement units as a result of item bias.

On the basis of the findings of this study, the researcher therefore recommends that the various items that displayed DIF should be investigated further and possibly be removed from the various sub-tests to increase the probability that an individual score is obtained on the basis of ability only and that language does not influence the individual's results. Furthermore, the construct equivalence of the sub-tests across the language groups needs to be investigated. From the findings of this study, and through innovative research (in the area of language testing), it is evident that, just as in the case of multilingual tests, monolingual tests need to be evaluated to determine whether bias exists. Should biases be found, these biases should be eliminated to ensure that test results can be compared meaningfully across groups.

Note

[1] We use the terms 'item bias' and 'DIF' interchangeably, but accept Van de Vijver and his co-researchers' position on item bias.

References

Allen MJ & Yen WM. 1979. *Introduction to measurement theory*. California: Brooks/Cole.

Anastasi A & Urbina S. 1997. *Psychological testing*. New Jersey: Prentice-Hall.

Foxcroft C & Roodt G. (Eds.). 2009. *An introduction to psychological assessment in the South African Context*. 3rd edition. Cape Town: Oxford University Press South Africa (Pty) Ltd.

Foxcroft & Roodt. 2001. [***AQ: details?***]

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**: 2633.

International Test Commission (ITC). 2001. International guidelines for test use. Available at:

www.intestcom.org/itc_projects.htm [accessed 12 November 2010].

Koch SE. 2005. Evaluating the equivalence, across language groups, of a reading comprehension test used for admission purposes. PhD thesis, University of Port Elizabeth, 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) *Multilingualism and exclusion. Policy, practice and prospects.* Pretoria: Van Schaik Publishers, pp 79–103.

Koch E. 2009. The case of bilingual language tests: A study of test adaptation and analysis. *Southern African Linguistics and Applied Language Studies* **27**(3): 301–317.

Nagelkerke. [***AQ: details?***]

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.

Swaminathon H & Rogers HJ. 1993. A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Journal of Educational Measurement* **33**: 215–230.

Van de Vijver FRJ & Leung K. 1997. *Methods and data analysis for cross-cultural research.* Sage Publishers.

Van de Vijver FJR & Poortinga YH. 1997. Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment* **13**(1): 29–37.

Van de Vijver FJR & Rothman YH. 2003. Assessment in multicultural groups: The South African case. *South African Journal of Industrial Psychology* **30**(4): 1–7.

Van de Vijver FRJ & Tanzer NK. 1998. Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology* **47**: 263–279.

Woodcock RW & Muñoz-Sandoval AF. 2001. *Woodcock-Muñoz Language Survey: Normative Update.* Itasca, IL: Riverside Publishing.