THE EVALUATION OF THE GROUP DIFFERENCES AND ITEM BIAS OF THE ENGLISH VERSION OF A STANDARDISED TEST OF ACADEMIC LANGUAGE PROFICIENCY FOR USE ACROSS ENGLISH AND XHOSA FIRST-LANGUAGE SPEAKERS

Genevieve Ruth Haupt

A mini-thesis submitted in partial fulfilment of the requirements for the degree of M.A. Psychology in the Department of Psychology, University of the Western Cape.

Supervisor: Prof. S. E. Koch

January 2010

Keywords: Language Proficiency, Woodcock Muñoz Language Survey, Language in Education, Language Assessment, Monolingual Assessment, Cross-cultural Assessment, Cross-linguistic Assessment, Bias, Equivalence, Differential Item Functioning.

ABSTRACT

South Africa's Language-in-Education Policy is one of additive multilingualism, but in reality this policy is not adhered to, in that most black children are being educated through the medium of English from Grade 4. This type of instruction affects the development of academic language proficiency in their primary language, as these children are not engaging in cognitively demanding tasks in their primary or first language. The Woodcock Muñoz Language Survey (WMLS) is a test to assess academic language proficiency in Additive Bilingual Education, and is extensively used in the United States of America (USA) for this purpose. It is important to note that the proposed study is a sub-study of a larger study, in which the original WMLS (American-English version) was adapted into English and Xhosa, to be used in South Africa to assess additive bilingual programmes. For this sub-study, the researcher was interested in examining the overall equivalence of the adapted English version of the WMLS. Owing to insufficient tests evaluating academic language proficiency in the South African context, the significance, as well as the overall aim, of the study is to ensure that the issues of group difference and item bias have been assessed to ensure that the adapted English version of the WMLS is suitable to be used across English first-language and Xhosa first-language speakers. Because this is a sub-study, the researcher (of the sub-study) has conducted an exploratory quantitative study with the use of Secondary Data. The researcher has used the framework of equivalence as a theoretical framework in order to examine the research question. Given the use of existing data, the procedures of the collection of the data by the researcher of the larger study have been outlined in the Methodology section of the present study. The sample consisted of 198 English and 197 Xhosa first-language speakers, who were selected using convenience purposive sampling from "ex Model C" and "previously disadvantaged" schools in Port Elizabeth and Grahamstown. The main findings of this study indicated that there were overall differences between the two language groups on the various subtests of the adapted English version of the WMLS with regard to group differences, reliability, item characteristics, and differential item functioning. Therefore the adapted English version of the WMLS cannot be used for score comparison across the two language groups, namely English and Xhosa first-language speakers.

DECLARATION

I declare that the "the evaluation of the group differences and item bias of the English version of a standardised test of academic language proficiency for use across English and Xhosa first-language speakers" is my own work, that it has not been submitted before for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged as complete references.

Genevieve Ruth Haupt January 2010

Signed: . . . . . .

ACKNOWLEDGEMENTS

I would firstly like to thank God for giving me the strength and perseverance to succeed far beyond my expectations.

Next I would like to thank my parents for their support and understanding during the course of my studies, and my fiancé, Roderigo Rix, for his patience and encouragement. I would also like to thank all my friends and colleagues who have supported me, and give a special thank you to my friend Mujaahidah Mohamed for all her encouragement and assistance over the last year.

Finally I would like to thank my supervisor, Prof. S. E. Koch, for all her guidance and encouragement, and for urging me on when it was needed. I would like to thank her as well as the National Research Foundation (NRF) for giving me the opportunity to be part of this larger, ground-breaking-study, and for awarding me an NRF bursary.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FORMULAS

CHAPTER 1

INTRODUCTION

1.1 Introduction

The South African policy on Language-in-Education is one of additive multilingualism or bilingualism (Cummins, 1984; 2000). Additive bi-or multilingual education is a form of education where the primary (and strongest) language of learners is used for cognitive and literacy development, while at the same time they are being taught a second and/or a third language (in most cases this would be English, Afrikaans, etc.). However, in reality most black children are being educated through the medium of English from Grade 4. This affects the development of academic language proficiency in their primary language as well as in their second language, since these children are not engaging in cognitively demanding tasks in their primary language, and often also not in their second language as they are not sufficiently proficient in this language to engage in higher-order academic tasks in this language (Cummins, 1984; 2000).

Only after children are sufficiently proficient in an additional language will they benefit from education thought the medium of the additional language. Issues related to testing in a language that is the first language of some children but the second or third language of others, is thus an issue that needs urgent attention in this context, as this will affect any recommendations made about a child's readiness to be instructed in his or her additional language. This form of testing is also called "monolingual testing".

This study focuses on the challenges inherent in using a monolingual test to assess multilingual or bilingual speakers. More specifically, the focus will be on some of the issues embedded in multicultural and multilingual assessment or testing, such as equivalence and bias.

1.2 Background

There are various reasons why most South African learners, as well as learners in other parts of Africa, have not been instructed through the medium of their primary language (as outlined above). The main reason (with regard to the whole of Africa) is the effect of colonialism. In addition, in the South African context, the most infamous reason is the previous political infrastructure, namely apartheid (as well as colonialism) which described certain racial groups as "different" and perpetuated these "differences" by discriminating against these groups politically, economically, and most importantly, educationally.

During apartheid mother-tongue education was enforced, and learners were instructed in their mother-tongue language until Grade 8. The black population of South Africa (SA) felt that this was just another ploy by the apartheid government to maintain power over the non-white population of SA, in that mother-tongue instruction formed part of the education system of "bantu education" (Apartheid South Africa: Bantu Education, paragraph 3, 2000). Historically, black learners were discriminated against educationally as they were instructed through an education system known as "Bantu education". Bantu education fundamentally limited their knowledge and aimed to direct black or non-white youth to the unskilled labour market, to ensure white control and prosperity (Apartheid South Africa: Bantu Education, paragraph 3, 2000). After the uprisings against the enforced introduction of Afrikaans as a medium of instruction in 1976, black speakers of the African languages therefore insisted on changing the policy to one of teaching through the medium of English from Grade 4, which left a legacy that continues to impact on attitudes about language to this day, and contributes to the lack of the implementation of the language-in-education policy by Government.

In addition, tests, such as achievement tests, used to assess these learners, initially were standardised only for white learners and sometimes "coloured" learners. This essentially meant that black learners would score less on these tests, as the items on the test were only standardised for white learners; therefore these tests could be said to be biased.

The trajectory of language-in-education in the rest of Africa followed a different path to the one in SA, but the use of the colonial language in education still epitomizes education throughout Africa south of the Sahara, with the same pattern of failures in school as in SA, and the resulting underdevelopment in the continent (Alexander, 2002).

In order to address the implementation of the SA's Language-in-Education Policy, which is one of additive multilingualism, the Additive Bilingual Education (ABLE) project was implemented in 2002 at a school consisting of mainly first-language Xhosa speakers in the rural areas of the Eastern Cape (Koch, 2009). This programme needed to be evaluated. Therefore, the Woodcock Muñoz Language Survey (WMLS), an American instrument, was selected, and adapted into South African English and Xhosa to be used to evaluate the academic language proficiency of the learners in these specific languages. The larger study has various sub-studies and the current study is one of the sub-studies which focuses on the equivalence of the adapted English version of the WMLS for use across English and Xhosa first-language speakers.

The larger study consisted of various phases. The first was the adaptation of the original WMLS (English version) into SA English and Xhosa. Second was the evaluation of the equivalence of the two language versions of the WMLS across English and Xhosa first-language learners, and third was the evaluation of the predictive, construct and content validity of both these adapted versions across English and Xhosa first-language learners in the SA context.

As previously mentioned, this sub-study focuses on examining the overall equivalence of the adapted English version of the WMLS, and more specifically, on the evaluation of item bias, as well as group differences, in order to determine whether this version is suitable to use across English first language as well as Xhosa first-language speakers. It is important to note that in using this test, the claim will be made that it can be used to make the same statements with regard to English academic language proficiency of both Xhosa first-language speakers and

English first-language speakers. In other words this test can be viewed as a monolingual test, as it is available in one language (English) and it was used to assess an addition language group (Xhosa first-language learners) as well as the English first-language group. To be able to do this, the equivalence of the test scores of these two groups on this test, needs to be established. These concepts will be explained further in the section on Theoretical Framework.

1.3 History of language testing

According to Kaplan and Saccuzzo (1997), testing is viewed both as recent and as pioneered by America. To some extent this is true, as the major developments in testing have occurred this century, and many of these developments have taken place in America. However, the origins of testing are neither recent nor originating in America (Kaplan & Saccuzzo, 1997). Historians have found evidence that the Chinese had a relatively sophisticated civil service testing programme more than 4000 years ago (Dubois, 1970, 1972, as cited in Kaplan & Saccuzzo, 1997). Oral examinations were set every third year in China and the results were used for work evaluations and promotion decisions.

During the mid-1960s through the 1970s, language-testing practice was reflected in large-scale institutional language testing. When consulting language textbooks of the time, it can be seen that these practices were informed essentially by a theoretical view of language ability including skills such as listening, speaking, reading and writing, as well as components of language such as grammar, vocabulary and pronunciation (Bachman, 2000). An approach to test design focused on testing isolated "discrete points" of language, while the primary concern was psychometric reliability. Fundamentally (during this era) language-testing research focused on the hypothesis that language proficiency consisted of a single unitary trait, giving rise to a quantitative, statistical research methodology (Bachman, 2000). In other words, language proficiency consisted of a single, global ability that was widely accepted. An example of this would be using one

of the components of language, such as vocabulary, testing it, and concluding that this is what language proficiency consisted of.

By the 1980s, language testing (as a subfield within the realm of applied linguistics) had come a long way. For instance, one of the areas within language testing which saw expansion was the influence of second-language acquisition (SLA) research which not only encouraged researchers to investigate a wide variety of factors such as field independence/dependence, academic discipline and background knowledge and discourse domains on language test performance, but also to examine the strategies involved in the process of test-taking itself (Bachman, 2000). According to Bachman (2000), a major research project which was of particular importance in this regard, was the Development of Bilingual Proficiency, as it provided an essential avenue of investigation into the nature of language proficiency and its acquisition in both educational and informal settings. This was done by using innovative language tests as the principle research instruments (Bachman, 2000).

Towards the end of the 1980s language testers were severely challenged in a paper by Pienemann *et al.* (1988) to explicitly take the language of learners' developmental sequence into consideration in the design of tests and in the interpretation of test scores. When reflecting on the developments in the 1980s it is clear that testers were moving towards a more mainstream approach to testing. In other words, the unitary view which was the common paradigm during the previous decades became obsolete, and a "multi-componential approach" to language proficiency was adopted. This refers to an approach that consisted of a number of interrelated specific abilities, as well as a general ability or set of general strategies and procedures (Bachman, 2000).

In the 1990s this trend continued. The 1990s saw further developments in the following realms: research methodologies; practical advances; factors that affect performance of language tests; authentic, or performance, assessments; and concerns with ethics of language testing and professionalising the field. With

regard to the practical advances during this time, one of the main advances made during this time was the testing of cross-cultural pragmatics. According to Bachman (2000), while much of the mainstream test development continued to focus on the traditional areas of linguistic components such as grammar and vocabulary, also known as the four skills, listening, speaking, reading and writing, researchers at the University of Hawaii broke new ground in assessing what they called the "cross-cultural pragmatics". In other words, these researchers, by drawing on research in linguistic pragmatics, SLA and sociolinguistics, Hudson *et al.* (1992, 1995), developed a framework and a set of assessment instruments that were aimed at providing information about two sources of variation in cross-cultural communication, firstly the variability associated with the social properties of the speech experience, as well as the speaker's choice for achieving various communication goals, and secondly, the variability due to the particular types of data collection procedures and associated instruments (Bachman, 2000). The various prototypes of instruments that the researchers developed in English for indigenous speakers of Japanese provided model assessments, both for research into the nature of cross-cultural pragmatics and for the practical assessment of pragmatic competence in cross-cultural communication (Bachman, 2000).

Despite the fact that these researchers were the first to address an area of language ability that had not yet been well researched in the language assessment literature, according to Bachman (2000) there were three additional features worth elaborating on regarding the researchers of this decade. Firstly, the research conducted during this time clearly involved collaboration among language testers and researchers in other areas of applied linguistics, exemplifying an interface between language testing and other areas of applied linguistics (Bachman & Cohen, 1998). Secondly, these researchers were committed to including a wide variety of task types, including multiple-choice or cued response items, structured oral interviews, self assessments, and direct observations, as part of the assessment procedure. Finally, the researchers followed sound test development procedures and conducted the project in carefully planned cyclical research and development mode, with monitoring, evaluation and revision built into each cycle.

According to Bachman (2000), some of the data produced from these researchers promised to provide some truly novel insights into the realm of language ability and language use. With regard to the history of testing, developments in language performance in language assessment, fed by related developments in language teaching and educational measurement, have led to a better understanding of the nature of the methods, or tasks, used to elicit performance in language assessments (Bachman, 2000). Furthermore, time has also brought about a better understanding of the ways in which we can design, develop and use such tasks and evaluate their usefulness (Bachman, 2000).

Overall, Bachman (2005) explains that there are two sets of determinants of test performance. The first attribute he identified is the various attributes of the test takers themselves which they bring to the assessment. More specifically, the attribute that most researchers are interested in, is language ability (Bachman, 2005). In other words, this refers to the attribute that was intended to be measured, the construct (Bachman, 2005). The second attribute, which completes this set of determinants, is the various aspects of the assessment tasks and the situation in which the assessment takes place (Bachman, 2005). In a review, Bachman (2005) concludes that investigating the relative effects of these determinants on test takers' performance and on the scores they receive in language tests, is effectively the process of validation.

1.4 History of testing in South Africa

Due to SA being previously under British rule, the introduction of psychological assessment in SA was predominantly related to our colonial heritage (Claassen, 1997). Initially, psychological assessments were characterised by an environment which had an unequal distribution of resources based on racial categories (black, coloured, Indian and white) (Foxcroft & Roodt, 2005). Therefore it was almost a certainty that psychological assessment would reflect a racially segregated society. Indeed, any account of psychological assessment in SA needs to be aware of the substantial impact that apartheid policies had on test development

and use (Nzimande, 1995).

Owing to apartheid, not only were the early measures of testing standardised for whites only, but these measures were also driven by political ideologies (Foxcroft & Roodt, 2005). This was the case, for example, where measures of intellectual ability were used in research studies to draw distinctions between races in an attempt to show the superiority of one group over another, particularly during the 1930s and 1940s when the government was grappling with the issue of establishing "Bantu Education". Fick (1929) as cited in Foxcroft and Roodt (2005), for example, administered individual measures of motor and reasoning abilities, which had been standardised only for white children, to a large sample of black, coloured, Indian and white school children. He found that the mean score of black children was inferior to that of Indian and coloured children, with whites mean scores superior to all groups. When completing his research, he attributed the difference in scores to various factors (present at the time), listing them as inferior schools and teaching methods, along with black children's unfamiliarity with the nature of the test tasks, and these factors could have disadvantaged their performance on the measures. However, when he extended his research he stated that the differences were due to innate differences in the original ability of the non-white test-takers (Fick, 1939, as cited in Foxcroft & Roodt, 2005).

With regard to Fick's extended research, its implications were that the various group differences (such as, culture, socioeconomic status as well as familiarity with the content of the test, etc.) had not been taken into account in the interpretation of the results as well as the fact that the test had been standardised for white children (in itself), the test should not have been considered to be suitable or applicable to non-white children. In other words, these extended findings perpetuated the apartheid ideology. From the above discussion of Fick's study it is clear that testing in education was also marked by inequality and segregation.

1.5 Language Testing in Education

Language testing almost never exists in isolation; rather it is done for a particular purpose and in a specific context (Bachman, 1990). According to Bachman (1990) there is an intrinsic reciprocal relationship between research in language acquisition and developments in language teaching on the one hand, and language testing on the other hand. In other words, language testing both serves and is served by research in language acquisition and language teaching (Bachman, 1990). According to Bachman (1990), language tests are frequently used as criterion measures of language abilities in second-language research. Similarly, language tests can be valuable sources of information about the effectiveness of learning and teaching (Bachman, 1990). An example of these tests would be teachers' evaluation of students' progress and assisting in student achievement, such as comprehension tests, which were suppose to be used to test learners' comprehension skills among other skills related to this type of testing.

According to Bachman (1990), the insights gained from language acquisition research and language-teaching practice can provide information for designing and developing more useful tests. Current research and development in language testing incorporates advances in several areas that include research in language acquisition and language teaching, theoretical frameworks for describing language proficiency and language use, as well as measurement theory (Bachman, 1990).

However, in the South African context, as is the case with many other contexts (or countries), there is a diverse cultural society that is rich with many different languages. Therefore in the realm of language testing, issues of cross-cultural and cross-linguistics testing are also pertinent.

1.6 Cross-cultural and Cross-linguistic Testing

The most common characteristic of cross-cultural studies is their comparative nature, which involves the comparison of at least two cultural populations (Van de Vijver & Leung, 1997). Many of these comparative studies involve different nation states, in sociology, education, political sciences, management, and psychology (Van de Vijver & Leung, 1997). However, comparative studies can also involve different ethnic groups from a single country. In these studies tests are used across diverse groups, such as blacks, whites and coloureds, as was the case with Fick's 1929 individually administered tests for motor and reasoning abilities. Studies conducted across groups (cross-culturally or cross-linguistically) where specific measures are used to evaluate individuals within each group, are marked by various issues. Two such issues are equivalence of test scores across and possible bias that could arise with the use of these tests across groups.

1.7 Bias and Equivalence

Equivalence and bias play a pivotal role in testing, even more so within the realm of cross-cultural and cross-linguistic testing. According to Van de Vijver and Leung (1997), the attainment of equivalence is perhaps the most central issue in cross-cultural comparative research. For measures to be equivalent, individuals with the same or similar standing on a construct, such as first-language speakers of Xhosa or English, should obtain the same or similar scores on the different language versions of the items or measures, or on a test that is available in only one language. If not, the tests are said to be biased, and the two versions of the measure are non-equivalent, or the scores on the test do not have the same meaning (Foxcroft & Roodt, 2005). The concepts of bias and equivalence will be elaborated on in the following chapter (Literature Review), and will form the central focus of this study on the English version of the WMLS.

1.8 Rationale and aim

Within the South African context, there is a need for more models regarding the effectiveness of additive bilingual education, as well as ways in which to assess the effectiveness of these models. These evaluation models should be able to assess the effectiveness of additive bilingual education to address the development of the academic language proficiency of children in their primary language as well as additional languages, especially the language of power and control in the wider community. However, in the South African context, this is hampered by the lack of unbiased and appropriate instruments to perform such a task efficiently. In addition, with the use of monolingual assessment measures, when used to assess more than one language group (where the target language of the test is not the first language of one group, but is the first language of the other group), there needs to be further investigation to establish whether the measure is equivalent across the groups.

The WMLS is a test which can efficiently measure the development of academic language proficiency in an individual's primary and secondary languages, and is used extensively in the USA to evaluate Additive Bilingual Education (Woodcock & Muñoz-Sandoval, 2001). The WMLS was therefore selected and adapted into SA English and Xhosa to assess the academic language proficiency of English first-language and Xhosa first-language speakers.

The overall aim of the study was to evaluate the equivalence of the adapted English version of the WMLS for use across English first-language and Xhosa first-language learners. The researcher has outlined four specific objectives in order to achieve the overall aim of this study.

1.8.1 Objectives

The specific research objectives are:

1. To evaluate mean group differences between English and Xhosa first-language learners on the English version of the WMLS.
2. To evaluate group differences in terms of the reliability of the test between English and Xhosa first-language learners on the English version of the WMLS.
3. To evaluate group differences in terms of the item characteristics of the test between English and Xhosa first-language learners on the English version of the WMLS.
4. To evaluate the item bias, across English and Xhosa first-language learners, of the English version of the WMLS.

The first chapter, namely the Introduction, has focused on the background of this study as well as the history of testing, the history of language testing, the history of testing in education, cross-cultural testing and cross-linguistic testing, as well as the rationale, aim and specific objectives of this study.

The second chapter, namely the Literature, will comprise the reviewing of various literature studies that have been conducting in the field of language testing or assessment with a specific focus on monolingual assessment with cross-cultural and cross-lingual elements. This chapter will be concluded with a discussion on the theoretical framework employed in order to outline the present study around a particular theoretical framework.

Chapter 3, Methodology, will outline the research methodology of the main study as well as of this study. The fourth chapter contains the results of the various statistical procedures as well as the null hypothesis stated for each specific objective. Chapter 5, Discussion and Conclusion, is an in-depth discussion generated from the Results chapter, as well as further recommendations.

CHAPTER TWO

MONOLINGUAL ASSESSMENT IN SOUTH AFRICA: PRACTICES AND ISSUES, AND THE THEORETICAL FRAMEWORK OF EQUIVALENCE

2.1 Introduction

The main focus of this study (also stated as the overall aim of this study in the preceding chapter) is the evaluation of equivalence of the adapted English version of a language proficiency test to be used across two language groups, namely English first-language speakers and Xhosa first-language speakers, in the South African context. This focus lends to testing in a cross-cultural and cross-linguistic setting (such as SA). With this in mind this chapter will highlight cross-cultural testing in general, monolingual testing specifically, and the use of monolingual language testing in a multilingual society. Within the realm of cross-cultural assessment the use of monolingual assessments has not really been explored before, and therefore there are limited resources available on this topic. In this chapter the terms "assessment" and "testing" will be used interchangeably. In addition, the theoretical framework employed will be discussed in greater detail..

2.2 Cross-cultural Assessment

According to Meiring (2007), cross-cultural assessment has emerged as an important area of research from the realm of cross-cultural psychology. Briefly, cross-cultural psychology is the systematic study of the relationships between the cultural context of human development and the behaviours that become established in the collection of individuals growing up in a particular culture (Meiring, 2007). When a psychometric measure, for example a measure of personality, or intelligence or some other construct, is used across various cultures, to compare test candidates and more particularly their test scores, this is known as cross-cultural testing or assessment (De Klerk, 2008). With cross-cultural testing, the need for multiple language versions of tests, questionnaires and surveys is continuously increasing, especially for comparative studies across

cultural and language groups (De Klerk, 2008). With the increasing need for multiple language versions of tests, many tests are adapted from one language and culture to another. However, the adapted measures should not be taken to be valid by face value alone. De Klerk (2008) emphasises that the influence of culture on assessing specific psychological constructs needs to be explored, and should be able to be adjusted to make them meaningful to the particular culture and to get equivalent or comparable measures across cultures.

In the context of South Africa and many other countries, one of the most apparent issues in cross-cultural assessment is language (De Klerk, 2008). For example, a test written in English cannot be expected to produce a sound measure of the same construct within the Xhosa population of SA. Therefore in order to give both English and Xhosa candidates the same starting point in completing a test, it should be adapted and made available in the native language of a specific group, in this case the Xhosa group. In addition, both versions of the test should be equivalent. According to de Klerk (2008), there is more to ensuring that a measure is equivalent than just the process of translating tests from one language to another. He elaborates that further consideration needs to be given when measuring psychological constructs across cultures in that the influence of culture on the measurement needs to be explored. That is, for each psychological process or construct to be measured in a new cultural population, it is necessary to determine the extent to which it is universal across cultures, and, if not, to specify the exact differences, and make the necessary adjustments in order to achieve this example of universality (De Klerk, 2008). Essentially, one cannot assume that what is relevant in one culture or language, when translated, is relevant in another culture or language. Therefore it is clear that culture influences the measurement of a psychological construct.

Within the realm of cross-cultural psychological research, de Klerk (2008) has identified three main theoretical orientations: 1) absolutism, 2) relativism and 3) universalism. Firstly, cultural absolutism assumes that in all cultures, the majority of observable psychological characteristics like personality traits or cognitive

abilities are the same (De Klerk, 2008). Secondly, cultural relativism assumes that culture determines almost all human behaviour (De Klerk, 2008); in other words, that the differences in human behaviour can be explained through differences in the cultural contexts of society (De Klerk, 2008). Finally, cultural universalism is a kind of "middle option" and assumes that basic psychological processes are universal for all people, and that culture will also influence the development and expression of psychological characteristics (De Klerk, 2008).

These three outlined theoretical approaches borrow from different methodological approaches to test construction, namely the "etic", the "emic" and the "derived emic" (De Klerk, 2008). In essence, the "etic" approach assumes that universals can be identified in intelligence or psychological constructs, and one specific standard test can be applied cross-culturally (which relates to the orientation of absolutism) (De Klerk, 2008). On the other hand the "emic" approach claims that for assessment, culture-specific measurements need to be developed, preferably by an indigenous psychologist, as these measurements should be based on culture meanings and value systems with regard to the psychological characteristic being measured (this approach relates to the theoretical orientation of relativism) (De Klerk, 2008). According to de Klerk (2008), the middle way in test construction is the "derived emic" whereby the two approaches of etic and emic test construction are combined. In this approach, a test with a relatively universal measurement potential might be constructed in one culture (etic) and subsequently indigenous researchers from other cultures derive culture-specific versions (emic) from the universal stem (De Klerk, 2008). Following these empirical comparisons between the response on the different tests will demonstrate whether the test measures the same construct or not across the different cultural or language groups (this approach lends from the theoretical orientation of universalism) (De Klerk, 2008).

In addition to the various theoretical stances and approaches to cross-cultural testing there are various issues that arise from cross-cultural assessment. These issues may arise from one of the main reasons for cross-cultural assessment,

namely comparison studies across cultural and language groups. Essentially, comparative studies need to ensure fairness, since comparisons are to be made between people from different backgrounds. These comparisons would therefore need to take place on the same "scale" in order to avoid comparing apples and oranges (De Klerk, 2008). For example, if a company is developing a criterion for managers who originate from different cultures, the same scale needs to be used to ensure fairness, in other words that no bias exists. It is important to note that when administering an instrument or test in the context of various linguistic and cultural groups, the psychological characteristics, or at least the roots of these characteristics, are assumed to be universal for all groups (De Klerk, 2008), thereby ensuring that the same construct as well as the process undertaken when administering the instrument is the same across groups. This is highlighted by Poortinga and Van der Flier (1988, as cited in de Klerk, 2008) who stated that to be able to use tests in different cultural populations, one has to assume that:

- The behaviour domain (ability or trait) as sampled by the items, has at least approximately the same meaning in each culture;
- The ability or trait measured, at least approximately plays the same role in organisation of behaviour in each culture, and
- The scores have, at least approximately, the same meaning for test takers (in a quantitative sense) independent of their cultural backgrounds.

According to Van de Vijver and Leung (1997), with the comparability of test scores between various cultural populations, the issues of bias and equivalence arise. Bias occurs when differences in scores on an indicator of a particular construct do not correspond to differences in the underlying trait or ability (Van de Vijver & Leung, 1997). "Equivalence", on the other hand, refers to the measurement level at which scores can be compared across cultures (Van de Vijver & Leung, 1997). These two concepts, namely bias and equivalence, play a pivotal role in cross-cultural assessment, and in order to make valid comparisons and to draw inferences across cultural groups, measures should have optimal equivalence and thus have minimal bias. Bias and equivalence will be explored in depth in the following section (Theoretical Framework).

Despite the need for various assessment measures to be adapted into multiple languages, there are some institutions or organisations that still use monolingual assessment. The next section will explore monolingual assessment and the issues and implications arising from this type of assessment.

2.3 Monolingual Assessment

"Monolingual assessment" refers to the use of assessment measures which are available only in one language across two or more language groups (Koch, 2005). Many countries, including SA, use monolingual tests to measure individuals on a particular trait.

An example of a type of test used in the SA context is admission tests for the purpose of selection into Higher Education (Koch, 2005). 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 if a monolingual test is not used to assess proficiency in a specific language, according to cross-cultural research, it should be made available in more languages of the given population in order to ensure fairness. Huysamen (2002) points out that unless a test is 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. According to Huysamen (2002), even though the American Psychological Association (APA) standards for educational and psychological testing for employment (referred to hereafter as the *Standards*) does not equate group differences with unfairness, it essentially goes further in that it recognises that such differences may be due to construct-irrelevant variance or construct under-representation. In light of this, when group differences are identified, the researcher needs to evaluate the construct being measured with more scrutiny for construct-irrelevance and construct under-representation.

Construct-irrelevance occurs when a construct being measured may be relevant in one group and not in another group. According to Huysamen (2002), owing to SA's multilingual society (11 official languages), one's home language often represents a potential source of construct-irrelevant variances, which should always be considered when measuring a construct cross-linguistically. Furthermore, Huysamen (2002) states that unless a test is used to measure language proficiency (testing an individual's proficiency in a given language), an individual should not be given a test that is not in a language that he or she is proficient in when measuring a construct other than language, as this might lead to construct-irrelevant variances.

Huysamen (2002) further states that irrelevant variances may not be restricted to language proficiency only, but could also extend to cultural differences that the test is not designed to measure. For example, there is an item in the HSRC's Group Test for Five-and-Six-year-olds in which the examinees are shown line drawings of a lion, a dog, a deer and a cow, and are required to identify the tame animal that "we do not eat" (Huysamen, 2002). Ramphal (1972) indicated that Indian children of the Hindu religion might fail this item because their religion would prohibit the slaughter of cows. Therefore this example illustrates that even a perfectly acceptable translation of a test from one language into another may still be subject to construct-irrelevant variances.

In light of the above discussion, it can be argued that tests that lack relevant construct variances or that contain irrelevant construct variance because of their language requirements, or because of the close association between language and culture, may result in differences in group means (Huysamen, 2002). On the other hand, construct under-representation occurs when a construct is being measured which is under-represented in a test or measure. An example occurs in measuring job performance, when productivity is assumed to be the only variable that measures an individual job performance. There are other variables such as conscientiousness and work ethic that are equally important in measuring job performance. Therefore it can be said that the construct of job performance is

under-represented, as a single attribute or variable will rarely be sufficient to predict the criterion of job performance (Huysamen, 2002).

With regard to the use of monolingual tests internationally, Koch (2005) states that in the USA the Standardised Achievement Test (SAT) and the American College Testing (ACT) are the most commonly used instruments for admission purposes. It is noteworthy that both of these achievement tests (SAT and ACT) are available in English only, and the Spanish-speaking minority population of the USA also have to complete these tests in English. However, in the USA there is increasing resistance to this practice and the testing of linguistic minorities on these tests (Valdés & Figuoroa, 1994).

Further research conducted on the use of monolingual tests internationally such as Escamilla (2006) indicates that there is still continual use of tests that are available only in one language to assess populations where the language of the test is the second language or even third language of the population being assessed. According to Escamilla (2006), a number of people are labelled as "language minority" people in the USA, when referring to people who have English as a second language or are bilingual, or people whose heritage language is not English (Escamilla, 2006). Escamilla (2006) further states that in 2003 the number of language minority people in the USA was estimated to be around 7.5 million.

The US Department of Education requires all schools in the USA to determine the extent to which students who are language-minority students are also limited in English proficiency (Escamilla, 2006). At one point in the USA, limited English proficient children were labelled Limited English Proficient (LEP), but are now more commonly referred to by federal, state and local school districts as English Language Learners (ELLs)(Escamilla, 2006). According to Escamilla (2006), the US Department of Education estimates the total number of ELLs in the US public schools to be about 5 million, and the US Department of Education further estimates there to be about 350 different language groups included in the

population of language minority students and ELL students. The State of Colorado, for example, has documented more than 120 language groups represented in its public schools (Escamilla, 2006). However, on closer inspection, this statistic is somewhat deceptive because in reality the vast majority of about 70% of ELLs in the USA speak one common language, which is Spanish (Escamilla, 2006).

Escamilla (2006) has conducted a study which focuses on Colorado – a State whose demographic situation is similar to that of the USA. About 70 000 public school students in Colorado were identified as ELLs, which constituted about 10% of the entire school population. Furthermore, it was found that, although there were 102 documented language groups in Colorado schools, the vast majority (56 000 students or 80%) spoke Spanish as a first language. According to Escamilla (2006), despite the national and state data showing that the overwhelming majority of the language minority and ELLs speak a common language (Spanish), language diversity is often used as a reason to promote monolingual English assessment policies and in many places to require that assessments be conducted only in English because, in their view, it would be "too expensive" and "not feasible" to develop assessments in 102 languages. Escamilla (2006) points out that with the creation of assessment measures in Spanish, these measures would account for 80% of the linguistic population, and possibly create a more equitable and accurate assessment system for ELLs. In other words the costs would be minimal, as the Spanish measures would account for 80% of the total population of ELLs.

Ruiz (1988, as cited in Escamilla, 2006) has suggested that there are three basic language orientations employed by the US nation, as well as communities and schools as they engage in the creation of language policies and planning, namely language as a problem, language as a right, and language as a resource. Ruiz postulates that ELLs in the US, more specifically those ELLs whose first language is Spanish, have historically been viewed as a problem, therefore US schools have been charged to create policies and practices to "fix the language problem" of

Spanish-speaking students. Escamilla (2006) further states that recent educational reform movements such as standards-based education and high-stakes testing have served to exacerbate the notion that speaking a language other than English in the USA is a problem that must be remediated by the schools. Furthermore, with the inception of standards-based education and high-stakes testing (in the USA), these assessments use the native English speakers as the norm group (Escamilla, 2006). An issue that has arisen from these types of testing is that for second-language learners, paper/pencil content assessment is used, and these assessments often measure students' lack of proficiency in English, not their knowledge of the content (Escamilla, 2006).

Another example of a critical study regarding monolingual assessment which has been conducted in the South African context is that of Foxcroft and Aston (2006). They critically examine the adaptation process undertaken by the Human Sciences Research Council (HSRC) when standardising the WAIS-III for English-speaking South Africans, by deliberating whether sufficient attention was paid to establishing whether this monolingual measure (as it was available in English only) was equivalent for various groups of English first- and second-language test-takers. Foxcroft and Aston (2006) have found that the equivalence of the WAIS-III across diverse language groups was not unequivocally established, and that there were indications that some bias may exist for English second-language test-takers, especially if these test-takers are black or Afrikaans-speaking. Recommendations proposed by Foxcroft and Aston (2006) include: first undertaking a more thorough, qualitative bias review of all of the subtests to detect items and instructions that require adaptation or replacement; secondly, the establishment of equivalence of the measure across the diverse language groups that it is intended for; thirdly, the re-examining of whether separate norms for the various combined language and cultural groups might result in the measure being able to be used more fairly with English second-language test-takers in particular; refining the Afrikaans translation of the verbal subtests and using judgemental and empirical methods to establish the equivalence of the English and Afrikaans version, and finally, exploring the possibility of adapting/translating the measure

into various African languages. According to Foxcroft and Aston (2006), having the WAIS-III available in multiple languages will allow psychologists to assess test-takers in the language in which they are most proficient.

However, Foxcroft and Roodt (2009) state that test adaptation in SA has not flourished in terms of developing multiple language versions of measures. The assumption is widely held that the majority of South Africans are educated in English from Grade 4 onwards, and as English is the language of instruction at most higher education institutions and in the workplace, assessing test-takers in English is acceptable (Foxcroft & Aston, 2006). From this assumption it is argued that if test-takers cannot perform well on measures administered in English, they will not be able to cope with English as the language of instruction at a higher education level or in the workplace (Foxcroft & Roodt, 2009). Therefore despite the SA context of multilingualism, it has not become standard practice to have multiple language versions for all measures. However, various SA researchers (such as Foxcroft & de Bruin, 2008; Koch, 2003, 2005, 2007, 2008 and 2009; Meiring, 2007; Saunders, 2002 as cited in Foxcroft & Roodt, 2009) have been critical of this assumption for a number of reasons.

An example would be found in the work of Watson, Davies and Foxcroft (2006, as cited in Foxcroft & Roodt, 2009) which indicates that merely being educated through the medium of English is no guarantee that learners are sufficiently proficient in English as a second language at the end of Grade 12. These researchers state that in fact the majority are not proficient (Foxcroft & Roodt, 2009). Furthermore, where second-language English speakers are assessed on a measure developed for first-language English speakers, there needs to be evidence that the measure is assessing comparable constructs for the two groups, and that the language is not a "nuisance factor" that impacts negatively on test performance (Foxcroft & Roodt, 2009). According to Foxcroft and Roodt (2009), in terms of the latter, a number of SA studies have highlighted language proficiency as a potential source of bias for SA measures of both cognitive and personality measures (these studies include Abrahams & Mauer, 1999b; Joseph &

van Lill, 2008; van de Vijver & Rothman, 2006; Meiring, van de Vijver, Rothman, & Barrick, 2005, as cited in Foxcroft & Roodt, 2009).

The preceding discussion on monolingual assessment has been concerned with monolingual assessment and the importance of having assessment measures available in more than one language to ensure fairness, especially in societies with multiple languages. The following section will focus on monolingual language assessment.

2.4 Monolingual Language Assessment

Language cannot be viewed in isolation (Bachman, 2000). Rather, language develops, and is maintained through social interactions and academic influences. McNamara (chapter in Hinkel, 2005) refers to this as "the social context of language". Language assessment or language testing is a field of study that falls under the umbrella of applied linguistics (Bachman, 2000). The main focus of language assessments is the measured judgement of first, second or other language in the school, college or university context; assessment of language use in the workplace, and assessment of language in the immigration, citizenship and asylum contexts (Honrberger, & Shohamy, 2008). In addition to uses of language assessment there is the utilisation of the results of language tests which include firstly making inferences about test takers' language abilities or making predictions about their capacity for using language to perform further tasks in contexts outside the test itself. Secondly, the results of assessment measures are used in decision-making; such as selection, diagnosis, placement, progress, grading, certification, and employment. These decisions are made on the basis of the level of ability or test takers' capacity for non-test language use.

According to Bachman (2000), the development of standards of professional competence and language testing practice that are appropriate across cultures is a challenging task, as it risks cultural imperialism on the one hand, and ethical relativism on the other hand. With regard to cross-cultural language assessment,

Kunnan (2000) places more emphasis on the fairness of language and testing practices. He states that there has always been concern among test developers, test users and test researchers within this realm, and the traditional manner of ensuring fairness has been through investigations of the tests' reliability and validity. However, in the past decade, educational and language assessment researchers have begun to focus directly on the fairness and related matters such as test standards and test bias, equity and ethics, for testing professionals (Kunnan, 2000).

When a test is standardised on a particular group (for example, a racial group), as was the case during the apartheid era where tests were standardised on the white population of SA and used to assess the non-white population of SA, it can lead to various issues around the fairness of the test (Foxcroft & Roodt, 2009).

2.5 Research on Monolingual Assessment

Various researchers have identified different issues with the use of monolingual assessment, such as issues around fairness. These researchers recommend solutions for the various issues, one of which is translation or adaptation. There have been only a few studies done in this area in SA. According to Kunnan (2000), it has been argued that although language test developers and researchers are concerned with the concept of fairness when they investigate tests for technical qualities like validity and reliability, the primacy of fairness has not been considered or acknowledged. Kunnan (2000) adds that fairness as a concept within a framework of social justice has not been developed or debated.

A study was conducted by Meiring, Rothmann and Barrick (2005) in which they investigated the adequacy of cognitive tests and the Fifteen Factor Questionnaire (15FQ+), a personality measure which was used to assess a group of police applicants from all major South African ethnic groups. This study specifically looked at various possible examples of bias regarding the specific instrument used (15FQ+). More specifically, construct, method and item bias were examined.

This test can therefore be regarded as a good example of a monolingual test that is used across language and cultural groups.

"Construct bias" refers to the question of whether the same underlying constructs are being measured in each ethnic group; "method bias" is the generic term for instrument-related and person-related factors that can systematically affect the size of cross-cultural score differences, such as differential social desirability; and "item bias" refers to the presence of items that do not measure the same in each cultural group (Meiring, Rothmann, & Barrick, 2005) (these concepts will be discussed in greater detail under the later heading, Theoretical Framework). The major finding was that several scales of the personality questionnaire revealed construct bias in the various ethnic groups assessed. Item bias existed on the personality scales, but this bias was low, and method bias did not have any impact on the cross-cultural differences in the personality scales (Meiring, Rothmann, & Barrick, 2005). Furthermore, a number of the personality scales revealed low internal consistencies, notably in the Black groups. Therefore it was concluded that the 15FQ+ was not suitable as an instrument in the SA multicultural context because of the low internal consistencies of some of the scales and the lack of construct equivalence (Meiring, Rothmann, & Barrick, 2005). According to these authors, one way in which to remedy bias is by adapting or translating a measure.

Further research that has been conducted in SA on bias, especially with regard to monolingual assessment, is that of Koch and Dornbrack (2008) on the use of language criteria for admission to higher education in the SA context. They indicated that many institutions have adopted a set of additional language criteria for admission, which have bearing on only one or two languages (the current languages of teaching and learning are English and Afrikaans), despite the fact that these institutions have adopted a multilingual language policy which includes one or two African languages as additional languages of teaching and learning (Koch & Dornbrack, 2008).

However, Koch and Dornbrack (2008) find that the effect of these criteria on Xhosa first-language students from disadvantaged educational backgrounds would be dramatic. They state that these criteria will be unfair – and also ironic, in view of the multilingual language policy that the particular university adopted in 2005, and that it is apparent that the educational backgrounds of the students applying to such institutions are not taken into account. Secondly, the authors identified that the use of students' performance in a single language as an indication of "academic literacy in the language of teaching and learning" can be viewed as biased and problematic (Koch & Dornbrack, 2008). This is even more apparent in the linguistically diverse context of SA. The legality of the use of language as a criterion for admissions can also be questioned, especially when the criterion has been proved to be biased (Koch & Dornbrack, 2008).

Finally, the authors state that the results of the case study clearly indicated that in the case of students who enter tertiary education with well-developed academic skills in their first language, this language can be viewed as a better predictor of academic literacy, or the ability to cope with academic work, than achievement or test results in the "language of teaching and learning" (Koch & Dornbrack, 2008). In summary, these authors point out that there is an inherent difficulty with implementing multilingualism at higher education institutions, especially institutions which view English as the primary indicator of success and status. Therefore if institutions of higher learning in SA want to redress past inequalities, and ultimately provide equal access to all (who are capable of succeeding in these institutions), they cannot be allowed to implement monolingual admission criteria.

Additional work in the area of monolingual testing for admission to university by Koch (2005, 2007a, 2007b), demonstrates evidence of high levels of item and construct bias across Afrikaans, English and Xhosa first-language speakers on a reading test in English. The implications for fair admission into university are empirically demonstrated, and possible alternatives to monolingual assessment are discussed extensively in Koch (2007b).

There are two important issues that arise when using tests, whether they are monolingual or multilingual measures, across groups (whether cultural or language) namely, bias and equivalence. These issues will now be discussed in greater detail and will be linked to the theoretical framework employed for the present study.

2.6 Theoretical framework of bias and equivalence

2.6.1 Introduction

From the above discussion in the Literature Review Chapter, two important aspects of monolingual assessment (and assessments in two languages) across cultures or language groups have come to the fore, namely, equivalence and bias. Theorists such as Poortinga and Van de Vijver and various others have made a clear theoretical link between test bias and test equivalence, a theoretical approach that greatly assists with the conceptualisation of a comprehensive evaluation of tests for use across groups, regardless of whether the tests are monolingual or available in more than one language version (Koch, 2005). Monolingual tests, accordingly to Van de Vijver and Tanzer (1998) and Poortinga (1989), specify the framework for equivalence as levels of equivalence, and link it to the types of bias. Therefore this theoretical framework will form the framework for the evaluation of the applicability of the adapted English version of the WMLS, across English and Xhosa first-language learners.

The constructs of the taxonomy of equivalence and the three types of bias will now be discussed in order to gain a deeper understanding of the theoretical link between test equivalence and test bias, as proposed by theorists such as Poortinga and Van de Vijver.

2.6.2 The Taxonomy of Equivalence

"Equivalence" is associated with the measurement level at which scores obtained in different cultural groups can be compared, whereas "bias" indicates the presence of factors that challenge the validity of cross-cultural comparison, and threatens equivalence (Van de Vijver & Leung, 1997). Van de Vijver and his co-researcher distinguish between various levels of equivalence; the various levels include *construct, unit,* and *scalar equivalence or full comparability* (Van de Vijver & Leung, 1997).

When an instrument measures different constructs in two cultures, no comparison can be made (Van de Vijver & Leung, 1997). In other words, there is no link between scores obtained in one cultural group and scores obtained in another group. When the same construct is being measured across groups, it is known as *construct equivalence* (also known as structural equivalence). For example, a construct such as "middle class" or "depression" may have different meanings in different cultures (Van de Vijver, 1997). In order to ensure that construct equivalence is achieved, the nomological networks of the instrument in each culture should be investigated.

The next (and higher) level of equivalence is known as *measurement unit equivalence* (Van de Vijver, 1997). Suppose that an intelligence test developed in Europe has been administered in Europe and a translation of it is used in South Africa. The test material may contain various implicit and explicit references to the European culture. These references will put South African subjects at a disadvantage as the material will be unfamiliar. The end result is that the origin of the scales across groups is affected, but the measurement unit will be the same (Van de Vijver, 1997).

The last and highest level of equivalence is known as *scalar equivalence* or *full comparability* (Van de Vijver, 1997). This level of equivalence can be achieved when the measurement instrument has the same measurement unit in each cultural

group, in other words measures the same constructs and has the same origin. The attainment of scalar equivalence is a prerequisite in any cross-cultural or cross-linguistic comparison (Van de Vijver, 1997).

Equivalence cannot exist when bias is found. In other words, an instrument needs to be free of bias in order for it to be considered equivalent. The next section will focus on the three types of bias.

2.6.3 Three types of Bias

"Bias" is a generic term for all nuisance factors threatening the validity of cross-cultural comparisons (Van de Vijver & Leung, 1997). Some of the reasons for bias include poor item translation, inappropriate item content, and lack of standardisation in administration procedures. The researcher will discuss three types of bias, which include construct, item and method bias with regard to cross-cultural research. If bias exists, then scalar equivalence is compromised.

Construct bias occurs when the construct being measured is not identical across cultures. According to Van de Vijver and Leung (1997), studies in Western and non-Western countries have shown that everyday conceptions of intelligence in non-Western countries are broader than the domain covered by most Western intelligence tests.

Construct bias can also be brought about by a lack or overlap in behaviours associated with the construct in the cultures studied. Ho (1996, as cited in Van de Vijver & Leung, 1997) has studied the concept of filial piety, in other words being "a good son or daughter". He finds that the various behaviours associated with being a good son or daughter, such as taking care of one's parents, conforming to their requests, and treating them well, is much broader in China than in most Western countries.

According to Van de Vijver and Leung (1997), construct bias can also arise as a

result of poor sampling of a domain in an instrument. Broad constructs are often represented by a small number of items in a questionnaire or test. Embretson (1983 as cited in Van de Vijver & Leung, 1997) has coined the term "construct underrepresentation" to refer to this insufficient sampling of a behavioural domain. For example, if most items of a measure of coping depict interpersonal situations, the instrument will yield a poor insight into intrapersonal coping mechanisms and will not generalise to instruments with a broader or differently focused item pool (Van de Vijver & Leung, 1997).

Leung and Zhang (1996 as cited in Van de Vijver & Leung, 1997) have identified another form of construct bias. This form refers to the exporting from Western to non-Western countries, and some of the issues examined in these studies have little relevance to non-Western cultures (Leung & Zhan, 1996 as cited in Van de Vijver & Leung, 1997).

It is important to note that even if a construct is well represented in an instrument, there is still no guarantee that there will be no bias in scores. There are two more types of bias to consider, namely method and item bias.

"Method bias" refers to the bias which occurs from the characteristics of the instrument and/or its administration. Differential response styles across groups, such as acquiescence and extremity ratings, can constitute method bias (Van de Vijver & Leung, 1997). A demonstration can be found in the work of Hui and Triandis (1989), as cited in Van de Vijver and Leung (1997). These authors find that Hispanics tended to choose extremes on a 5-point rating scale more often than do White Americans, while no significant cross-cultural differences are found for 10-point scales. Ross and Mirowsky (1984) report more acquiescence and "socially desirable" responses among Mexicans than among Anglo-Americans in a mental health survey.

A common source of method bias in mental testing is differential familiarity with the stimuli used (Van de Vijver & Leung, 1997). Cattell attempted to resolve this

issue by using simple geometric stimuli; he reasoned that tests that use very simple stimuli are free from cultural influences. This led to the development of "culture-free tests" (Cattell, 1940). However, when it was discovered that this assumption was untenable, "culture-fair tests" were developed (Cattel & Cattell, 1963). The stimuli in these tests were developed in such a way that they would not be differentially affected by cultural background (Van de Vijver & Leung, 1997).

Other issues that could lead to method bias are differing response style, differences in physical conditions, and communication problems. According to Van de Vijver and Leung (1997), method bias usually affects scores at the level of the whole instrument. With regard to statistical terms, method bias will be found in the data as a significant effect for a cultural group in a *t* test, or a significant main effect of a cultural group in an analysis of variance (assuming that the method bias is sufficiently large to reach statistical significance). According to Van de Vijver and Leung (1997), significant cross-cultural differences can be a mixture of valid cross-cultural differences and bias effects, in particular method bias.

Lastly, "item bias" refers to measurement artefacts at item level. According to Van de Vijver and Leung (1997) in the Anglo-Saxon literature the term has been largely replaced by *differential item functioning*, or its acronym, DIF. DIF occurs when people from different groups (commonly gender or ethnicity) with the same latent trait (the same ability/skill, for example) have a different probability of giving a certain response on a questionnaire or test (Van de Vijver, 1997).

Item bias can be produced by various sources such as incidental differences in appropriateness of the item content (for example, some items of an educational test are not in the curriculum in one cultural group), inadequate item formulation (for example, complex wording), and inadequate translation (Van de Vijver & Leung, 1997). According to Van de Vijver and Leung (1997), an example of item bias can be found in the European Value Survey (Halman, May 1996, personal

communication). This scale of value orientations contains an item about loyalty, which in the Spanish version shows cross-cultural differences that are very different from those on other items. Furthermore, upon closer examination it appeared that, unlike in other languages, the Spanish word that was used for *loyalty* has the connotation of sexual faithfulness.

According to Van de Vijver and Leung (1997), another source of item bias is incidental inappropriateness of item content. Amirkhan's coping questionnaire has the item *Watched more television than usual* as one of the items measuring avoidance, one of the three coping strategies assessed. When the questionnaire was applied to groups of Sahel dwellers without electricity in their homes, the item had to be removed (Van Haaften & Van de Vijver, 1996).

In summary, in general, bias will lower the level of equivalence. However, in the case of item bias, a distinction is made between uniform and non-uniform bias (DIF) (Van de Vijver & Leung, 1997). "Uniform bias" refers to influences of bias on scores that are more or less the same for all score levels, and "non-uniform bias" or DIF refers to the influences that are not identical for all score levels (Van de Vijver & Leung, 1997).

Uniform and non-uniform bias may be harmless for construct equivalence, as numeric score comparisons across cultures are not permitted anyway, as stated by Van de Vijver and Leung (1997). In addition, uniform bias will not threaten measurement unit equivalence, as unbiased scores at this level of equivalence cannot be directly compared across cultures (because of differences in the origin of each group), and adding a constant to all scores in a single group does not affect this type of equivalence (Van de Vijver & Leung, 1997). However, the introduction of uniform bias to scores that show (or need) scalar equivalence will lead to the loss of this type of equivalence (that is, the loss of scalar equivalence). Furthermore, when a constant is added to all scores in one group but not 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. 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 or cross-linguistic (in the case of this study) score comparisons are likely to generate incorrect results (Van de Vijver & Leung, 1997).

The next chapter will focus on the methodology of this study, namely the design of the study, the participants recruited, the procedures followed, and the statistical processes that were used.

CHAPTER 3

METHODOLOGY

3.1 Design of the Study

It is important to be aware that this study forms part of a much larger study. The larger study consists of various phases, firstly, the adaptation of the original WMLS (English version), into SA English and Xhosa, secondly, the evaluation of the equivalence of the two language versions of the WMLS across the English and Xhosa groups, and thirdly the evaluation of the predictive, construct and content validity of both these adapted versions across English first-language and Xhosa first-language speakers in the SA context. This sub-study focuses specifically on the equivalence of the adapted English version of the WMLS for use across English first-language and Xhosa second-language speakers. The data that was used in this study had been previously collected by the researcher of the main study. The researcher of the current study will indicate what procedures the researcher of the main study used in collecting the data in this section of this paper.

The researcher (of the current study) conducted a comparative quantitative study with the use of Secondary Data (SD). SD or Secondary Data Analysis (SDA) can be defined as the re-analysis of data for the purpose of answering the original research question with better statistical techniques, or answering new questions with the use of old data (Glass, 1976). With regard to this study, the researcher has not re-evaluated the original question posed by the researchers of the main study, and therefore this study will make use of SD and not SDA.

The researchers of the main study have not analysed the data as yet, although the researchers had this sub-study in mind when collecting the data that has been used for this study.

3.2 Procedure

Ethical clearance was received from the Ethics Committee of the University of Port Elizabeth (UPE) which is now Nelson Mandela Metropolitan University (NMMU), in 2005. Psychology Honours students from the UPE received training, and once trained, they collected the data. The data was captured in Excel spreadsheets and combined by the main researcher.

The researchers of the main study received permission from the Eastern Cape Department of Education (DET). Thereafter, they made contact with the principals of the various schools once they (the schools) had been identified. Information sheets as well as Informed Consent forms (Appendix A and B) were given to the principals of these schools to be given to the parents of the learners; the consent forms were translated into Xhosa as well. Only children whose parents completed these forms participated in the study.

3.3 Sampling

According to the researcher of the main study, the sampling constituted convenience purposive sampling, which allowed the researcher to control for a homogeneous group of participants. This ensured that an equal number of males and females from the various types of schools were selected (ex-model C and Ex-DET schools). This ensured that the educational backgrounds of the two groups were controlled for, that is, the two groups consisted of Grade 6 and 7 learners only. This type of sampling also allowed the researcher to control for confounding factors, such as gender and grade.

The researcher chose convenience sampling in conjunction with purposive sampling, firstly as normative data was not collected, and secondly as the researcher was interested in evaluating the test for the specific age groups (Grade 6 and 7 learners) at this stage.

Tables 1 to 5 presents the sample in terms of the two language groups, gender and grade, for each language group, namely English first-language-speaking learners and Xhosa first-language-speaking learners.

Table 1:
Sample per language group

Language Group	**Sample Size (n)**
English	192 (50.60)
Xhosa	197 (49.40)
Total	**389 (100)**

**percentages in parenthesis*

The above table indicates the number of participants from each language group. There are only a few more participants in the Xhosa first-language-speaking group.

Table 2:
Sample per gender

Gender	**Sample Size (n)**
Male	174 (44.73)
Female	215 (55.27)
Total	**389 (100)**

**percentages in parenthesis*

Table 2 disaggregates the sample by gender. This table indicates that the sample consisted of more females than males.

Table 3:
Sample per grade

Grade	Sample Size (n)
Grade 6	180 (46.30)
Grade 7	209 (53.70)
Total	**389 (100)**

**percentages in parenthesis*

Table 3 disaggregates the sample by grade. There are more Grade 7 learners in the sample than Grade 6 learners.

Table 4:
Language group represented by gender

Language Group	Gender		Total
	Male	Female	
English	98 (49.75)	99 (50.25)	197 (100)
Xhosa	76 (39.58)	116 (60.42)	192 (100)
Total	174	215	389 (100)

**percentages in parenthesis*

Table 4 provides a view of the each language group disaggregated by gender.

Table 5:
Language group represented by grade

Language Group	Grade		Total
	Grade 6	Grade 7	
English	82 (42.70)	110 (57.30)	192 (100)
Xhosa	98 (49.75)	99 (50.25)	197 (100)
Total	180	209	389

Table 5 indicates the language groups disaggregated by grade (Grade 6 and 7).

As stated previously (in this section) the use of purposive convenience sampling allows for the control of confounding factors, such as gender and grade. However, despite the use of this type of sampling, there were some discrepancies in maintaining equal numbers for gender and grades (as indicated in Tables 1 to 5). In other words, there were unequal numbers of males and females, and there were unequal numbers of English and Xhosa first language-speakers in each grade (i.e. there were not equal numbers of language speakers in Grade 6 and Grade7).

These discrepancies, especially with regard to the two grades, could possibly impact on the results, as it is assumed that the Grade 7 children will do better on the test than the Grade 6 children, which will be taken into account in the discussion of the results.

3.4 Participants

Because the researcher was using SD, the participants of the main study will remain the same for the current study. The participants consisted of 198 English first-language-speaking learners and 197 Xhosa first-language-speaking learners. The English and Xhosa groups were selected from "ex-model C" and "previously disadvantaged" schools in the Port Elizabeth and Grahamstown areas. The learners were selected from these schools in order to maintain validity for the learners' different educational levels of English as well as the differing levels of their teaching in English.

3.5 Measurement Tool

The measurement tool that was used is the adapted English version of WMLS. The WMLS is a test used to measure academic language proficiency of learners, and has been extensively used in the USA to evaluate Additive Bilingual Education (in its original form that is available in English and Spanish). The WMLS is an individually administered test that takes approximately 40 minutes to administer. The WMLS consists of four sub-tests, namely Picture Vocabulary, Verbal Analogies, Letter-Word Recognition, and Dictation. The test requirements

as well as what each test measures are presented in Table 6. The content of each test was selected to represent important skills needed for language proficiency for a diverse population, covering a broad range of development (2 years to adulthood). The items of the WMLS are not made available as it is a commercially purchased test, and items are therefore confidential.

Table 6:
Test Requirements and Test Measurement of the WMLS

TEST	TEST REQUIREMENTS	MEASURES	RESPONSE STYLE
Picture Vocabulary	Subject names the familiar and unfamiliar pictured objects that involve breadth and depth of School-related knowledge and experience.	Oral language, including language development and lexical knowledge.	Oral (word)
Verbal Analogies	Subject completes oral analogies requiring verbal comprehension and reasoning.	Reasoning using lexical knowledge.	Oral (word)
Letter-Word Identification	Subject reads familiar and unfamiliar letters and words.	Letter-Word Identification skills.	Oral (letter, word, name)
Dictation	Subject responds in writing to questions which require verbal comprehension, knowledge of letter forms, spelling, punctuation, capitalisation, and word usage.	Prewriting Skills (for early items), Ability to respond in writing to a variety of questions.	Motoric (Writing)

3.5.1 Psychometric Properties of the WMLS

There are two important issues that need to be considered when selecting an instrument for the purpose of research, namely reliability and validity.
With regard to the reliability of the WMLS (the American version), internal consistency reliability coefficients (r*n*) and standard errors of measurement (SEMs) were calculated for all English forms and clusters across their range of intended use (Woodcock & Muñoz-Sandoval, 2001). The test reliabilities were calculated by the split-half procedure, using odd and even raw scores, and were corrected for length by the Spearman-Brown formula (Woodcock & Muñoz-Sandoval, 2001). The median reliabilities range from .80 to .93 for the tests from .88 to .96 for the clusters (Woodcock & Muñoz-Sandoval, 2001).

With regard to the validity of the WMLS (the American version), it was evaluated on content, concurrent, as well as construct validity (Woodcock & Muñoz-Sandoval, 2001). Content validity is the extent to which the content of a test represents the domain of content that it is designed to measure. Concurrent validity is the extent to which scores on a test are related to scores on a certain criterion measure which is typically expressed as a correlation coefficient between the test and the criterion (Woodcock & Muñoz-Sandoval, 2001). Low correlations imply that measures are dissimilar, and *vice versa.*

Correlations among the WMLS Normative Update tests and clusters at selected age levels have been done, where most of these tests showed inter-correlations at a moderate level of .4 or above (Woodcock & Muñoz-Sandoval, 2001).

The above-mentioned reliability and validity are based on the original American versions of the WMLS. Therefore it is important to note that no psychometric properties are available for the SA adapted versions (English and Xhosa versions) of the WMLS. The current study forms part of a larger study investigating the psychometric properties for the adapted SA versions of the WMLS.

3.6 Data Analysis

Owing to the use of SD, the researcher has used the existing data (of the main study) to conduct various statistical tests using the statistical program of SPSS (Statistical Program for the Social Sciences). The hypothesis and the statistical tests used to test each of the four objectives, in order to achieve the overall aim of the study, are as follows:

3.6.1 Objective 1: To evaluate the mean group differences between English and Xhosa first-language learners on the English version of the WMLS

Statistical Test: Descriptive statistics were used to illustrate the mean and standard deviations per group. Thereafter inferential statistics were examined namely, the Hotellings'T^2 statistic and the post hoc t-tests. The Hotellings'T^2 statistic is used when there are two groups and there are several dependent variables, in order to test the differences between the groups in terms of all the outcome measures simultaneously. The post hoc t-tests are conducted to indicate where the differences are and whether these differences are significant, if any should arise.

Null Hypothesis: There are no group differences between the English and Xhosa first-language learners with regard to their mean scores on the English version of the WMLS.

3.6.2 Objective 2: To evaluate group differences in terms of the reliability of each subtest between English and Xhosa first-language learners on the English version of the WMLS

Statistical Test: Firstly the Cronbach's Alpha was calculated for each group per subtest in order to test differences in the reliability of the various subtests between English and Xhosa first-language learners on the English version of the WMLS. The equality of reliability was calculated by using the statistic $(1\text{-alpha}_1)/(1\text{-alpha}_2)$, following the approach proposed by Van de Vijver and Leung (1997).

This difference follows an F-distribution with N_1-1 and N_2 -1 degrees of freedom.

Null Hypothesis: There is no significant group difference with regard to the Cronbach's Alpha for each of the subtests between English and Xhosa first-language learners on the English version of the WMLS.

3.6.3 Objective 3: To evaluate group differences in terms of the item characteristics of the test between English and Xhosa first-language learners on the English version of the WMLS

Statistical Test: Objective 3 has been evaluated descriptively, evaluating the item characteristics, namely mean item difficulty as well as mean item discrimination; therefore no null hypothesis has been tested. This specific research objective will be analysed by means of evaluating the item characteristics which include item difficulty level as well as the item discriminating power of each item per subtest for each language group.

According to Foxcroft and Roodt (2009), item analysis adds value to the process of item development of a measure. In other words the purpose of item analysis is to determine whether an item is measuring what it is suppose to measure. Item analysis assists in the process of determining how difficult an item is, whether it discriminates between good and poor performers, whether it is biased against certain groups, and whether there are any other shortcomings in it (Foxcroft & Roodt, 2009). Item difficulty refers to the proportion or percentage of individuals who answer an item correctly (Foxcroft & Roodt, 2001). In other words, the higher the percentage of correct responses the easier an item, and the lower the percentage of correct responses the more difficult an item.

According to Schalfz and Whitney (2005, as cited in Foxcroft & Roodt, 2005), item difficulties that are 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. In other words, the difficulty level of an item restricts the discriminatory power of the item (Foxcroft & Roodt, 2001). Item-total correlations or item discrimination refer to the potential of a good item to discriminate well between the low and high achievers on a particular test. It is expected that individuals who do well on a measure as a whole tend to answer a good item correctly, while individuals who do poorly on a measure as a whole tend to answer a good item incorrectly (Foxcroft & Roodt, 2001). 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.

3.6.4 Objective 4: To evaluate the item bias across English and Xhosa first-language learners, of the English version of the WMLS.

Statistical Test: Logistic regression has been used to evaluate the item bias. According to Swaminathan and Rogers (1990), logistic regression calculates the probability of a correct response to an item using the following model for DIF detection:

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

where the parameters τ_0, τ_1, τ_2, and τ_3 represent the intercept and the weights for the ability, group difference, and the ability and group interaction terms, respectively, θ is ability denoted by the total test score, and g is the group membership, in this case coded as 0 for the reference group and 1 for the focal group.

When using logistic regression (stepwise analysis), three steps are entered: step 1 the total score of subtest as the conditioning variable; step 2 the group membership is added to the analysis; step 3 the interaction of the group membership and the conditioning variable (total score on the subtest).

Following this, the model fit was evaluated as well as the null-hypothesis. This was done by evaluating the differences of the Chi-square by step 1, step 2, and step 3 at 2 degrees of freedom. Owing to the repeated nature of the analyses, and to control for increased type I error, the more stringent criterion of $p < 0.01$ was used at the critical value (at 2 degrees of freedom) of 9.55.

The next step in the interpretation was to evaluate the effect size using R^2 differences between step 1 and step 3. The effect size was categorised into three dimensions;

1) Negligible DIF: $R^2\Delta<0.35$
2) Moderate DIF: $0.035<R^2\Delta<=0.060$
3) Large DIF: $R^2\Delta>0.060$

With regard to the effect size, only moderate and large DIF items were evaluated. The null hypothesis of no-DIF was rejected only for moderate and large DIF effect sizes.

The final step with regard to the interpretation was to determine whether uniform and non-uniform DIF existed, as well as the direction of DIF. This was done by evaluating differences between R^2 from step 1 to step 2 (which indicates uniform DIF) and R^2 from step 2 to step 3 (which indicates non-uniform DIF) (as suggested by Zumbo, 1999). Once uniform or non-uniform DIF had been identified, the next and final step was to determine the direction of these items. The direction of DIF was determined by evaluating the sign of the beta value or that of the τ in the logistic regression equation in the last step of the logistic regression analysis (with all the variables entered). Uniform DIF favoured the reference group (English group) if $\tau_2 <0$, and the focal group (Xhosa) if $\tau_2>0$. Non-uniform DIF favoured high ability (in terms of the total score) members of the reference group, and low ability (in terms of the total score) members of the focal group if the $\tau_3<0$. Non-uniform DIF favoured high ability (in terms of the total score) members of the focal group, and low ability (in terms of the total score) members of the reference group if the $\tau_3>0$. The total score was a

continuous score and therefore the terms “high ability” and “low ability” do not refer to categories on this variable, but rather to a range of scores on the total score.

Null Hypothesis: The probability of scoring 1 on item *i* for all the subscales will be the function of ability only, in other words, there is no item bias across English and Xhosa first-language learners on the subscales of the WMLS.

3.7 Ethical Considerations

3.7.1 The Overall Study

The researcher of the main study used the Adapted English version as well as the adapted Xhosa version of the WMLS, with English first-language learners and Xhosa first-language learners respectively, to collect the data that will be used to evaluate the psychometric properties of both the adapted English and the adapted Xhosa version of the WMLS test (that will be used in the SA context). However, this study will focus on the adapted English version of the WMLS in order to evaluate the equivalence of this version to be used across English first-language learners and Xhosa first-language learners.

The researcher of the main study explored all of the necessary ethical considerations that were required from the NMMU (the former UPE). These considerations have been outlined in the above section under “Procedure” in the Methodology section of this study. The researcher (of this current sub-study) has obtained permission from the researcher of the main study to use the data that had been collected.

3.7.2 Ethics of this study

With the use of SD, voluntary and informed consent is required, but the scope of this study falls within the scope of the main study. Therefore the voluntary and informed consent was obtained by the researchers of the main study.

The researcher of this study ensured that the data was handled only by herself, as well as ensuring that the data was stored in a safe and secure place which was accessible only by herself. No reference to specific children and schools was made, in order to ensure anonymity.

CHAPTER 4

RESULTS

4.1 Introduction

This chapter focuses on the overall aim, to evaluate the equivalence of the adapted English version of the WMLS for use across the English first-language-speaking group and Xhosa first-language-speaking group of the SA population, with a specific focus on reliability and differential item functioning. The overall aim consists of four objectives. Each objective was analysed by means of either descriptive or inferential statistics or both. The statistics that were utilised in this chapter consisted of means, standard deviations, reliability analysis, and logistic regression analysis. These statistical procedures were chosen in order to test the null hypothesis for each specific research objective.

4.2 Evaluating group differences on total mean scores of the language groups

Specific Research Objective 1: to evaluate mean group differences between English and Xhosa first-language-speaking group on the English version of the WMLS. The null hypothesis tested for this objective was: there are no group differences between the English and Xhosa first-language-speaking group with regard to their mean scores on the English version of the WMLS. This specific objective has been evaluated by means of descriptive statistics presented in the table below.

Table 7:
Mean score and standard deviations for the English and Xhosa first-language-speaking group on the various subtests of the WMLS

Language Group	Picture Vocabulary		Verbal Analogies		Letter-word Identification		Dictation	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
English	30.11	3.31	13.66	4.56	46.46	5.80	32.31	4.64
Xhosa	24.67	4.12	8.94	4.47	41.69	8.06	26.07	7.10

The above table indicates that the overall performance of the Xhosa first-language-speaking group was lower than that of the English first-language-speaking group. The standard deviations of both groups were relatively small, indicating that the scores were clustered around the mean score; however, the Xhosa first-language-speaking group displayed a higher standard deviation on the Letter-word Identification and Dictation subtests than the English first-language-speaking group. Figure 1 depicts a graphical representation of the mean scores for the language groups on the various subtests of the WMLS.

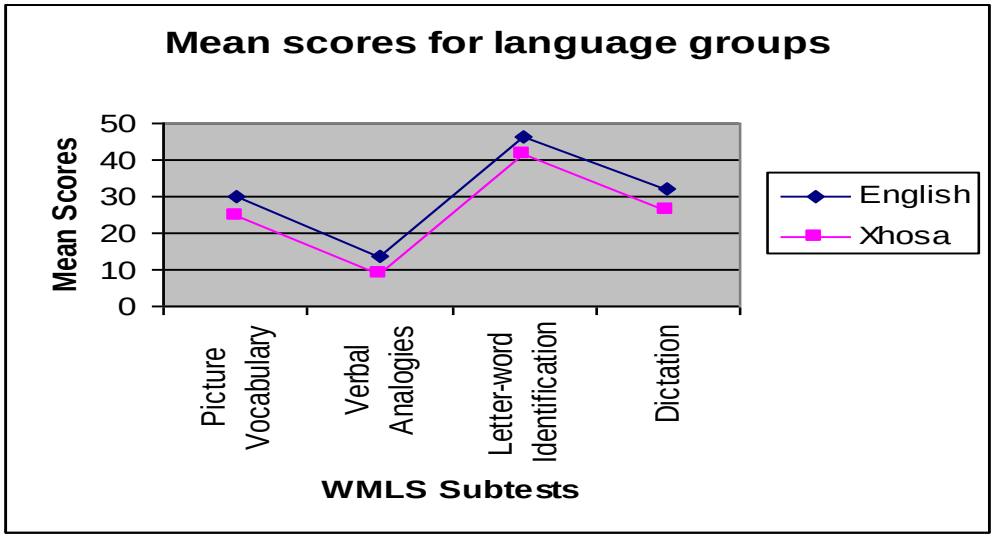

Figure 1: Graphical representation of mean scores of the English and Xhosa first-language-speaking groups

This representation clearly indicates that the Xhosa first-language-speaking group displayed an overall lower mean score than the English first-language-speaking group. However, this does not indicate whether the difference between the English and Xhosa first-language-speaking groups are significant, and therefore a Hotellings'T^2-test was conducted to identify whether the differences between the two language groups are significant.

Table 8:
Hotellings'T²-test results for the English and Xhosa first-language speakers

T^2(casewise MD) =0.554 F (0.55)=53.217[a] p< 0.00				
Subtests	**Mean differences**	**Post-hoc t-value**	**Df**	**p**
Picture Vocabulary	2882.30	205.91	384	0.00
Verbal Analogies	2168.40	106.52	384	0.00
Letter-word Identification	2215.31	44.75	384	0.00
Dictation	3793.31	105.08	384	0.00

The above results of the Hotellings'T²-test indicate that there were significant overall differences between the English and Xhosa first-language-speaking groups on the adapted English version of the WMLS ($p<0.05$). Furthermore the *post hoc* t-test demonstrates that the differences displayed between the two language groups are significant on all of the subtests, and therefore the null hypothesis has been rejected, as group differences do exist. However, these differences are not sufficient evidence to determine that the adapted English version of the WMLS is not equivalent, and further analysis is required.

4.3 Group differences in terms of reliability for the language groups across the subtests

Specific Research Objective 2: to evaluate group differences in terms of reliability on the adapted English version of the WMLS between the English and Xhosa first-language-speaking groups on the English version of the WMLS. The null hypothesis was: there are no significant group differences with regard to the Cronbach's Alpha for each of the subtests between the English and Xhosa first-language-speaking groups. The researcher used the Cronbach's Alpha method to compare the reliability for each language group on the various subtests of the adapted English version WMLS, the results of which are displayed in Table 9.

Table 9:
Indexes of reliability of the English and Xhosa first-language-speaking group for each subtest

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

In the above table, the Cronbach's Alpha was calculated in order to measure the internal consistency of each subtest of the WMLS. 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.

According to the above results, the trend was that the Xhosa first-language-speaking group displayed an overall higher reliability than the English first-language-speaking group on the various subtests of the WMLS. Furthermore the reliability coefficient as displayed by the Cronbach's alpha indicated a good internal consistency for both language groups as it was above .80, except for the English first-language-speaking group on the Picture Vocabulary subtest (0.73).

In addition to the above analysis, both the English first-language-speaking group and the Xhosa first-language-speaking group were compared on their equality of reliability by using the following statistic (1-alpha1)/(1-alpha2). (This method has been outlined in the methodology chapter). This statistic was used in conjunction with the critical value of 1.26. Table 10 indicates the equality of reliability figures.

Table 10:
The test of equality of reliability for the two language groups

	Picture Vocabulary	Verbal Analogies	Letter-word Identification	Dictation
Test of equal reliability	1.46	1.24	0.13	2.05
p-value	<0.05	>0.05	>0.05	<0.05

The above table indicates that there were significant differences between the English first-language-speaking group and Xhosa first-language-speaking group in terms of their reliability on the Picture Vocabulary and Dictation subtests. There were no differences between the two language groups on reliability on the remaining subtests (Verbal Analogies and Letter-word Identification subtests). The Verbal Analogies and Letter-word Identification subtests are thus similar in their equality of reliability for both language groups.

The above outlined significant differences indicate that the null hypothesis tested for this Objective is rejected for the Picture Vocabulary and the Dictation subtests, as group differences do exist between the two language groups in terms of their reliability coefficients on these subtests. The null hypothesis is not rejected for the Verbal Analogies and Letter-word Identification subtests as no group differences exist between the two language groups with regard to their reliability coefficients on these subtests.

However, while these results are indicating problems with regard to the reliability of the Picture Vocabulary and Dictation subtests, at this stage the findings are not sufficient to conclude that the English version of the WMLS is equivalent for the other two subtests (i.e. Verbal Analogies and Letter-word Identification) across the English first-language-speaking group and the Xhosa first-language-speaking group, and further analyses on all the subtests need to be conducted.

4.4 Group differences on item characteristics for the four subtests between the English first-language-speaking group and the Xhosa first-language-speaking group.

Specific Research Objective 3: to evaluate mean group differences in terms of item characteristics of the WMLS between the English first-language-speaking group and the Xhosa first-language-speaking group. This objective will be evaluated descriptively and therefore no null hypothesis was stated. This objective was investigated by analysing the item characteristics, that is, the mean item difficulties (p-values) and item-total correlations (item discrimination indexes) for each subtest of the WMLS. The findings are presented descriptively in Tables 11 to 16.

Table 11:
Mean item difficulty values and mean standard deviation across the two language groups for each subtest

Language Groups	Picture Vocabulary		Verbal Analogies		Letter-word Identification		Dictation	
	Item difficulty	*Std dev.	Item difficulty	Std dev.	Item difficulty	Std dev.	Item difficulty	Std dev.
English	0.52	0.14	0.39	0.31	0.82	0.20	0.58	0.20
Xhosa	0.43	0.15	0.25	0.26	0.73	0.26	0.46	0.25

**Standard deviation*

Table 11 indicates the mean item difficulty values and standard deviations for the two groups per subtest. In light of this, the Picture Vocabulary subtest indicates a higher item difficulty value for the English first-language-speaking group (0.52) than the Xhosa first-language-speaking group (0.43). Furthermore the Verbal Analogies subtest seems to be more difficult for the Xhosa first-language-speaking group (0.25) than for the English first-language-speaking group (0.39). With regard to the Dictation subtest, the English first-language-speaking group displayed a higher difficulty value (0.58) than the Xhosa first-language-speaking

group (0.46). Therefore the trend is that the various subtests were more difficult for the Xhosa first-language-speaking group than for the English first-language-speaking group.

With regard to the above mean standard deviation of the item difficulty scores, the trend is that the standard deviation values displayed on the Picture Vocabulary, Letter-word Identification and Dictation subtests (0.14, 0.20 and 0.20 respectively) are smaller for the English first-language-speaking group than for the Xhosa first-language-speaking group (0.15, 0.26 and 0.25 respectively), but on the remaining subtests, the Verbal Analogies subtest, the Xhosa first-language-speaking group displayed a lower variance (0.26) than the English first-language-speaking group (0.31). The standard deviation results invariably indicate the variances from the mean scores on a test. With regard to the above standard deviation results, the English first-language-speaking group had a lower variance than the Xhosa first-language-speaking group with regard to the item difficulties on the various subtests.

However, the difficulty value is basically a behavioural measure related to the frequency of choosing the correct response. and it therefore does not indicate anything about the intrinsic characteristics of the item itself, also in terms of similarity of item characteristics across the language groups (Foxcroft & Roodt, 2001). Therefore further analyses are necessary. The following table will look at the item-total characteristics to determine the discrimination power of the overall subtests.

Table 12:
Mean item-total correlations (Item discrimination) across the language groups for each subtest

Language Groups	Picture Vocabulary	Verbal Analogies	Letter-word Identification	Dictation
English	0.13	0.32	0.26	0.19
Xhosa	0.15	0.34	0.35	0.30

The above table indicates that the mean item-total correlations (item discrimination) for Picture Vocabulary and Verbal Analogies are similar for both language groups; however, the mean item-total correlations (discriminating power) are different between the language groups with regard to the Letter-word Identification. The mean item-total correlations for Picture Vocabulary for both groups are lower than 0.20. The mean item-total correlation of this subscale is below the acceptable range.

With regard to the English first-language-speaking group, the mean item total correlation of the Dictation subtest indicates that the items of this subtest generally do not discriminate well between the high and low achievers of this language group (mean item-total correlation, 0.19). However, the Xhosa first-language-speaking group displayed an acceptable mean item total correlation (0.30).

In summary, the Picture Vocabulary subtest was easier for the English first-language-speaking group than the Xhosa first-language-speaking group, though both language groups displayed poor mean discrimination on this subtest. The Verbal Analogies subtest was easier for the English first–language-speaking group than it was for the Xhosa first-language-speaking group, and in both groups the items displayed a satisfactory mean discrimination level. With regard to the Letter-word Identification subtest, this subtest was easier for the English first-language-speaking group than for the Xhosa first-language group, and in both language groups the items displayed a satisfactory mean discrimination level. With regard to the final subtest, Dictation, this subtest was easier for the English

first-language-speaking group than for the Xhosa first-language-speaking group, but the items in the Xhosa first-language-speaking group displayed a good mean discrimination level while in the English first-language-speaking group the items did not display a good mean discrimination level.

The item characteristics for each individual item per subtest will be presented in Tables 13 to 16.

Table 13:
Item difficulties (p-values) and item-total correlations (item discrimination values) for each language group on the Picture Vocabulary subtest

Items	**Item difficulty (p-values)**		**Item discrimination (item-total correlations)**	
	English	**Xhosa**	**English**	**Xhosa**
1	1.00	1.00	0.00	0.00
2	1.00	1.00	0.00	0.00
3	1.00	1.00	0.00	0.00
4	1.00	1.00	0.00	0.00
5	1.00	1.00	0.00	0.00
6	1.00	1.00	0.00	0.00
7	1.00	1.00	0.00	0.00
8	1.00	0.99	0.00	0.01
9	1.00	1.00	0.00	0.00
10	1.00	0.99	0.00	0.12
11	1.00	0.96	0.00	0.26
12	1.00	0.99	0.00	0.05
13	1.00	1.00	0.00	0.00
14	1.00	1.00	0.00	0.00
15	0.81	1.00	0.17	0.00
16	0.99	1.00	0.07	0.00
17	1.00	0.96	0.00	0.25
18	1.00	0.97	0.00	0.10
19	0.96	0.64	0.23	0.51
20	0.93	0.50	0.23	0.52
21	0.91	0.54	0.20	0.52
22	0.94	0.56	0.12	0.61

Items	Item difficulty (p-values)		Item discrimination (item-total correlations)	
	English	Xhosa	English	Xhosa
23	0.93	0.38	0.16	0.44
24	0.91	0.62	0.09	0.48
25	0.73	0.44	0.39	0.54
26	0.78	0.21	0.39	0.33
27	1.00	0.81	0.00	0.42
28	0.65	0.37	0.41	0.37
29	0.72	0.15	0.21	0.41
30	0.34	0.66	0.26	0.36
31	0.76	0.28	0.54	0.63
32	0.36	0.35	0.52	0.43
33	0.39	0.07	0.31	0.31
34	0.11	0.08	0.34	0.30
35	0.34	0.10	0.28	0.36
36	0.18	0.02	0.43	0.14
37	0.19	0.02	0.44	0.20
38	0.01	0.00	-0.00	0.00
39	0.02	0.01	0.17	0.04
40	0.01	0.00	0.24	0.00
41	0.11	0.00	0.36	0.00
42	0.02	0.00	0.23	0.00
43	0.01	0.00	0.22	0.00
42	0.01	0.00	0.09	0.00
45	0.00	0.00	0.00	0.00
46	0.01	0.00	0.24	0.00
47	0.00	0.00	0.00	0.00
48	0.00	0.00	0.00	0.00
49	0.00	0.00	0.00	0.00
50	0.00	0.00	0.00	0.00
51	0.00	0.00	0.00	0.00
52	0.00	0.00	0.00	0.00
53	0.00	0.00	0.00	0.00
54	0.00	0.00	0.00	0.00
55	0.00	0.00	0.00	0.00
56	0.00	0.00	0.00	0.00

Items	Item difficulty (p-values)		Item discrimination (item-total correlations)	
	English	Xhosa	English	Xhosa
57	0.00	0.00	0.00	0.00
58	0.00	0.00	0.00	0.00

With regard to the above results, on the whole the English first-language-speaking group performed better on the items of the Picture Vocabulary subtest than the Xhosa first-language-speaking group. The two groups performed relatively similarly until item 18, but from item 19 to item 33 the English group performed better (with more of the groups getting the items correct). However, more first-language speakers of the Xhosa group got item 30 correct than the English first-language-speaking group, in other words indicating that this item was easier for the Xhosa first-language-speaking group than for the English first-language-speaking group.

As mentioned above, item difficulty values below 0.3 are viewed as very difficult items. With regard to the above results, it is indicated that items 26, 29, 31, and 33 were more difficult for the Xhosa first-language-speaking group than for the English first-language-speaking group. Item 30 is indicated as being more difficult for the English first-language-speaking group than for the Xhosa first-language-speaking group. These differential item difficulty values could indicate possible DIF items.

With regard to the above results and specifically to the items that the language groups got correct, items 1 to 10 are items which do not discriminate well for both language groups. In other words, these items do not discriminate between the high and low achievers within the given group, either the English first-language speakers or Xhosa first-language speakers). While items 11, 17, 22, 23, 24, and 27 do not discriminate well between the English first-language-speaking group, these items discriminate well in the Xhosa first-language-speaking group. On the other hand, items 40, 41, 42, and 46, do not discriminate well between the high and low achievers in the Xhosa first-language-speaking group, though these items

indicate a good discrimination between the high and low achievers within the English first-language-speaking group.

Table 14:
Item difficulties (p-values) and item-total correlations (item discrimination values) for each language group on the Verbal Analogies subtest

Items	Item difficulty (p-values)		Item discrimination (item-total correlations)	
	English	Xhosa	English	Xhosa
1	0.83	0.72	-0.02	0.31
2	0.95	0.95	0.06	0.01
3	0.89	0.85	0.20	0.17
4	0.92	0.72	0.27	0.33
5	0.93	0.92	0.09	0.25
6	0.94	0.93	0.14	0.24
7	0.90	0.57	0.32	0.60
8	0.78	0.28	0.37	0.56
9	0.71	0.19	0.30	0.55
10	0.68	0.45	0.35	0.53
11	0.81	0.34	0.40	0.63
12	0.69	0.28	0.38	0.51
13	0.68	0.34	0.44	0.67
14	0.54	0.37	0.38	0.61
15	0.33	0.12	0.36	0.30
16	0.33	0.12	0.46	0.42
17	0.30	0.12	0.49	0.41
18	0.24	0.19	0.36	0.47
19	0.16	0.07	0.47	0.41
20	0.20	0.05	0.55	0.38
21	0.11	0.04	0.41	0.48
22	0.11	0.03	0.42	0.27
23	0.13	0.03	0.58	0.29
24	0.09	0.02	0.48	0.33
25	0.11	0.02	0.51	0.31
26	0.07	0.01	0.38	0.18
27	0.03	0.01	0.27	0.20
28	0.10	0.05	0.51	0.47

Items	Item difficulty (p-values)		Item discrimination (item-total correlations)	
	English	Xhosa	English	Xhosa
29	0.04	0.02	0.30	0.32
30	0.01	0.00	0.18	0.00
31	0.01	0.01	0.13	0.17
32	0.04	0.01	0.29	0.25
33	0.01	0.01	0.16	0.17
34	0.01	0.00	0.13	0.00
35	0.00	0.00	0.00	0.00

With regard to the above results, the English first-language-speaking group predominantly did better on the Verbal Analogies subtest than the Xhosa first-language-speaking group (from item 7). In other words, more of the English-language group got items 1 to 14 correct (seemingly easier items) than the Xhosa first-language-speaking group. With regard to items 8, 9, 12, 15, 16 and 17, the Xhosa first-language-speaking group displayed an item difficulty level below 0.30 and these items were much more difficult for this group than for the English first-language-speaking group. In addition, from item 18 the items become increasingly more difficult for both language groups.

With regard to discriminating power, items 1, 2, 30, 31, 33, 34 and 35 displayed a low discriminating power for both language groups. Items 3 and 26 are items that did not discriminate well between the high and low achievers of the Xhosa first-language-speaking group but did for the English first-language-speaking group. Items 5 and 6 discriminated better between the high and low achievers in the Xhosa first-language-speaking group than between the high and low achievers in the English first-language-speaking group.

Several reasons can be suggested for these differences in results. These reasons could range from genuine differences in ability between the language groups and cultural differences between the two groups, to possible DIF existence, therefore further exploration is necessary to determine what exactly the reasons are for these differences.

With regard to the Verbal Analogies subtest, both language groups displayed an acceptable item discrimination value overall. However, there are certain items which displayed a lower discriminating power. With regard to items 1, 5 and 6, the English first-language-speaking group displayed a lower discriminating power than the Xhosa first-language-speaking group on this subtest, while the Xhosa first-language-speaking group displayed a lower discriminating power on item 26 than the English first-language-speaking group. The low discriminating power (of these items) essentially makes it difficult to discriminate between the low and high achievers in each language group. Thus further analysis needs to be conducted on the various items identified by the item characteristics as being problematic, to determine whether these items are biased.

Table 15:
Item difficulties (p-values) and item-total correlations (item discrimination values) for each language group on the Letter-word Identification subtest

Items	Item difficulty (p-values)		Item discrimination (item-total correlations)	
	English	Xhosa	English	Xhosa
1	1.00	1.00	0.00	0.00
2	1.00	1.00	0.00	0.00
3	1.00	1.00	0.00	0.00
4	1.00	1.00	0.00	0.00
5	1.00	1.00	0.00	0.00
6	1.00	1.00	0.00	0.00
7	1.00	1.00	0.00	0.00
8	1.00	1.00	0.00	0.00
9	1.00	0.99	0.00	0.21
10	1.00	1.00	0.00	0.00
11	.99	0.99	0.16	0.23
12	1.00	1.00	0.00	0.00
13	1.00	1.00	0.00	0.09
14	1.00	1.00	0.00	0.00
15	1.00	0.99	0.00	0.23
16	1.00	0.99	0.00	0.26
17	1.00	0.95	0.00	0.28
18	1.00	0.99	0.00	0.23

Items	Item difficulty (p-values)		Item discrimination (item-total correlations)	
	English	Xhosa	English	Xhosa
19	1.00	0.99	0.00	0.30
20	1.00	0.96	0.00	0.36
21	0.99	0.98	0.23	0.24
22	1.00	0.90	0.00	0.43
23	0.99	0.94	0.35	0.44
24	0.99	0.97	0.27	0.33
25	1.00	0.89	0.00	0.50
26	0.98	0.88	0.37	0.55
27	0.99	0.96	0.27	0.39
28	1.00	0.89	0.00	0.52
29	0.99	0.87	0.11	0.55
30	0.98	0.82	0.13	0.61
31	0.93	0.56	0.39	0.61
32	0.93	0.82	0.38	0.65
33	0.96	0.85	0.41	0.49
34	0.96	0.89	0.37	0.46
35	0.92	0.91	0.33	0.24
36	0.94	0.85	0.41	0.59
37	0.90	0.55	0.43	0.69
38	0.83	0.83	0.50	0.59
39	0.92	0.48	0.44	0.53
40	0.80	0.32	0.49	0.44
41	0.51	0.24	0.55	0.34
42	0.44	0.44	0.37	0.61
43	0.68	0.51	0.53	0.59
42	0.63	0.41	0.55	0.60
45	0.65	0.60	0.63	0.64
46	0.79	0.68	0.67	0.62
47	0.68	0.35	0.60	0.50
48	0.51	0.21	0.60	0.27
49	0.40	0.07	0.55	0.22
50	0.42	0.44	0.53	0.61
51	0.50	0.42	0.58	0.58
52	0.48	0.19	0.59	0.31

Items	Item difficulty (p-values)		Item discrimination (item-total correlations)	
	English	**Xhosa**	**English**	**Xhosa**
53	0.18	0.21	0.42	0.32
54	0.21	0.39	0.35	0.59
55	0.26	0.29	0.49	0.53
56	0.14	0.10	0.40	0.31
57	0.07	0.01	0.29	0.03

With regard to the above results, both language groups performed similarly from item 1 to item 28. However items, 41, 48, 49, and 52 were more difficult for the Xhosa first-language-speaking group than for the English first-language-speaking group. On the other hand despite items 53 and 54 being difficult for both language groups, they were more difficult for the English first-language-speaking group than for the Xhosa first-language-speaking group.

With regard to the discriminating power of items 1, 2, 3, 4, 5, 6, 7, 8, 10, 12, 13, and 14, these items displayed a low discriminating power in both language groups, although it should be mentioned that this particular test was not standardised for language groups, rather it was standardised to be administered to individuals from age 2 to 99 years old. In addition, items 9, 11, 15, 16, 17, 18, 19, 20, 22, 25, 28, 29, and 30 displayed a lower discriminating power in the English first-language group than in the Xhosa first-language group. In other words, these items discriminate well between the low and high achievers in the Xhosa first-language group, while they do not discriminate between the high and low achievers in the English first-language group. This therefore indicates that individuals from the Xhosa first-language-speaking group who did well on the overall test got these items correct, and vice versa (that is if they did poorly on the test overall, they were likely to get these items incorrect). On the other hand, it would not be as easy to distinguish between the individuals of the English first-language-speaking group, because these items are poor discriminators (that is should an individual have done well on the test overall, it does not necessitate that that individual would have got any of the above items correct).

Table 16:
Item difficulties (p-values) and item-total correlations (item discrimination values) for each language group on the Dictation subtest

Items	Item difficulty (p-values)		Item discrimination (item-total correlations)	
	English	Xhosa	English	Xhosa
1	1.00	1.00	0.00	0.00
2	1.00	1.00	0.00	0.00
3	1.00	1.00	0.00	0.00
4	1.00	0.99	0.00	-0.07
5	1.00	0.99	0.00	-0.07
6	1.00	0.99	0.00	-0.07
7	1.00	1.00	0.00	0.00
8	1.00	0.99	0.00	0.09
9	1.00	0.99	0.00	0.10
10	1.00	0.80	0.00	0.33
11	1.00	0.97	0.00	0.25
12	1.00	0.98	0.00	0.14
13	0.98	0.94	0.18	0.06
14	0.99	0.96	0.13	0.27
15	0.99	0.95	0.12	0.25
16	0.99	0.97	0.13	0.22
17	0.99	0.88	0.10	0.40
18	0.88	0.61	0.38	0.57
19	0.95	0.50	0.20	0.59
20	0.98	0.90	0.03	0.31
21	0.99	0.80	0.11	0.45
22	0.98	0.53	0.18	0.60
23	0.85	0.46	0.33	0.55
24	0.90	0.47	0.29	0.59
25	0.93	0.70	0.32	0.60
26	0.63	0.33	0.31	0.61
27	0.97	0.64	0.08	0.56
28	0.57	0.47	0.41	0.61
29	0.46	0.27	0.57	0.56
30	0.69	0.47	0.35	0.62
31	0.52	0.20	0.22	0.45
32	0.14	0.14	0.31	0.36

Items	Item difficulty (p-values)		Item discrimination (item-total correlations)	
	English	Xhosa	English	Xhosa
33	0.63	0.34	0.60	0.63
34	0.69	0.39	0.54	0.62
35	0.38	0.20	0.40	0.43
36	0.16	0.09	0.19	0.25
37	0.47	0.30	0.47	0.52
38	0.30	0.10	0.34	0.47
39	0.19	0.07	0.33	0.40
40	0.35	0.13	0.51	0.49
41	0.15	0.06	0.38	0.38
42	0.10	0.04	0.34	0.29
43	0.28	0.12	0.54	0.46
42	0.01	0.01	0.09	0.15
45	0.08	0.05	0.29	0.35
46	0.09	0.05	0.28	0.36
47	0.02	0.01	0.28	0.18
48	0.01	0.01	0.14	0.07
49	0.01	0.02	0.01	0.26
50	0.02	0.01	0.16	0.15
51	0.01	0.00	0.14	0.00
52	0.01	0.01	0.03	0.12
53	0.00	0.01	0.00	0.11
54	0.00	0.00	0.00	0.00
55	0.00	0.00	0.00	0.00
56	0.00	0.01	0.00	0.14

With regard to the above results, both groups performed relatively similarly until item 28. Items 29, 31, 35, 38 and 40 were more difficult for the Xhosa first-language-speaking group (<0.30) than the English first-language-speaking group.

With regard to the discrimination power, items 1, 2, 3, 7, 8, 9, 12, 13, 42, 48, 50, 51, 52, 53, 54, 55, and 56 have a discrimination value of <0.20 for both groups and therefore do not discriminate well between high and low achievers for both

the English first-language-speaking group and the Xhosa first-language-speaking group.

In addition, items 4, 5, and 6 displayed a negative discrimination value for the Xhosa first-language-speaking group. A negative discrimination value indicates that these items are measuring a trait which is the opposite of what it is intended to measure. In addition, this negative value might suggest that an error was made in the scoring of the item or that the item was poorly worded (Allen & Yen 1979). According to Allen and Yen (1979), an item with a negative or low value should be improved or eliminated. In addition, item 47 had a lower discriminating value for the Xhosa first-language-speaking group than for the English first-language-speaking group, therefore this item does not discriminate well between high and low achievers in the Xhosa first-language-speaking group.

With regard to the English first-language-speaking group, this group displayed a low discriminating power (<0.20) on items 10, 11, 14, 15, 16, 17, 20, 21, 22, 27, 36 and 49 in comparison to the Xhosa first-language-speaking group. Therefore these items (with low discriminating values) do not discriminate well between the low and high achievers in the English first-language group.

The above results of item characteristics indicate various differences between the English first-language-speaking group and the Xhosa first-language-speaking group. These differences can be attributed to various reasons, such as differences in ability levels between the two language groups, possible cultural differences, or various other biases (such as item bias or Differential Item Functioning-DIF). Therefore a further analysis of the items needs to be conducted to determine from where these differences arise.

4.5 Item Bias (Differential item functioning) displayed on the four subtests for both language groups

Specific Research Objective 4: to evaluate the item bias across the English and Xhosa first-language-speaking group for all the subtests of the WMLS. The null hypothesis tested for this objective is: the probability of scoring 1 on item *i* for all the subtests will be the function of ability only, in other words there will be no item bias across English and Xhosa first-language-speaking groups on the subtests of the WMLS. This specific research objective was evaluated by means of logistic regression analysis which was conducted on each subtest.

Tables 17 to 24 present the results of the DIF analyses of all the subtests of the English version of the WMLS across the two language groups in the study. Table 17 presents the summary of the stepwise logistic regression DIF procedure for the Picture Vocabulary subtests for both language groups, while Table 18 presents the direction of DIF for the Picture Vocabulary subtest for both language groups.

Table 17:
Summary of Stepwise Logistic Regression DIF procedure: Picture Vocabulary

Items	Stepwise R^2			DIFF* Chi Square	DIFF* R^2	Size of DIF
	Step 1	Step 2	Step 3			
8	0.06	0.11	0.11	0.65	0.05	No DIF
10	0.36	0.36	0.36	0.02	0.00	No DIF
11	0.49	0.49	0.49	0.18	0.00	No DIF
12	0.19	0.21	0.21	0.16	0.02	No DIF
15	0.01	0.37	0.37	70.77	0.36	Large
16	0.01	0.22	0.22	2.99	0.21	No DIF
17	0.46	0.47	0.47	0.17	0.01	No DIF
18	0.16	0.20	0.20	2.38	0.04	No DIF
19	0.55	0.56	0.56	1.58	0.01	No DIF
20	0.56	0.58	0.58	9.03	0.02	No DIF
21	0.53	0.53	0.53	2.22	0.00	No DIF
22	0.60	0.61	0.62	9.28	0.02	No DIF
23	1	0.61	0.62	34.57	0.38	Large

Items	Stepwise R^2			DIFF* Chi Square	DIFF* R^2	Size of DIF
	Step 1	Step 2	Step 3			
24	0.41	0.41	0.42	4.59	0.01	No DIF
25	0.46	0.48	0.48	6.45	0.02	No DIF
26	0.51	0.57	0.58	30.34	0.07	Large
27	0.48	0.52	0.52	10.05	0.04	Moderate
28	0.37	0.37	0.39	7.34	0.02	No DIF
29	0.49	0.56	0.57	37.27	0.08	Large
30	0.03	0.34	0.34	107.05	0.31	Large
31	0.72	0.72	0.72	2.41	0.00	No DIF
32	0.29	0.43	0.45	60.91	0.16	Large
33	0.40	0.42	0.42	5.56	0.02	No DIF
34	0.28	0.34	0.35	14.91	0.07	Large
35	0.37	0.37	0.37	1.90	0.00	No DIF
36	0.43	0.43	0.46	5.31	0.03	No DIF
37	0.52	0.52	0.52	0.16	0.00	No DIF
38	0.05	0.11	0.11	0.81	0.06	No DIF
39	0.18	0.19	0.20	1.42	0.02	No DIF
40	1.00	1.00	1.00	0	0	No DIF
41	0.48	0.51	0.51	5.17	0.03	No DIF
42	0.51	0.51	0.51	0.26	0.00	No DIF
43	0.56	0.56	0.56	0.13	0.00	No DIF
44	0.23	0.25	0.25	0.28	0.02	No DIF
46	1.00	1.00	1.00	0	0	No DIF

The above table indicates that eight items display DIF, of which seven items (15, 23, 26, 29, 30, 32 and 34) had a large effect size, $R^2\Delta>0.06$. Furthermore, one item (27) showed a moderate effect size of $0.035<=R^2\Delta0.06$. There were no items with negligible DIF. The null hypothesis of "no DIF" was rejected for the items that displayed moderate and large DIF.

Table 18:
Direction of DIF for Picture Vocabulary subtest

Item	Variables in equation	Beta	Direction
15	Total	0.24	Uniform DIF. Xhosa first Language speaking group favoured.
	Language group	26.77	
	Interaction term	-0.24	
	Constant	-5.57	
23	Total	0.31	Non-uniform DIF. HA Xhosa first language group is favoured, LA English first Language Speaking group favoured.
	Language group	-2.29	
	Interaction term	0.01	
	Constant	-6.51	
26	Total	0.50	Non-uniform DIF. HA English First language speaking group, LA Xhosa first Language Speaking group favoured.
	Language group	4.81	
	Interaction term	-0.23	
	Constant	-13.18	
27	Total	0.00	Uniform DIF. HA. English first language speaking favoured.
	Language group	-29.26	
	Interaction term	0.40	
	Constant	21.20	
29	Total	0.26	Uniform DIF. English first language speaking
	Language group	-7.97	
	Interaction term	0.22	
	Constant	-6.65	

Item	Variables in equation	Beta	Direction
			group.
30	Total	0.29	Uniform DIF. Xhosa first Language Speaking group favoured.
	Language group	2.99	
	Interaction term	0.01	
	Constant	-9.65	
32	Total	0.66	Non-Uniform Bias. HA English First language speaking group, low ability Xhosa first language speaking group favoured.
	Language group	11.65	
	Interaction term	-0.32	
	Constant	-20.89	
34	Total	0.60	Non-uniform DIF. HA English First language speaking group favoured, LA Xhosa first language speaking group favoured.
	Language group	7.55	
	Interaction term	-0.19	
	Constant	-21.57	

**HA-high ability and LA-low ability*

The above table indicates that four items displayed uniform DIF and four items displayed non-uniform bias. Two items with large DIF (15 and 30), displayed uniform DIF. These items favoured the Xhosa first-language-speaking group. In addition item 29 (with large DIF) displayed uniform DIF, which favoured the English first-language-speaking group. Item 27 had moderate DIF which

displayed uniform DIF. This item favoured the English first-language-speaking group.

In addition, items 23, 26, 32 and 34 had large DIF and displayed non-uniform DIF. Items 26, 32 and 34 favoured the high ability English-first-language speaking group and low ability Xhosa first-language-speaking group, while item 23 favoured the high ability Xhosa first-language-speaking group and low ability English first-language-speaking group.

Table 19 presents the summary of the stepwise logistic regression DIF procedure for the Verbal Analogies subtests for both language groups, while Table 20 presents the direction of DIF for the Verbal Analogies subtest for both language groups.

Table 19:
Summary of Stepwise Logistic Regression DIF procedure: Verbal Analogies

Items	Stepwise R^2			DIFF* Chi Square	DIFF* R^2	Size of DIF
	Step 1	Step 2	Step 3			
1	0.12	0.12	0.12	13.57	0.00	Negligible
2	0.02	0.08	0.03	1.01	0.01	No DIF
4	0.35	0.35	0.35	0.41	0.00	No DIF
5	0.11	0.14	0.17	10.12	0.06	Moderate
6	0.15	0.18	0.20	9.16	0.05	No DIF
7	0.63	0.63	0.64	4.30	0.01	No DIF
8	0.57	0.61	0.61	21.15	0.04	Moderate
9	0.50	0.57	0.58	37.52	0.08	Large
10	0.42	0.43	0.43	4.26	0.01	No DIF
11	0.64	0.65	0.66	9.04	0.02	No DIF
12	0.49	0.50	0.50	8.11	0.01	No DIF
13	0.62	0.62	0.62	2.92	0.00	No DIF
14	0.43	0.43	0.43	14.44	0.00	Negligible
15	0.30	0.31	0.31	1.75	0.01	No DIF
16	0.43	0.43	0.43	0.70	0.00	No DIF
17	0.43	0.43	0.43	0.57	0.00	No DIF

Items	Stepwise R^2			DIFF* Chi Square	DIFF* R^2	Size of DIF
	Step 1	Step 2	Step 3			
18	0.29	0.32	0.33	10.88	0.04	Moderate
19	0.42	0.43	0.43	1.67	0.01	No DIF
20	0.53	0.53	0.53	1.20	0.00	No DIF
21	0.49	0.50	0.54	9.76	0.05	No DIF
22	0.39	0.39	0.39	0.19	0.00	No DIF
23	0.61	0.61	0.62	2.30	0.01	No DIF
24	0.57	0.57	0.57	0.50	0.00	No DIF
25	0.55	0.56	0.56	0.08	0.01	No DIF
26	0.42	0.43	0.43	0.61	0.01	No DIF
27	0.44	0.44	0.49	3.01	0.05	No DIF
28	0.59	0.61	0.61	3.96	0.02	No DIF
29	0.44	0.46	0.49	3.96	0.05	No DIF
30	0.21	0.26	0.34	3.22	0.13	No DIF
31	0.33	0.39	0.41	2.14	0.08	No DIF
32	0.41	0.41	0.42	1.22	0.01	No DIF
33	0.39	0.49	0.49	2.48	0.10	No DIF
34	0.35	0.37	0.37	0.27	0.02	No DIF

The above table indicates that six items display DIF, of which two items (8 and 9) had a large effect size, $R^2\Delta>0.06$. In addition, two items of the six identified with DIF (5 and 18) showed a moderate effect size of $0.035<=R^2\Delta 0.06$. The remaining two items displayed negligible effect size, $R^2\Delta <0.03$. The null hypothesis of "no DIF" was rejected for the items displaying moderate and large DIF.

Table 20:
Direction of DIF for Verbal Analogies subtest

Item	Variables in the equation	Beta	Direction
5	Total	0.14	Non-uniform Bias. HA Xhosa first language speaking group favoured, LA English first language speaking group favoured.
	Language group	-1.49	
	Interaction term	0.34	
	Constant	0.75	
8	Total	0.40	Uniform DIF. English first language speaking group are favoured.
	Language group	-1.71	
	Interaction term	0.03	
	Constant	-3.51	
9	Total	0.26	Uniform DIF English first language speaking group favoured.
	Language group	-3.84	
	Interaction term	0.19	
	Constant	-2.39	
18	Total	0.25	Non-uniform DIF. HA Xhosa first language speaking group favoured, LA English first language speaking group favoured.
	Language group	-0.17	
	Interaction term	0.09	
	Constant	-4.80	

*HA-high ability and LA-low ability

With regard to the above table, the two items with moderate DIF (5 and 18) displayed non-uniform DIF. These items favoured the high ability Xhosa first-language-speaking group and the low ability English first-language-speaking group. The remaining two items with large DIF item (8 and 9) displayed uniform DIF. Both of these items favoured the English first-language-speaking group.

Table 21 presents the summary of the stepwise logistic regression DIF procedure for the Letter-word Identification subtests for both language groups, while Table 22 presents the direction of DIF for the Letter-word Identification subtest for both language groups.

Table 21:
Summary of Stepwise Logistic Regression DIF procedure: Letter-Word Identification

Items	Stepwise R^2			DIFF* Chi Square	DIFF* R^2	Size of DIF
	Step 1	Step 2	Step 3			
9	0.40	0.42	0.42	0.54	0.02	No DIF
11	0.25	0.38	0.51	8.74	0.26	No DIF
13	0.10	0.18	0.18	2.44	0.08	No DIF
15	1.00	1.00	1.00	0	0.00	No DIF
16	0.62	0.63	0.63	0.22	0.01	No DIF
17	0.28	0.35	0.35	5.88	0.07	No DIF
18	1.00	1.00	1.00	0	0.00	No DIF
19	1.00	1.00	1.00	0	0.00	No DIF
20	0.48	0.51	0.51	2.32	0.03	No DIF
21	0.37	0.37	0.40	1.61	0.03	No DIF
22	0.41	0.48	0.48	10.61	0.07	Large
23	0.57	0.57	0.62	5.46	0.05	No DIF
24	0.47	0.49	0.49	1.94	0.02	No DIF
25	0.52	0.59	0.59	10.53	0.07	Large
26	0.61	0.61	0.63	4.28	0.02	No DIF
27	0.54	0.55	0.55	1.05	0.01	No DIF
28	0.55	0.60	0.60	9.58	0.05	Moderate
29	0.51	0.53	0.54	5.83	0.03	No DIF
30	0.52	0.55	0.56	9.62	0.04	Moderate
31	0.54	0.61	0.61	30.71	0.07	Large
32	0.57	0.58	0.58	3.16	0.01	No DIF
33	0.50	0.50	0.52	5.05	0.02	No DIF
34	0.44	0.45	0.45	2.17	0.01	No DIF
35	0.39	0.40	0.40	2.27	0.01	No DIF
36	0.55	0.56	0.56	2.02	0.01	No DIF
37	0.58	0.64	0.64	27.85	0.06	Moderate

Items	Stepwise R^2			DIFF* Chi Square	DIFF* R^2	Size of DIF
	Step 1	Step 2	Step 3			
38	0.43	0.50	0.50	22.86	0.07	Large
39	0.50	0.62	0.62	54.30	0.12	Large
40	0.45	0.57	0.57	55.48	0.12	Large
41	0.42	0.44	0.44	15.77	0.02	Negligible
42	0.41	0.46	0.48	28.99	0.07	Large
43	0.51	0.51	0.51	1.48	0	No DIF
44	0.59	0.59	0.59	1.08	0.00	No DIF
45	0.56	0.60	0.61	25.42	0.05	Moderate
46	0.58	0.60	0.63	20.21	0.05	Moderate
47	0.56	0.58	0.58	11.14	0.02	Negligible
48	0.42	0.44	0.50	31.44	0.08	Large
49	0.48	0.56	0.57	33.31	0.09	Large
50	0.49	0.56	0.56	32.73	0.07	Large
51	0.54	0.56	0.57	10.35	0.03	Moderate
52	0.46	0.49	0.52	25.56	0.06	Moderate
53	0.27	0.32	0.37	29.37	0.10	Large
54	0.33	0.52	0.52	78.10	0.19	Large
55	0.47	0.56	0.56	38.79	0.09	Large
56	0.44	0.46	0.47	6.42	0.03	No DIF
57	0.29	0.33	0.35	7.42	0.06	No DIF

The above table indicates that twenty-two items display DIF, of which thirteen (22, 25, 31, 38, 39, 40, 42, 48, 49, 50, 53, 54 and 55) had a large effect size, $R^2\Delta>0.06$. In addition, seven items (28, 30, 37, 45, 46, 51, and 52) showed a moderate effect size of $0.035<=R^2\Delta 0.06$. And the remaining two items with DIF (41 and 47) displayed negligible effect size, $R^2\Delta <0.03$. The null hypothesis was rejected for the large and moderate DIF items.

Table 22:
Direction of DIF for Letter-Word Identification

Item	Variables in equation	Beta	Direction
22	Total	0.00	Uniform DIF English first language speaking group favoured.
	Language group	-26.39	
	Interaction term	0.20	
	Constant	21.20	
25	Total	0.00	Uniform DIF. English first language speaking group.
	Language group	-28.82	
	Interaction term	0.27	
	Constant	21.20	
28	Total	0.00	Uniform DIF. English first language speaking group was favoured.
	Language group	-29.34	
	Interaction term	0.29	
	Constant	21.20	
30	Total	0.14	Uniform DIF. English first language speaking group was favoured.
	Language group	-7.75	
	Interaction term	0.17	
	Constant	-2.21	
31	Total	0.27	Uniform DIF English first language speaking group favoured.
	Language group	-2.12	
	Interaction term	0.00	
	Constant	-9.12	
37	Total	0.28	Uniform DIF English first language speaking group favoured.
	Language group	-3.78	
	Interaction term	0.05	
	Constant	-9.93	
38	Total	0.30	Uniform DIF Xhosa first language speaking group favoured.
	Language group	2.58	
	Interaction term	-0.02	
	Constant	-11.49	

Item	Variables in equation	Beta	Direction
39	Total	0.33	Uniform DIF Xhosa first language speaking group favoured.
	Language group	0.94	
	Interaction term	-0.08	
	Constant	-11.46	
40	Total	0.28	Uniform DIF English first language speaking group favoured.
	Language group	-0.03	
	Interaction term	-0.04	
	Constant	-10.85	
42	Total	0.21	Uniform DIF English first language speaking group favoured.
	Language group	-5.38	
	Interaction term	0.14	
	Constant	-10.36	
45	Total	0.42	Uniform DIF Xhosa first language speaking group favoured.
	Language group	6.31	
	Interaction term	-0.11	
	Constant	-18.71	
46	Total	0.56	Non-uniform DIF HA English first Language speaking group favoured, and LA Xhosa first Language speaking group favoured.
	Language group	13.63	
	Interaction term	-0.30	
	Constant	-23.19	
48	Total	0.44	Non-uniform DIF HA English first Language speaking group favoured, and LA Xhosa first Language
	Language group	13.13	
	Interaction term	-0.30	
	Constant	-20.71	

Item	Variables in equation	Beta	Direction
			speaking group favoured.
49	Total	0.46	Uniform DIF Xhosa first language speaking group favoured.
	Language group	6.84	
	Interaction term	-0.18	
	Constant	-22.35	
50	Total	0.40	Uniform DIF Xhosa first language speaking group favoured.
	Language group	4.39	
	Interaction term	-0.06	
	Constant	-19.28	
51	Total	0.43	Uniform DIF Xhosa first language speaking group favoured.
	Language group	5.93	
	Interaction term	-0.11	
	Constant	-20.05	
52	Total	0.54	Non-uniform DIF HA English first language speaking group favoured, and LA Xhosa first language speaking group favoured.
	Language group	19.41	
	Interaction term	-0.37	
	Constant	-28.63	
53	Total	0.54	Non-uniform DIF HA English first Language speaking group favoured, and LA Xhosa first language speaking group favoured.
	Language group	19.41	
	Interaction term	-0.37	
	Constant	-28.63	
54	Total	0.31	Uniform DIF

Item	Variables in equation	Beta	Direction
	Language group	-0.12	English first language speaking group favoured.
	Interaction term	0.06	
	Constant	-16.53	
55	Total	0.53	Uniform DIF. Xhosa first language speaking group favoured.
	Language group	8.32	
	Interaction term	-0.13	
	Constant	-26.98	

**HA-high ability and LA-low ability*

The items with large DIF (22, 25, 31, 40, 42 and 54), and the items with moderate DIF (28, 30 and 37) displayed uniform DIF. These items favoured the English first-language-speaking group. Furthermore, five items with large DIF (38, 39, 49, 50 and 55), and two moderate items (45 and 51) displayed uniform DIF which favoured the Xhosa first-language-speaking group.

With regard to the remaining two large DIF items (49 and 53), and the remaining two moderate DIF items (46 and 52), these items displayed non-uniform DIF. Furthermore these items favoured the high ability English first-language-speaking group and the low ability Xhosa first-language-speaking group.

Table 23 presents the summary of the stepwise logistic regression DIF procedure for the Dictation subtests for both language groups, while Table 24 presents the direction of DIF for the Dictation subtest for both language groups.

Table 23:
Summary of Stepwise Logistic Regression DIF procedure: Dictation subtest

Items	Stepwise R^2			DIFF* Chi Square	DIFF* R^2	Size of DIF
	Step 1	Step 2	Step 3			
4	0.09	0.25	0.25	2.15	0.16	No DIF
5	0.09	0.25	0.25	2.15	0.16	No DIF
6	0.09	0.25	0.25	2.15	0.16	No DIF
8	0.19	0.21	0.21	0.58	0.02	No DIF
9	0.38	0.38	0.38	0.02	0.00	No DIF
10	0.34	0.42	0.42	19.09	0.08	Large
11	0.41	0.42	0.42	0.37	0.01	No DIF
12	0.27	0.28	0.28	0.98	0.01	No DIF
13	0.08	0.09	0.13	5.20	0.05	No DIF
14	0.47	0.51	0.51	3.09	0.04	No DIF
15	0.32	0.33	0.33	0.98	0.01	No DIF
16	0.36	0.37	0.38	1.46	0.02	No DIF
17	0.53	0.53	0.53	0.14	0.00	No DIF
18	0.53	0.53	0.53	1.41	0.00	No DIF
19	0.56	0.61	0.61	25.05	0.05	Moderate
21	0.47	0.48	0.48	3.40	0.01	No DIF
22	0.57	0.65	0.65	34.08	0.08	Large
23	0.48	0.50	0.50	11.21	0.02	Negligible
24	0.56	0.58	0.58	13.42	0.02	Negligible
25	0.53	0.54	0.54	1.49	0.01	No DIF
26	0.45	0.45	0.46	3.16	0.01	No DIF
27	0.55	0.57	0.58	12.10	0.03	Negligible
28	0.39	0.42	0.42	16.30	0.03	Negligible
29	0.52	0.52	0.54	7.41	0.02	No DIF
30	0.44	0.44	0.45	3.20	0.01	No DIF
31	0.31	0.33	0.33	8.32	0.02	No DIF
32	0.24	0.28	0.29	12.55	0.05	Moderate
33	0.63	0.63	0.64	3.67	0.01	No DIF
34	0.57	0.57	0.57	3.84	0.00	No DIF
35	0.33	0.33	0.34	3.40	0.01	No DIF
36	0.16	0.16	0.17	0.26	0.01	No DIF
37	0.41	0.42	0.43	5.59	0.02	No DIF
38	0.37	0.38	0.38	3.26	0.01	No DIF

Items	Stepwise R^2			DIFF* Chi Square	DIFF* R^2	Size of DIF
	Step 1	Step 2	Step 3			
39	0.38	0.38	0.38	1.15	0.00	No DIF
40	0.57	0.57	0.57	1.63	0.00	No DIF
41	0.39	0.39	0.40	1.62	0.01	No DIF
42	0.38	0.38	0.39	2.04	0.01	No DIF
43	0.52	0.52	0.53	6.01	0.01	No DIF
44	0.19	0.19	0.24	0.45	0.05	No DIF
45	0.38	0.38	0.38	0.13	0.00	No DIF
46	0.33	0.33	0.34	0.43	0.01	No DIF
47	0.52	0.52	0.53	0.34	0.01	No DIF
48	0.19	0.19	0.24	1.71	0.05	No DIF
49	0.28	0.37	0.40	6.67	0.12	No DIF
50	0.25	0.25	0.28	1.45	0.03	No DIF
51	0.64	0.73	0.73	1.18	0.09	No DIF
52	0.22	0.23	0.41	4.67	0.19	No DIF
53	0.23	0.38	0.38	2.07	0.15	No DIF
56	0.42	0.54	0.54	1.78	0.12	No DIF

The above table indicates that eight items display DIF, of which two items (10 and 22) had a large effect size, $R^2\Delta>0.06$. Two items (19 and 32) showed a moderate effect size of $0.035<=R^2\Delta 0.06$. And the remaining four items (23, 24, 27 and 28) displayed negligible effect size, $R^2\Delta <0.03$. The null hypothesis of "no DIF" for the above was rejected for the items that displayed moderate and large DIF.

Table 24:
Direction of DIF for the Dictation subtest

Item	Variables in equation	Beta	Direction
10	Total	0.00	Uniform DIF English first language speaking group favoured.
	Language group	-23.86	
	Interaction term	0.17	
	Constant	21.20	
19	Total	0.28	Uniform DIF English first language speaking group favoured.
	Language group	-1.71	
	Interaction term	-0.01	
	Constant	-5.24	
22	Total	0.36	Uniform DIF English first language speaking group favoured.
	Language group	-0.07	
	Interaction term	-0.09	
	Constant	-6.65	
32	Total	0.32	Non-uniform DIF HA English first language speaking group favoured, and LA Xhosa first language speaking group favoured.
	Language group	5.44	
	Interaction term	-0.13	
	Constant	-12.97	

**HA-high ability and LA-low ability*

The above table indicates that two of the items with large DIF (10 and 22) and one item with moderate DIF (19) displayed uniform DIF, which favoured the English first-language-speaking group. The remaining item with moderate DIF (32) displayed non-uniform DIF, which favoured the high ability English first-language-speaking group and the low ability Xhosa-speaking-group. Table 25 presents a summary of the results of the logistic regression method for all of the subtests.

Table 25:
Summary results of Logistic Regression method

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

Subscale	Effect Size	Type of DIF	Direction of DIF favour	Number of Items	Items
Dictation	Large DIF	Uniform	English	2	10, 22
			Xhosa	0	
		Non-uniform	HA English LA Xhosa	0	
			HA Xhosa LA English	0	
	Moderate DIF	Uniform	English	1	19
			Xhosa	0	
		Non-uniform	HA English LA Xhosa	1	32
			HA Xhosa LA English	0	

**HA-high ability and LA-low ability*

With regard to the above summary table, Picture Vocabulary subtest displayed eight items with DIF, of which seven items displayed large DIF and one item moderate DIF. Three of the large DIF items displayed uniform DIF, two of which favoured the Xhosa first-language-speaking group and the remaining one item favoured the English first-language-speaking group. The remaining four items with large DIF displayed non-uniform DIF, of which three items favoured the high ability English first-language-speaking group and low ability Xhosa first-language-speaking group, while the remaining item favoured the high ability Xhosa first-language-speaking group and the low ability English first-language group.

With regard to the Verbal Analogies subtest, four items displayed DIF with two items having large DIF and two items having moderate DIF. The two large items displayed uniform DIF, which favoured the English first-language-speaking group. The remaining two moderate items, displayed non-uniform DIF, which favoured the high ability Xhosa first-language-speaking group and the low ability English first-language-speaking group.

The Letter-word Identification subtest displayed twenty DIF items, thirteen of which were large DIF items and seven were moderate DIF items. Eleven of the large DIF items displayed uniform DIF five of which favoured the Xhosa first-

language-speaking group and the remaining six items favoured the English first-language-speaking group. The two remaining large DIF items displayed non-uniform DIF, which favoured the high ability English first-language-speaking group and the low ability Xhosa first-language-speaking group. With regard to the seven moderate DIF items, five displayed uniform DIF, of which three favoured the English first-language-speaking group and two favoured the Xhosa first-language-speaking group, while the remaining two moderate items displayed non-uniform DIF which favoured the high ability English first-language-speaking group and the low ability Xhosa first-language-speaking group.

The final subtest, Dictation, displayed four DIF items of which two displayed large DIF and two displayed moderate DIF. The two large items displayed uniform DIF, which favoured the English first-language-speaking group. One of the moderate items displayed uniform DIF, which favoured the English first-language-speaking group, and the remaining item displayed non-uniform DIF which favoured the high ability English first-language-speaking group and the low ability Xhosa first-language-speaking group.

From the above results of the logistic regression analysis it is clear that item bias or differential item functioning (DIF) exists for all of the subtests. In light of these findings, the null hypothesis is rejected for all of the subtests, as the probability of scoring one on an item is not due to ability only, but due to the membership of the language group as well.

4.6 Summary

This chapter focused on statistical analysis in order to evaluate four specific objectives, outlined in Chapter 1 of this study. The researcher has outlined three null hypotheses (the third specific objective does not require a null hypothesis as it was analysed descriptively).

The first specific objective's null hypothesis stated that there are no group differences between the English first-language-speaking group and Xhosa first-language group with regard to their mean scores on the adapted version of the English version of the WMLS. The specific objective was evaluated by means of a mean score evaluation and Hotellings'T^2-test, and group differences were discovered between the English first-language-speaking group and the Xhosa first-language-speaking group on all of the subtests. Therefore the null hypothesis was rejected.

Specific objective two's null hypothesis stated that there are no group differences in terms of reliability for each language group on the various subtests of the WMLS. This objective was evaluated by means of Cronbach's Alpha and it was discovered that two of the subtests displayed significant differences in reliability between the English first-language-speaking group and the Xhosa first-language-speaking group. These subtests were the Picture Vocabulary and Dictation subtest. With regard to these differences, the null hypothesis was rejected for the Picture Vocabulary and Dictation subtest. There were no differences observed in reliability between the two language groups on the Verbal Analogies and Letter-word Identification subtests, and therefore the null hypothesis was not rejected for these subtests. With regard to specific objective three, no null-hypothesis was stated as the item characteristics were analysed descriptively.

Specific objective four's null hypothesis stated that the probability of scoring 1 on an item will be the function of ability only, in other words there will be no item bias across the English and Xhosa first-language-speaking groups. It was observed that item bias or Differential Item Functioning (DIF) exists for all of the subtests and within each language group (English first-language speaking-group and Xhosa first-language-speaking group). In other words, the probability of scoring 1 was not due to ability only. In light of this, the null hypothesis was rejected for several of the items in each subtest.

From these findings it can be concluded that the English version of the WMLS cannot be regarded as equivalent across the two language groups, and therefore the scores of the adapted English version of the WMLS cannot be used for comparison across the two languages groups with all the items included in the scores.

In the following chapter, Discussion and Conclusion, the results of this chapter will be discussed in light of the overall aim of the study which is evaluating the overall equivalence of the adapted English version of the WMLS to be used across the English first-language-speaking group and the Xhosa first-language group. Recommendations will be made to possibly remedy the problems that have arisen in this chapter.

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1 Introduction

The overall aim of this study was to evaluate the equivalence of the adapted English version of the WMLS for use across English first-language-speaking and Xhosa first-language-speaking groups. The overall aim was evaluated by means of four specific objectives, 1) to evaluate mean group differences between English and Xhosa first-language-speaking groups on the English version of the WMLS, 2) to evaluate group differences in terms of the reliability of the test between English and Xhosa first-language learners on the English version of the WMLS, 3) to evaluate group differences in terms of the item characteristics of the test between English and Xhosa first-language learners on the English version of the WMLS, and 4) to evaluate the item bias, across English and Xhosa first-language-learners, of the English version of the WMLS.

These objectives were analysed and the results were presented in the previous chapter (Chapter 4, Results). In essence, these objectives assessed group differences, reliability, item characteristics and lastly differential item functioning, in order to make certain inferences about the scalar equivalence of full comparability of the adapted English version of the WMLS to be used across the English first-language-speaking group and the Xhosa first-language-speaking group.

This chapter will focus on the major findings of the results, and a discussion will be presented in detail to identify the implications of the results and finally make recommendations based on these results.

5.2 Discussion of the results

5.2.1 Evaluating group differences on total mean scores of the language groups

As indicated above, significant group differences were discovered with the analysis of the total mean scores for each language group per subtest, and the null hypothesis was thus rejected for this objective.

According to Huysamen (2002) (discussed in more detail in Chapter 2 of this study), mean group differences may result from tests which are lacking in relevant construct variances or that contain irrelevant construct variances owing to the close association between languages and culture. This is known as construct-irrelevance or construct-underrepresentation. Therefore, when group mean differences are discovered between groups, an investigation should be done into the construct representation and construct relevance as possible reasons. With reference to the various levels of equivalence (as discussed in chapter 2), when an instrument measures different constructs in two different cultures, or in this case language groups, no comparison can be made (Van de Vijver & Leung, 1997). According to Van de Vijver and Leung (1997), this is known as construct inequivalence or structural inequivalence (discussed in depth in Chapter 2, Taxonomy of Equivalence section). In other words, there will be no link between the scores obtained in one group and scores obtained in another group.

With regard to the mean group differences displayed between the English and Xhosa first-language-speaking groups (from the above discussion); the English first-language-speaking group had a higher mean group score than the Xhosa first-language-speaking group. These differences may be due to construct irrelevance in the Xhosa first-language-speaking group, as the Xhosa first-language learners may not have been familiar with the images in the Picture Vocabulary subtest or the wording of the Verbal Analogies, Letter-word Identification and the Dictation subtests. We can therefore assert that the concepts or constructs may be different within each language group and the items that were included in the test were more

relevant for the English first-language-speaking group than the Xhosa first-language-speaking group. However, despite the Xhosa first-language learners having English as their second language and being instructed through the medium of English in school (which therefore might be viewed as their dominant language), their home language might remain Xhosa, and according to Huysamen (2002), might be their more proficient language. This therefore could affect their mean group scores on the various subtests, and these differences may thus be due to real differences on the constructs of interest.

At this stage of the findings, the evidence was not sufficient to determine whether the adapted English version of the WMLS is, or is not, equivalent across the two language groups.

5.2.2 Group differences in terms of reliability for the language groups across the subtests

With regard to this specific objective, it was determined that there were significant group differences between the two language groups on the Picture Vocabulary subtest and the Dictation subtest in the matter of reliability, and therefore the null hypothesis was rejected for these subtests only.

According to Anastasi and Urbina (1997), an important condition influencing the size of the reliability coefficient, is the nature of the group for which reliability is being measured. "The nature of the group" could possibly refer to the range of individual differences within the group, the range of scores within the group, and the variances between the individuals within the group.

Firstly, any correlation coefficient is affected by the range of individual differences in the group. For example, if all the members of a particular group were fairly similar in spelling ability, then the correlation of spelling with any other ability would be close to zero in that group (Anastasi & Urbina, 1997). Therefore it would be virtually impossible (within such a group) to predict an

individual's standing on any other ability from a knowledge of her or his spelling score. In addition, Anastasi and Urbina (1997) indicate that the reliability coefficient depends on the variability of the sample within which they are found. In other words, if the reliability coefficient reported in a test manual is calculated for a particular group ranging from Grade four learners to high school learners, it cannot be assumed that the reliability will be equally high for Grade eight learners.

Additionally, according to Van de Vijver and Leung (1997), cross-cultural comparisons of scores presuppose accurate psychometric properties of the measures in all groups. The accuracy of psychometric properties is measured at an item or instrumental level (Van de Vijver & Leung, 1997). With regard to measuring the psychometric properties at an instrumental level, it is common practice to compare reliability coefficients. If significant differences are observed, the sources of these differences should be scrutinised. In addition, these differences can be explained by deviations of items from the norm, which result in what is known as the "floor and ceiling" effect. In other words, this occurs when the items of a given test do not measure the same construct in any of the groups (Van de Vijver and Leung, 1997). According to Allen and Yen (1979), when evaluating psychological and educational tests, the observed score distribution is approximately normal if the bulk of examinees have scores near the average score and fewer examinees have scores near the extremes. However, it is not necessary for a test to produce a normal curve, and although many score distributions are not normal, deviations from normality sometimes indicate a problem with the test (Allen & Yen, 1979).

Allen and Yen (1979) have identified two possible issues arising from deviations from the norm, namely the floor and the ceiling effects. When an ability test is administered that is too easy and most people get a very high score, the trait being measured might have a normal distribution, but all the people who would be in the upper tail (of the distribution curve) if the test were harder are crowded into the upper range of scores and therefore discrimination between the individuals (who

have taken a easy test) with higher ability and lower ability would not be possible. This is known as the ceiling effect. On the other hand, an ability test that is too hard has the opposite effect, and all the people who would be in the lower tail of the trait distribution are crowded into the lower range of the test-score distribution, resulting in the floor effect (Allen & Yen, 1979).

There could thus be various reasons why the two language groups displayed different reliabilities on the various subscales. In light of this discussion, one of the most likely contributing factors to the differences in reliability between the two language groups on the given subtests (Picture Vocabulary and Dictation subtest), was the variability of scores. With regard to the above findings, the English first-language-speaking group displayed a lower reliability coefficient than the Xhosa first-language-speaking group on the Picture Vocabulary displayed a lower Cronbach's Alpha (0.73). This could have been due to the lower variability (in the mean scores) found in the English first-language-speaking group. This test was easier for this group than for the Xhosa group, but it was not too easy, as can be seen from the mean score (30 out of a possible total of 57). The Picture Vocabulary subtest measures oral language, including language development and lexical knowledge. However, owing to the language of the test, namely English, the English first-language-speaking group might be at an advantage in that their first language is English while they are more familiar with the culture represented by the items as well. This therefore may have led to this group having a lower variability in their scores.

5.2.3 Group differences on item characteristics for the four subtests between the English first-language-speaking group and the Xhosa first-language-speaking group

The item characteristics results indicated that there were various differences between the English first-language speakers and the Xhosa first-language speakers on various items in the various subtests These differences were observed with regard to the mean item totals (item difficulty) and item-total correlations

(item discrimination). Furthermore item characteristic analysis can be impacted by many aspects such as the position of an item or administration instructions.

However the total mean item score or item difficulty scores indicated that both language groups performed satisfactorily in terms of mean item difficulty values on the Picture Vocabulary, Letter-word Identification and Dictation subtests. Both groups displayed lower (than the suggested mean of 0.5) mean item difficulty levels on the Verbal Analogies subtest. However, the Xhosa first-language learners displayed much lower mean item difficulty levels on this scale than the English group. It is important to note that the English first-language group displayed higher mean item difficulty values on all the subtests.

The mean item total correlations or item discrimination values displayed on the Picture Vocabulary subtest for both language groups had lower than the recommended values (of around 0.3). Thus, with regard to the overall subtest, the items did not discriminate well between high and low achievers in both language groups on this scale. With regard to the Verbal Analogies and Letter-word Identification subtests, overall both language groups displayed good discriminating values on the items of this subtest. However, both of these subtests displayed lower discriminating power in the English first-language-speaking group, essentially indicating that items in these scales succeed better in discriminating between high and low achievers (on the total scores). Furthermore, the English first-language-speaking group displayed unsatisfactory discriminating power between high and low achievers on the items of the Dictation subtest, while the items of the Dictation subtest displayed a good discriminating power between high and low achievers in the Xhosa first-language learners. These results indicated differences in terms of item characteristics, and further analyses thus needed to take place in order to determine whether these items were biased.

5.2.4 Item Bias (Differential item functioning) displayed on the four subtests for both language groups

The fourth objective was evaluated by means of the statistical procedure of logistic regression. The results obtained indicated that all of the subtests displayed differential item functioning (DIF) or bias for both groups. Therefore the null hypothesis was rejected for several items of each subtest.

Van de Vijver and Leung (1997) state that items bias (DIF) can be produced by various sources such as incidental differences in appropriateness of the item content (for example, some items of an educational test are not in the curriculum in one cultural group), inadequate formation (such as complex wording) and inadequate translation. Another possible source of item bias could be the incidental inappropriateness of item content (Van de Vijver & Leung, 1997).

It is thus clear that there could be various explanations for the various DIF items which were identified on the different subtests. For example, when investigating the possible reasons for DIF on the Picture Vocabulary subtest, it was found that some of the images or pictures are potentially irrelevant in the Xhosa first-language-speaking group. In other words, some of these images were not familiar to the Xhosa first-language-speaking group, but were familiar to the English first-language-speaking group. An example of this could be item 29 (an item that displayed large uniform DIF which favoured the English first language speaking group) which is an image of an igloo. As the Xhosa group included a large number of Xhosa learners from the semi-rural areas, it could be argued that the Xhosa-speaking learners had not been exposed to this type of imagery. On the other hand the large uniform items (15 and 30, images of sneakers and a theatre respectively) favoured the Xhosa first-language-speaking group. It can be argued that these images were familiar to the individuals from the Xhosa first-language-speaking group.

When analysing the Verbal Analogies subtests, all of the large uniform DIF items favoured the English first-language-speaking group. The two items which favoured the English first-language-speaking group were items 8 and 9. Item 8 is as follows, "train is to track as car is to _____" (answer: road, highway, street or lane). With regard to this item it can be argued that many of the Xhosa first-language-speaking individuals came from rural areas in the Eastern Cape, where there are not many roads or highways, therefore it could be said that the cultural background of the Xhosa first-language learners was not taken into account and therefore they were put at a disadvantage as this item could be said to form part of cultural irrelevance.

With regard to the Letter-word Identification subtest, there are several items which displayed DIF for both the English first-language-speaking group and the Xhosa first-language-speaking group. With regard to the large uniform DIF items which favoured the English first-language-speaking group, it could be argued that many of the words used on this subtest (where each word needed to be pronounced aloud) were more familiar to the English first-language-speaking group. An example of one of the items is item 31, which was the word "island". In other words, the use of the word "island" may be more prominent in the English first-language-speaking group and therefore this language group may be more familiar with this word. With regard to this argument it can be said that the items in this subtest were less familiar to the Xhosa first-language-speaking group and therefore cultural irrelevance occurs for this language group.

With regard to the final subtest of Dictation, all of the large uniform DIF items favoured the English first-language-speaking group. An example of this is item 10, which required that the testee write down the letter "Y" as a capital letter. A possible reason why the English first-language-speaking group was favoured could be that the Xhosa first-language group were not familiar with the concept of capital letters.

It is important to note that the existence of both uniform and non-uniform DIF is regarded as hazardous for scalar equivalence (which is the highest level of equivalence). Therefore with the finding of both uniform and non-uniform DIF, the scalar equivalence of the instrument, namely the adapted English version WMLS, is not equivalent across both language groups.

Van de Vijver and Leung (1997) have identified various ways to deal with DIF or item bias. Firstly these authors state that the analysis of item bias can be viewed as an indicator that an instrument is inadequate for cross-cultural comparison, which can guide a researcher in not partaking in cross-cultural comparisons or in using the instrument across cultural groups. However, Van de Vijver and Leung (1997) caution against this, as any study with highly dissimilar cultural groups may contain item bias. Secondly, item bias can be viewed as providing important clues about cross-cultural differences (Van de Vijver & Leung, 1997). In other words, unbiased items define 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 is developed based on universal and culture-specific elements of a construct (Van de Vijver & Leung, 1997). 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 with the use of item bias analysis, biased items can be identified and removed.

5.3 Conclusion

The major findings (as outlined in Chapter 4 of this thesis) indicate that there were overall differences between the English first-language-speaking group and the Xhosa first-language-speaking group on their mean scores (objective 1), their reliability displayed on each subtest for both language groups (objective 2), there were significant differences between the language groups on the item characteristics of each subtest (objective 3) and differential item functioning (item

bias) was displayed on all of the subtests (objective 4). Because of these significant findings it can be concluded that the scores of the adapted English version of the WMLS cannot be used for comparison across the two language groups with all the items included in the scores. Therefore the scores cannot be regarded as displaying scalar equivalence across language groups.

5.4 Limitations

The first limitation of this sub-study and the larger study is that sampling was not done with generalisability in mind, which affects the external validity of the adapted WMLS versions. However, the main researcher and the researcher of the sub-study do not view this as a major limitation as the main interest at this stage is the appropriateness of the instrument. Issues of internal validity are thus regarded as more important at this stage.

Secondly, there is a problem that arises with any type of DIF analysis, namely that of small sample sizes. In other words, group ability differences and small groups within ability groups can lead to unstable results and over-identification of DIF items (Koch, 2009). The suggestion is to cross-validate the findings using other DIF analysis methods.

5.5 Contribution and Recommendation

The researcher would like to recommend that the various items which displayed DIF be further investigated, as the scope of this study did not allow for in-depth investigation and interpretation of these items. In addition, a further investigation needs to be conducted around the scalar inequivalence in order to attain scalar equivalence so that this measure, the English version of the WMLS, can be used across the two language groups.

Finally, it is important to note that the main study as well as this sub-study constitute one of the first studies of their kind regarding monolingual language

tests and their use across language groups. This study has thus contributed to the need to evaluate the equivalence of monolingual language tests for use across different language groups.

REFERENCES

Abrahams, F., & Mauer, K. F. (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.

Alexander, N. (2002). Linguistic rights, language planning and democracy in post-apartheid South Africa. In S. Baker (Ed.), Language policy: Lessons for global models. Moneterey: Moneterey Institute of International Studies.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. California: Brooks/Cole.

Anastasi, A., & Urbina, S (1997). *Psychological Testing.* New Jersey: Prentice-Hall.

Apartheid South Africa: Bantu Education. Retrieved: 28 May 2008 at 3:45pm http://www.rebirth.co.za/apartheid_segregation_bantu_education3.htm

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing, 17*(1), 1-42.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1-34.

Bachman, L. F., & Palmer, A. S. (1981). The construct validation of the FSI oral interview. *Language Learning, 31*(1), 67-86.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*(4), 449-465.

Bachman, L.F., & Cohen, A.D. **(**1998). Language testing-SLA interfaces: An update. In Bachman, L.F. and Cohen, A.D. (Eds.). *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press.

Cattell, R.B. (1940). A culture-free intelligence test, I. *Journal of Educational Psychology, 31,* 176-199.

Cattell, R.B., & Cattell, A.K.S. (1963). *Culture fair intelligence test.* Champaign, IL: Institute for Personality and Ability Testing.

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

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy.* San Diego, CA: College Hill Press.

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

De Klerk, G (2008). Cross cultural testing. In M. Born, C.D. Foxcroft, and R. Butter (Eds.), *Online reading in testing and assessment, international test.* http://www.intestcom.org/Publications/ORTA/cross+cultural+testing.php

Escamilla, K. (2006). Monolingual Assessment and Emerging Bilinguals: A Case Study in the US. In O. Garcia, T. Skutnabb-Kangas, & M. Torres Guzman (Eds.), *Imagining Multilingual Schools.* Ch. 9: pp. 184-199. Clevedon, England: Multilingual Matters Ltd.

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

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

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

Foxcroft, C., & Roodt, G. (2005). *An introduction to psychological assessment in the South African Context* (2nd ed.). Cape Town: Oxford University Press South Africa (Pty) Ltd.

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

Glass, G. V. (1976). Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher, 5*(10), 3-8.

Hinkel, E. (2005). *Handbook of research in second language teaching and learning.* London: Lawrence Erlbaum Associates, Publishers.

Hornberger, N. H., & Shohamy, E. (2008). Encyclopedia of Language and Education. *Language Testing and Assessment, 7.* Berlin: Springer.

Hudson, T., Detmer, E., & Brown, J.D. (1992). *A framework for testing cross-cultural pragmatics*. Honolulu: Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.

Hudson, T., Detmer, E., & Brown, J.D. **(**1995). *Developing prototypic measures of cross-cultural pragmatics*. Honolulu: Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.

Hui, C. H., & Triandis, H.C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural. Psychology, 20,* 296-309.

Huysamen, G. K. (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.

Kaplan, R. M., & Saccuzzo, D. P. (1997). *Psychological testing: Principles, applications and issues* (4th ed.). Pacific Grove: Brooks/Cole.

Kunnan, A. J. (2000). *Fairness and validation in language assessment: selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida.* Cambridge: University of Cambridge.

Koch, S.E. (2005). *Evaluating the equivalence, across language groups, of a reading comprehension test used for admission purposes*. Unpublished doctoral thesis. University of Port Elizabeth, Port Elizabeth.

Koch, E. (2007a). Die evaluering van ’n eentalige toelatingstoets wat vir toelating tot hoër onderwys in ’n meertalige konteks gebruik word. *SA Joernaal vir Industriële Sielkunde, 33*(1), 90–101.

Koch, E. (2007b). 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.

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 2008, 26*(3), 333–350.

Koch, E., Landon, J., Jackson, M. J., & Foli, C. (2009). First brushstrokes: Initial comparative results on the additive bilingual education project (ABLE). *South African Linguistics and Applied Language Studies, 27*(1), 109-127.

Meiring, D. (2007). *Bias and equivalence of psychological measures in South Africa.* Published doctoral dissertation. Netherlands: Labyrint Publication.

Meiring, D., Van de Vijver, F., Rothmann, S., & Barrick, M.R. (2005). Construct, item and method bias of cognitive and personality tests in South Africa. *South African Journal of Industrial Psychology, 31*(1), 1 – 8.

Nzimande, B. (1995). *Culture fair testing? To test or not to test? Paper delivered at the Congress on Psychometrists for Psychologists and Personnel Practitioners,* Pretoria 5-6 June 1995.

Pienemann, M., Johnson, J., & Brindley, G. (1988). Constructing an acquisition based procedure for language assessment. *Studies in Second Language Acquisition 10,* 217–43.

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

Ramphal, A. (1972). *An investigation into the suitability of the National Bureau Group Test for Five-and Six-year Olds as an instrument for measuring school readiness among a group of Indian children in Durban.* Unpublished master's dissertation, University of Durban-Westville.

Ross, C.E., & Mirowsky, J. (1984). Socially-desirable response and acquiescence in a cross-cultural survey of mental health. *Journal of Educational Measurement, 17*, 1-10.

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

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

Valdés, G., & Figueroa, R. (1994). *Bilingualism and testing. A special case of bias.* Norwood, NJ: Ablex Publishing Corporation.

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

Van de Vijver, F.R.J., & Tanzer, N.K. (1998). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47*, 263-279.

Van de Vijver, A. J. R., & Rothmann, S. (2004). Assessment in multicultural groups: The role of acculturation. *Applied Psychology: An International Review, 53*(2), 215-236.

Van Haaften, E. H., & Van de Vijver, F. J. R. (1996). Psychological consequences of environmental degradation. *Journal of Health Psychology, 1*.

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

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

ADDENDUM A

Information Sheet

2008

Dear Parent

Your child has been selected as a possible participant in a research project of the Nelson Mandela Metropolitan University, called "A translation of a test of academic language proficiency into Xhosa".

The test is available in both English and Xhosa. This year we need to test the Xhosa speaking children on the English version of the test. The purpose of this part of the research project is to ensure that the English version of the test does not bias against Xhosa speaking children. The testing will take about one hour, and will be conducted at the school. Permission for this research project has been obtained from both the district manager and the school principal.

We cannot proceed with this research unless you give your permission for your child to be tested. We would therefore appreciate it if you would be kind enough to read the attached consent form, sign it and send it back to the school ASAP. If you have any questions concerning the research, please contact Elize Koch at 0824439311.

Regards

Dr. Elize Koch

Main Researcher.

ADDENDUM B

Informed Consent Form

1. The ABLE research team (consisting of Elize Koch, M-J Knoetze and Cordelia, Foli, who are working as researchers at the Nelson Mandela Metropolitan University, and Rhodes University) has requested my child to be part of a research study. The title of the research is *"An adaptation of a test of academic language proficiency into Xhosa."*

2. "I have been informed that the purpose of the research is to determine the psychometric properties of the instrument for the South African population as well as of the equivalence of the English and Xhosa versions of the test."

3. "I give permission for my child to be assessed on the test used in the study. The testing will involve about 1 hour of testing"

4. "I understand that the results of the research may be published but that my name or that of my child or our identity will not be revealed."

6. "I have been informed that any questions I have concerning the research study or my participation in it, before or after my consent, will be answered by Elize Koch at 0824439311."

7. "The above information has been explained to me. I understand everything. The nature, demands, risks and benefits of the project have also been explained to me. I understand that I may withdraw my consent and discontinue my participation at any stage without any penalty or loss of benefit to myself. In signing this consent form, I am not waiving any legal claims, rights or remedies. "

Participant

name:……

Participant signature

(parent):……Date……

"I certify that I have explained to the above individual the nature and purpose, the potential benefits, and possible risks associated with participation in this research study, have answered any questions that have been raised, and have witnessed the above signature."

Signature of researcher…….Date……