University of the Western Cape

MINI-THESIS

Title: *Evaluating the structural equivalence of the English and isiXhosa versions of the Woodcock Munoz Language Survey on matched sample groups*

Name:	Danille Arendse
Student number:	2513873
Degree:	M.A. (Research Psychology)
Faculty:	Community and Health Sciences
Department:	Psychology Department
Supervisor's name:	Professor Elize Koch
Date:	November 2009

Submitted in partial fulfilment for the requirements of the M.A. (Research Psychology) degree at the University of the Western Cape.

Keywords: language policy, language proficiency, cognitive academic language proficiency, bilingual education, test adaptation, test translation, language testing, bias, structural equivalence, factor analysis.

DECLARATION

The author hereby declares that the following final research report, is of her own work and that all the sources she has used or quoted have been indicated and acknowledged by means of complete references

…………………………………………….
D. E. Arendse

ACKNOWLEDGEMENTS

I would like to convey my heartfelt appreciation and sincere thanks to the following individuals for their priceless contribution to the completion of this final research report.

My heavenly father, whom I am nothing without and who helped me to endure throughout a demanding Masters year. My precious supervisor, Prof. Elize Koch, whose academic brilliance and wisdom not only encouraged me, but also served to stimulate the production of my thesis. Her astuteness and profound knowledge made her a privilege to work with. I am humbled by her kindness and appreciate all that she has done for me during the writing of my thesis. I thus salute her in all her brilliance.

A special thanks to my mother, Magdalene, for her invaluable encouragement and continuous voice of wisdom. She is an inspiration to me and has guided me throughout my life. Her persistent support throughout my Masters year has been priceless and is dually appreciated. I would also like to thank my family: JPH ‘Pa’ Lewin, Zelda, Maxine, Daniel, Daniel Snr, whose valuable support was crucial to my work. To my faithful and cherished friends, namely: Unity, Lerico ‘Stepper’, Adrian , Candice, Sidney, Nazneen, Tremaine, Fiona, Ghurswin ‘biggy’, Wesley, Johann, Yentl, Abdul and my departed friend, Andre Lakey. They have not only supported me throughout the year, but made my Masters year unforgettable. They are the pillars on which good friendships rest and one which I will treasure forever.

A sincere thanks to my esteemed research class, namely: Nondumiso, Crystal, Candice, Guia, Taryn and Chernelle, whose friendship and insightful knowledge was much appreciated. A immense thanks to the Psychology Department at the University of the Western Cape and its honoured lecturers who graced me with their teaching as well as its staff members, a special word of thanks for your support and motivation throughout the year. It was truly a memorable experience. Finally, a word of thanks and sincere gratitude to all those whom I have not mentioned above, but whose knowledge and support was much appreciated.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Page

LIST OF FIGURES

LIST OF FORMULAS

ABSTRACT

The diversity embodying South Africa has emphasized the importance and influence of language in education and thus the additive bilingual programme is being implemented in the Eastern Cape by the ABLE project in order to realize the South African Language in education policy (LEiP). In accordance with this, the Woodcock Munoz Language Survey (which specializes in measuring cognitive academic language proficiency) was chosen as one of the instruments to evaluate the language outcomes of the programme and was adapted into South African English and isiXhosa. The current study was a subset of the ABLE project, and was located within the bigger project dealing with the translation of the WMLS into isiXhosa and the successive research on the equivalence of the two language versions. This study evaluated the structural equivalence of the English and isiXhosa versions of the WMLS on matched sample groups (n= 150 in each language group). Thus secondary data analysis (SDA) was conducted by analyzing the data in SPSS as well as CEFA (Comprehensive Exploratory Factor Analysis). The original data set was purposively sampled according to set selection criteria and consists of English and isiXhosa first language learners. The study sought to confirm previous research by cross-validating the results of structural equivalence on two subscales, namely the Verbal Analogies (VA) and Letter-Word Identification (LWI) subscale. The research design reflects psychometric test theory and is therefore located in a bias and equivalence theoretical framework. The results of the exploratory factor analysis found that one can only accept structural equivalence in the first factor identified in the VA subscale, while structural equivalence was found in the factor for the LWI subscale. The use of scatter-plots to validate the results of the exploratory factor analysis indicated that one can tentatively accept these results. The study thus contributed to the literature on the translation of the WMLS, and the adaptation of language tests into the indigenous languages of South Africa, as well as additive bilingual programmes.

CHAPTER ONE

INTRODUCTION

1.1 Preamble

The current study makes use of exploratory factor analysis to evaluate the structural equivalence of the isiXhosa and English versions of the Woodcock Munoz Language Survey (WMLS) on matched sample groups. The present study is a subset of the Additive Bi-lingual Education (ABLE) project being realized by a number of researchers attached at several universities in the Eastern Cape and Western Cape, and is located within the research dedicated to the translation of the WMLS into isiXhosa and the successive research on the equivalence and validity of the two language versions. Moreover, the study serves to further research on the above-mentioned aspects as well as improve on methodological limitations noted in previous research on the equivalence of the two language versions. This will be elaborated on later in this paper in order to create a better understanding.

This chapter will serve to provide the background of the study and explores the rationale for why this study was pursued. This is followed by the aims and objectives of the study, which is the central focus of the methodology section. Along with this, the theoretical framework is mentioned which will be elaborated on in a later chapter. The next section is a brief review of bilingual education and testing which forms the background of the subsequent chapters regarding the literature on cross-lingual testing. Lastly, a brief description of the chapters to follow will be presented. This will allow the study to be presented in a coherent manner, enabling for better understanding of the present study.

1.2 Background of the study

South Africa is a multicultural society and thus each of the languages, cultures and traditions contributes to representing the South African nation. As a result, the South African constitution pledges equal status for all the eleven languages in South Africa and allows one to use the language of choice (Constitution of the Republic of South Africa, 1996). This then acts as a means by which linguistic diversity is acknowledged and respected. In addition, the equality clause states that schools must "provide education to all South Africans in their mother tongue" (Mangena, 2002, 1). Ironically, public institutions such as schools emerge as the instigators behind disadvantaging learners in their development and progress of their already marginalized languages. This makes the existence of functional multilingualism invisible and English continues to dominate as the language of education. This is where language rights is important and should be exercised; otherwise, language injustices will occur (Mangena, 2002).

As a means of eliminating the language injustices that occur and recognizing linguistic diversity, the government implemented the use of a new language policy in education, which would encourage multilingualism. The new policy was strongly informed by research and is known as the Language- in- Education policy (www.kzneducation.gov.za), which was accepted in 1997 and advocates that learners benefit emotionally and cognitively from bilingual education in a dual medium programme in schools (Mangena, 2002). The Language in Education policy (LiEP) endorses multilingualism in the form of additive bilingualism. This allows learners to maintain the use of their primary language as well as utilize it in their school, while simultaneously learning English. Furthermore, the policy permits the school and teacher to use the language of choice for learning and teaching and allows them to decide at which time to introduce an additional language (Asmal, 1999). The argument behind the LiEP is that children will learn more effectively and learn English more efficiently if their primary language was maintained and they simultaneously learnt English as a subject (Asmal, 1999).

There is however resistance to the usage of this policy in South Africa, as some schools do not use the additive bilingual approach and force the use of English as the medium of instruction from grade one (or grade 4), which tends to yield poor results. Much to the education department's dismay, the 2008 statistics revealed that some of the worst pass rates belonged to the Eastern Cape, which has also been identified as a rural and poverty-stricken province. Their learners had a pass rate of 50, 6% and only 14, 3% of those learners met the requirements for tertiary education (Author unknown, 2009). According to theory, there is a strong correlation between the language of instruction and learner's academic performance (Cummins, 1979). Thus one can speculate that since they did not receive their education in their primary language (Koch, 2006; 2007), this contributed to their weak academic performance.

To address this issue, the ABLE project was implemented in 2003 in order to help realize the LIEP. Thus, the ABLE project forms the background and setting from which the proposed study will be proceeding. This project, known as the ABLE project (Additive Bilingual Education) was developed with the purpose of implementing the model of additive bilingual education and seeks to namely: empower learners' primary language, encourage literacy of learners in two languages, and ensure learners become academically strong and competent in two languages.

These objectives are being implemented in a rural area of the Eastern Cape, South Africa, where isiXhosa has been identified as the primary language of the community and is the language of instruction of the school, with English as an additional language (Koch et al., 2009; Koch2009).

The Woodcock Munoz Language Survey (WMLS) is an instrument, which can sufficiently measure the development of academic language proficiency in an individual's primary and second languages, and is used extensively in the United States of America to evaluate children's Additive Bilingual Education programmes and language proficiency in English and Spanish (Woodcock & Munoz-Sandoval, 2005). The ABLE project members intentionally selected the WMLS to evaluate the language development of learners, because it allows them to assess the language outcomes of the project, thereby measuring the selected participants' performance in language as well as evaluating the effectiveness of the additive bilingual programme. The WMLS was therefore adapted into South African English and isiXhosa to assess the academic language proficiency in English and isiXhosa of the learners in the project. Research on the equivalence of the two language versions of the test was the next step in the process of translating and adapting the test into isiXhosa.

Within this realm of testing in multicultural societies, there are two pivotal aspects to consider, namely, equivalence and bias. If tests are unbiased, then the scores are equivalent and can be compared across cultures or language groups. When large group differences on tests are found, it is the first point of analysis in the examination of bias and further investigation into bias is necessitated before the results can be accepted (Van der Vijver & Leung, 1997). These aforementioned concepts of equivalence and bias will be discussed in more detail in the subsequent chapter of this paper, as the researcher aims to evaluate the equivalence of the adapted English and isiXhosa language versions of the WMLS.

It is important to be aware that this paper forms part of a larger study which consists of various phases, firstly, the adaptation of the original WMLS (English version), into South African English and isiXhosa, secondly, the evaluation of the equivalence of the two language versions of the WMLS, and thirdly the evaluation of the predictive, construct and content validity of both these adapted versions across English first language and isiXhosa first language students within the South African context. This sub-study is located in the bigger project dealing with the translation of the WMLS into isiXhosa and the subsequent research on the equivalence and validity of the two language versions.

The WMLS test was adapted in that some of the items were translated literally, and others were adapted. Thus, the words as well as the content of certain items were changed in order to improve the appropriateness of items in the isiXhosa cultural context. The processes involved in the adaptation of the WMLS test into isiXhosa adhered to the guidelines specified by the International Testing Commission (www.intestcom.org). This involved the WMLS being adapted through a process of two workshops, which took the form of one two-day workshop in 2004 and a one-day workshop in 2006 (Koch, 2009). The use of a multidisciplinary team to facilitate the process of adaptation assisted in advancing the quality of the test.

The first workshop involved most of the adapting of the test, after which data was collected and exploratory analysis of the statistical equivalence of the two language versions were conducted, which included item bias analysis. The second workshop involved changing the specific items in the subscales according to the results of the analysis. One subtest that was rewritten completely after the first round of the analysis, was the Verbal Analogies test, as the whole subtest was found to be problematic (Koch, 2009). The subsequent evaluation of the equivalence of English and isiXhosa version of the WMLS provided tentative evidence of the scalar equivalence of two of the scales of the test, namely VA and LWI (Koch, 2009). However, a number of the items on both subtests still displayed DIF.

In the study by Koch (2009, in press), the analysis of the WMLS using weighted multidimensional scaling (WMDS) provided the researcher with critical results pertaining to the strengths of these scales. The scales, Picture Vocabulary (PV) and Dictation (DIC) indicated in-equivalence and it was recommended that they both be readapted in order to expect better results. The remaining scales, Verbal Analogies (VA) and Letter-Word Identification (LWI), however, indicated the presence of biased items yet they appeared to structurally equivalent across the language versions (Koch, 2007).

The results of the WMDS for the VA subscale indicated the presence of two dimensions present with slight, uninterpretable differences being noticed in the weights of the language versions with these dimensions. These dimensions were namely: Dimension 1: isiXhosa - 0, 279 and English - 0, 641. Dimension 2: isiXhosa – 0, 255 and English – 0, 271. These differences are not interpretable even though the differences on the first dimension was quite large and as a result, construct equivalence was provisionally accepted (Koch, 2009).

The results of the WMDS for the LWI subscale revealed evidence of construct equivalence across the language versions of the scale. Two dimensions were selected with no differences in weights of language versions on these two dimensions. These dimensions were namely: Dimension 1: isiXhosa – 0, 694 and English – 0, 694. Dimension 2: isiXhosa – 0, 139 and English – 0, 086. In addition to this, there was a finding of DIF items on both these subscales, with DIF items being attributed to group differences on ability (Koch, 2009).

Based on these results, further research was needed to validate the results obtained by the WMDS and thus exploratory factor analysis was suggested. In addition to this, the researcher was unable to establish where these differences lie due to the analysis method used and recommended further analyses using exploratory factor analysis as this technique allows one to explore exactly where these differences lie. The differences are observable due to the fact that exploratory factor analysis allows one to analyze each scale individually across the language versions, while WMDS simultaneously analyses the scales. The exploration of the PV and LWI scales were not considered as they need to be readapted in another workshop and thus are not appropriate for further analysis. Since the VA and LWI scales indicated promising results, further analysis was prompted to confirm the existing results (Koch, 2009). Moreover, these results motivated the use of matched sample groups by matching on ability as a means of eliminating methodological weakness observed in the previous study (Sireci & Khaliq, 2002).

1.3 Rationale

This study will advance and cross-validate previous research by addressing specific methodological limitations in the design of the previous research (Koch, 2009) such as the effect of differences in ability on structural equivalence, by matching the groups on their total scores on one of the subscales of the tests well as by using a different statistical technique to evaluate the construct bias (structural equivalence) by means of exploratory factor analysis to cross validate the previous findings. The concept of structural equivalence will be explained and discussed further under the literature review.

This research will increase knowledge and information on the advancement of these two subtests, which will then serve to identify the effectiveness of the isiXhosa version of the test, as well as the adaptation of language tests in general, especially where the indigenous languages of SA are concerned.

1.4 The aims of the study

The study aimed to evaluate the structural equivalence of the English and isiXhosa version of the Woodcock Munoz Language Survey on matched sample groups.

The specific objectives of the study were:

1) To compare the Verbal Analogies and Letter-Word Identification scales of the two language versions of the WMLS on reliability using sample groups matched on their total scores on the subscales of the test.

The hypothesis of the present study is formulated as follows:

H_{o}: There is no significant difference between the Verbal Analogies and Letter-Word Identification scales of the two language versions of the WMLS on reliability on matched sample groups on their total scores.

H_{1}: There is a significant difference between the Verbal Analogies and Letter-Word Identification scales of the two language versions of the WMLS on reliability on matched sample groups on their total scores.

2) To compare the Verbal Analogies and Letter-Word Identification scales of the two language versions of the WMLS on mean item characteristics using sample groups matched on their total scores on the subscales of the test.

No hypotheses will be formulated as this objective will be explored descriptively.

3) To evaluate the structural equivalence of the Verbal Analogies and Letter-Word Identification scales two language versions of the WMLS using sample groups matched on their total scores on the subscales of the test.

1.5 The Theoretical Framework

The theoretical framework used in this study is that of equivalence and bias (Van der Vijver & Leung, 1997). This theoretical framework will form the basis for the research design and subsequent analysis of the data.

1.6 Overview of the literature

In this overview, concepts that were briefly mentioned in the introduction will be discussed. This will allow one to understand the theory underpinning both the LiEP and the ABLE project, which is essential, as this will assist in clarifying complicated concepts related to language learning, education and testing. In addition to this, the relevance of and importance of language testing and issues pertaining to testing in two languages will be explored in the subsequent chapters in order to substantiate the emphasis placed on this study.

1.6.1 The prominence of language proficiency in bilingual education

South Africa is currently going through a post-liberation period in which multilingualism is encouraged (Wright, 2002) and thus the acquisition and teaching of primary and additional languages in education is an issue relating to language proficiency and deserves the necessary attention. In addition to this, the international literature regarding language proficiency encourages the evaluation of English language proficiency among learners in order to provide equal educational opportunities, before making decisions regarding the language that is to be used for instruction

Cummins (1984 in Laija-Rodrigues, Ochoa & Parker, 2006), postulated that a Common Underlying Proficiency (CUP) aids the transfer of language between learners first and second language. This implies that there is a universal underlying principle, which is present in the first language and can be transmitted to the second language, by maintaining the use of the first language and ensuring that the learner receives sufficient exposure to the second language.

Much research exploring the first and second language proficiency in America has found that high levels of proficiency in the first language leads to high academic achievement in the second language as well as good reading skills in the first language predicts good reading in the second language (Koda, 2005; Laija-Rodrigues et al, 2006; Parker, Louie & O'Dwyer, 2009). Hence, the language skills acquired in the first language will be transferred to the additional language and this serves as a means by which individuals can become bilingual and bi-literate.

Cummins (1978 in Laija-Rodrigues et al, 2006), proposed a threshold hypothesis by which to resolve the visible discrepancies in the relationship between bilingualism and cognition. It provides a framework by which one can predict the academic and cognitive effects of bilingualism. Additive bilingualism will lead to a high threshold with a high level of bilingual compliance and high levels of proficiency in both languages with positive cognitive effects (Cummins, 1978 & 1979).

The threshold hypothesis is interrelated to the developmental interdependence hypothesis, which states that the second language's level of competency is a function of the nature of competency developed in the first language when exposure to the second language commences. This hypothesis stipulates that there is an interaction between the language of instruction and the level of competency developed in the first and the second language (Cummins, 1979). Essentially, when learners have a firm foundation of knowledge and skills in their first language, these skills and knowledge can be transferred to the second language. Along with this, CALP (cognitive academic language proficiency) is a particular threshold, which includes language for cognitively demanding tasks and is needed for advanced cognitive and academic growth (Laija-Rodrigues et al, 2006). Moreover, advanced conceptual understanding in the first language predicts similar conceptual understanding in the second language (Cummins, 1992).

International studies have investigated the effectiveness of bilingual education and provided sufficient proof in favour of bilingual education, especially that of additive bilingual education as an effective means of encouraging bilingualism in a multicultural context (Cummins, 1979 & 1992). South African research in this regard is needed, though, while tests that are available in the two languages of instruction in such programmes (also language proficiency tests), are essential for the evaluation of the effectiveness of these programme**.**

1.6.2 Psychological testing and language testing in South Africa

In South Africa, psychological testing is still a new method of assessment when compared to other countries. However, South Africa’s heritage is an aspect that affects assessments and is considered in conjunction with psychological testing (Meiring, Van der Vijver, Rothmann & Barrick, 2005). This heritage and our diversity in terms of languages and cultures therefore make the testing of, and in, different languages very important and the translation and adaptation of such tests are not only inevitable, but also indispensable. As a result, one cannot overlook the fact that scores obtained in one culture or language cannot merely be compared with scores obtained in another culture or language.

In the past, when psychological tests were used to assess non-white learners, the tests used were initially standardized for white learners (Van der Vijver & Leung, 1997). This essentially means that non-white learners would score less or differently on these tests as the items on the test, only standardized for white learners; the lower scores may have been because of test bias or incorrect norming. A biased test may under-predict standard scores for non-whites.

By being cognizant of the South African diversity, one can address previous inequalities observed in testing. Consequently, the government implemented the Employment Equity Act 55 of 1998, section 8, in which it states that "psychological testing and other similar assessment are prohibited unless the test or assessment being used a) has been scientifically shown to be valid and reliable, b) can be applied fairly to all employees and c) is not biased against any employee or group" (Republic of South Africa, 1998*).* This act thus encourages adherence to regulations as well as good quality research findings (Meiring, Van der Vijver, Rothmann & Barrick, 2005). One way of addressing this issue, is to adapt tests into the primary languages of learners.

The distinction between psychological testing and language testing is in that educational or psychological tests are methods created to deduce particular behaviour from which one can make conclusions about particular characteristics of an individual

(Carroll in Bachman, 1990) whereas language testing enables the researcher to focus on particular language abilities of individuals (Bachman, 1990, Pray, 2005).

As a result, tests, including language tests, must be useful and meaningful in the context of their use, as well as valid and reliable for use in diverse groups. Validity is when one is able to measure what one intends to measure without threats filtering in (internal validity) and reliability is understood as the quality of test scores in their perfect reliable score which is free from measurement errors (Bachman, 1990).

Educational testing in South Africa

Educational testing in South Africa, which is often referred to as a multilingual context due to the diversity of languages and cultures, can be quite problematic. The availability of instruments used to test learners is not always available in the primary language of the test-taker. This can cause serious implications as the learner is being discriminated against in the form of the test's language. Thus, test developers have been beseeched to develop tests in the language of the learner. This places pressure on test developers to change the items and restructure the format or phrasing in order to ascertain that the same construct and content is covered in the various language versions of the test. More training of how to appropriately adapt tests in a multilingual context should be given, as the expertise in this field of testing is limited and should increase. The increase of expertise will then serve to increase the quality and eventually the quantity of multilingual tests in South Africa (Foxcroft, Paterson, Le Roux & Herbst, 2004).

It should also be noted that research institutes such as the HSRC are actively trying to solve issues related to testing and have theorized the existence of an encompassing body to govern testing and to ensure that there is proper control and regulate the use of tests in South Africa. Such a body would then monitor test adaptation and translation as well as issues relating to tests such as bias and fairness. The controlling body which is referred to as the Centre of Excellence for testing, is not only an excellent idea but will advance the skills and capabilities of test developers in South Africa (Foxcroft, Paterson, Le Roux & Herbst, 2004).

The ethical procedures involved in testing is essential to the adaptation and translation processes and issues mentioned above can be linked to fairness in testing and issues around bias. International standards for psychological educational testing (www.ipacweb.org), the International Testing Commission (www.intestcom.org) as well as the Code of conduct in South Africa (www.ipacweb.org), all act as points of reference to ensure that good ethical testing is done which does not discriminate or disadvantage any test taker. Such bodies help to guide the work done by test developers and provide useful guidelines to guarantee that tests are valid and reliable as well as fair and unbiased. Under the International Standards for Educational Testing (www.ipacweb.org), there are certain obligations which must be followed to encourage fairness such as: if scores across languages differ, one must obtain evidence of validity for each subgroup, a test may only be used for a group if it is valid for that specific group and one must evaluate the group differences and establish that it is not based on content-irrelevant (content validity) and construct-related problems (construct equivalence) (www.ipacweb.org).

The regulations regarding testing individuals from different linguistic backgrounds emphasize the following: the test should be designed in order to reduce threats to validity and reliability of test score conclusions which may occur from language differences, the administration of the test should be done in the test taker's most proficient language unless proficiency in the less proficient language is part of the measurement, equitable treatment of test across languages should be ensured and the design of the test should also attempt to reduce invalidity based on language differences (www.ipacweb.org). When considering the rules and regulations implored by such institutes, one is conscious of the appropriate governing bodies in charge to tackle the current testing-related issues, as well as stressing the importance of testing fairly across groups.

1.6.3 Test Adaptation and Translation

As argued before, tests (including language proficiency tests) in multicultural societies such as South Africa must undergo a process of adaptation and translation in order to address the issue of language testing in more than one language. Test adaptation is defined as a process in which translators seek to replace constructs of one language to another language, which must be psychological, culturally and linguistically equivalent to the original language. Thus, it is not merely literally translating the test content from one language to another language, because by doing so one is not sure that the constructs that are being tested are the same in both languages (Hambleton, Merenda & Spielberger, 2005).

When one is not sure of the accuracy of these constructs, one cannot make comparisons between language versions because one is not sure that the same concept in both languages is being measured. Thus if the concepts and tests are in-equivalent, then the results yielded would be flawed and it would be unethical to draw conclusions based on such findings. Hence, structural equivalence exists when two tests are viewed as being structurally similar and they have reciprocal ties. Consequently, the items on the different tests should be perfectly exchangeable or substitutable (Hambleton, Merenda & Spielberger, 2005).

The International Test Commission (ITC) is a commission comprised of an international committee of cross-cultural psychologists, which assist researchers in the test adaptation process. They published a set of guidelines known as the ITC guidelines and by following these guidelines, one can ensure test adaptation and development is done accurately (www.intestcom.org). The guidelines encourage test developers also to make use of statistical techniques in order to evaluate structural equivalence and to identify whether the test is adequate for the intended learners (Sireci & Gonzalez, 2003). Thus the test adaptation guidelines serves to contribute to the quality of test adaptation as well as increase the validity of cross-language and cross-cultural findings. Examples of these guidelines are namely (www.intestcom.org):

> *D1: Instrument developers / publishers should apply appropriate statistical techniques to, 1) establish the equivalence between the different versions of the instruments, and 2) identify problematic components or aspects of the instruments, which may be inadequate to one or more of the intended populations.*

> *D9: Instrument developers / publishers should provide statistical evidence of*

the

equivalence of questions for all intended populations.

In addition to this, the quality of translations of different language versions is done by, namely: proofreading, back-translation, inspecting for clear meaning of the sentences and checking the level of wording. Quantitative methods such as item analyses and item bias analyses, and the evaluation of reliability, validity, and construct-bias need to follow good translation practises to allow researchers to assess the quality of translations (Beller, 1995; Hannemann & Riddle, 2005; Leong & Austin, 2001; Van der Vijver & Rothmann, 2004).

1.7 Outline of chapters

With the introduction of chapter one, the commencement of the following chapters serve to substantiate the themes briefly mentioned in the above sections.

Chapter 2 provides a description of the theoretical framework being used and studies that have used the same theoretical guidance. This will ensure a deeper understanding of the theory underpinning the present study.

Chapter 3 provides an overview of language testing in both an international context and national context, indicating the misuse and abuse of testing in two languages. This chapter will serve to explain the nature of research that the present study is embarking on.

Chapter 4 is a discussion of the methodology used in the present study and includes the brief overview of how the data was analyzed. This section also includes the ethical considerations that the study as well as the larger study adhered to.

Chapter 5 involves the results and discussion of the different analysis techniques stipulated in the objectives of the study as well as the factorial analysis of the data. Thereafter, the interpretation of the results will follow.

Chapter 6 is the discussion of the implications of the results previously observed as well as the conclusion of the study in which the limitations and future recommendations are made accordingly.

CHAPTER TWO

EQUIVALENCE AND BIAS

2.1 Introduction

The theoretical framework of equivalence and bias (Van der Vijver & Rothmann, 2004), as discussed extensively in the literature dealing with multicultural and multilingual psychometric theory, was chosen for this study because it provides a good model within which to interpret the ITC's guidelines (www.intestcom.org); thus the researcher is following the international conventions and requirements.

Within this framework, procedures ensuring linguistic equivalence are needed to be followed as a first step to support the accuracy of the translation and adaptation process, which is then verified by the process of statistical equivalence. The verification of equivalence, more specifically the structural equivalence of the two language versions of the test, is what the present study addresses, as establishing structural equivalence ascertains the presupposition that equivalent constructs are present in both language versions.

This chapter will thus explore the concepts of equivalence and bias, and will discuss some research studies related to equivalence, specifically the structural equivalence of tests. This will enforce a greater understanding and knowledge of related issues in the field of language testing. The comprehension of equivalence will also serve to initiate appreciation for the statistical techniques used in the methodology chapter.

2.2 Equivalence

Test adaptation is accurately understood as adapting the language of the original test to another language by attending to linguistic equivalence issues, such as psycholinguistic processes in measurement across languages and cultures. Awareness of different languages and cultures, i.e. isiXhosa, assists in the adaptation process, as the researcher's sensitivity to the IsiXhosa culture will act as a means of ensuring validity and prevent such problems filtering in. By being cognizant of the different meanings that people attach to different constructs, one is able to adapt tests which then specifically measure the desired construct. If they are both proved to be measuring the same constructs, then equivalence has been achieved. If not, then the two versions of the test are declared non-equivalent (Foxcoft & Roodt, 2006).

According to Van der Vijver & Rothmann (2004), equivalence is viewed in terms of a hierarchy and consists of three levels, namely: structural (construct or functional equivalence), measurement unit equivalence and scalar equivalence. Structural equivalence is the lowest and initial level of equivalence and it emphasizes the entire validity of the underlying psychological construct across versions of the test. The second is measurement unit equivalence, which is obtained when two language versions of a test have different origins, but have identical measurement units (such as for example Celsius and Fahrenheit measures of temperature). Finally, scalar equivalence is the highest level of equivalence and is when measures have identical measurement units and origins across cultural groups, and measure the same constructs (Meiring et al, 2005).

Structural equivalence is achieved when two tests are viewed as being structurally similar and they are noted to have reciprocal ties with one another. This implies that the items on the different tests should be perfectly exchangeable or substitutable when structural equivalence is achieved. Furthermore, structural equivalence allows us to make correct conclusions based on the fact that the tests were proven to be identical in nature (Hannemann & Riddle, 2005). Hence structural equivalence is when the same construct is measured in both groups (Van der Vijver et al, 2004).

The equivalence of concepts is ensured with the use of psychometric properties of the instrument used to measure the concepts. Construct equivalence measures operational definitions of constructs in each language and cultural group and is therefore a precondition when doing cross-cultural and cross-language comparisons. The use of statistical techniques will then be used to establish the equivalence of the two versions of the test and will assist in locating any problematic constructs found in these versions (Leong &.Austin, 2001).

2.3 Bias

Equivalence is evaluated by assessing bias, which is defined as the presence of nuisance factors affecting the test scores of different groups differentially. The importance of assessing and identifying bias is that the presence of bias prevents one from comparing scores; if bias exists, equivalence are severely threatened. There are three types of bias, namely: construct, method and item bias. Construct bias refers to the differences in constructs across cultures and languages. Method bias refers to all the conflicting factors that are associated with methods and includes instruments and administration bias. Item bias (Differential Item Functioning) refers to nuisance factors identified in the items of the test (Foxcroft & Roodt, 2006; Hambleton, Merenda & Spielberger, 2005; Koch, 2005; Van der Vijver & Rothmann, 2004).

In this study, the researcher will be specifically focusing on construct bias, as it affects structural equivalence (Koch, 2006; Hambleton, Merenda & Spielberger, 2005). Construct bias occurs when there is only a partial overlap of the construct across cultures or languages (Meiring et al, 2005; Van der Vijver & Rothmann, 2004).

The subsequent section assists one in developing an understanding of how research on structural equivalence has been conducted in the past and what the implications are.

2.4 The Spanish PAA test and the English SAT: an example of the practical implications of evidence of the equivalence of two language versions

The translation, and subsequent scaling, of a test from one language to another is rather complex and must be done with the use of proper psychometric equating methods. When translating from one language to another language and culture, the words and concepts do not constantly take on equivalent meanings or difficulty levels (Beller, 1995). This then necessitated the evaluation of equivalence.

Angoff and Modu (1973 in Beller, 1995) evaluated the equivalence of a test given in two languages between the verbal and mathematical scores on the College Board Spanish-language Prueba de Aptitud Academica (PAA), and the English SAT by conducting an exploratory factor analysis. In this study, the data resulting from the common items were used to equate the tests, the tests were standardized, and the results used to explain the differences in abilities between two groups of students.

The fundamental assumption underlying this and other studies that evaluate equivalence of tests is that the difference between the means of the difficulty values for the two groups is an indication of the difference in their ability levels. Researchers then assume that the test measures the same trait or construct for both groups, and they specifically focus on items that do not conform to the general pattern (Beller, 1995). Underlying the equating methods is thus the assumption that the relationship between the common items and the whole test is the same for the two groups (Beller, 1995). Consequently, structural equivalence exists when two tests are viewed as being structurally similar and they are noted to have reciprocal ties with one another. This implies that the items on the different tests should be exchangeable or substitutable when structural equivalence is achieved. Furthermore, structural equivalence allows us to make correct conclusions because the tests were proven identical in nature (Hannemann & Riddle, 2005; Leong &Austin, 2001; Van der Vijver et al, 2004).

2.5 The nature of structural equivalence across different language versions of tests

Cross-cultural researchers who emphasize the use of tests in multiple languages also emphasize the awareness of bias entering such tests, as they can skew results. The adherence to the regulations of the ITC and their guidelines for adapting educational and psychological tests, especially in which they state “instrument developers / publishers should apply appropriate statistical techniques to establish the equivalence of the different versions of the instrument…” (Sireci, Harter, Yang & Bhola, 2000, 3; www.intestcom.org).

Presenting descriptive statistics for each language version is very important and assists the researcher in assessing the structure of the test and initially comparing the means across the language versions. The ultimate aim of such research is to confirm that any differences noted in the tests are not linked to the language version of the test, but rather the test itself. This then informs the researcher that the test was successfully translated and adapted. The insistence on evaluating the structural equivalence is thus impressed on all cross-lingual researchers, because it is primarily a validity issue and considers the different cultural nature of the different language versions (Sireci, Harter, Yang & Bhola, 2000)

To echo this, Messick (in Sireci, Harter, Yang & Bhola, 2000), emphasized the imperative nature of evaluating structural equivalence in that it represents "the extent to which a measure displays the same properties and patterns of relationships in different population groups and under different ecological conditions" (in Sireci, Harter, Yang & Bhola, 2000, 18). Thus, comparisons across different language groups cannot be made if equivalence if not established (Sireci, Harter, Yang & Bhola, 2000).

There are two broad divisions within which to define equivalence, such as the interpretative and procedural components (Welkenhuysen-Gybels & Van de Vijver, 2001). The interpretative component focuses on commonalities in interpretations while procedural equivalence deals with constructs and investigates the operationalization of underlying concepts. Structural equivalence is thus the initial level of procedural equivalence. Exploratory factor analysis is a common method for evaluating structural equivalence and allows one to adequately assess the structure of the data (Welkenhuysen-Gybels & Van de Vijver, 2001).

Part in parcel with factor analysis is the use of the Tucker's Phi index as an agreement index. It is often utilized in studies done on the structural equivalence of different language versions of tests and becomes quite vital in the sampling and re-sampling of data. The re-sampling of data with regards to the use of the Tuckers Phi is done in order to generate high critical values, and is done at the start of either the bottom-up or top-down approach to sampling multiple groups (Welkenhuysen-Gybels & Van de Vijver, 2001). The use of Tucker's Phi is essential in order to establish if tests are structurally equivalent by assessing their value obtained, which should be above 0, 9. The Tucker's Phi as an agreement index in the present study will be discussed and explained in greater detail in the methodology chapter.

When taking the above into consideration, one is conscious of the reality that tests adapted for use across languages is not by any means a simple process and involves highly technical experts to be part of the process. The complexity of evaluating construct equivalence in adapted tests across languages can present researchers with a multitude of problems and can discourage intensive investigation. Exploratory factor analysis that includes common factor analysis and principal components, are commonly used methods for evaluating for construct equivalence and involves analyzing the data for each language group separately and then comparing the results (Sireci, Bastari, Xing, Allalouf & Fitzgerald, 1998).

The current study attempts to defeat the methodological limitation of prior research (Koch, 2009), by matching groups. As a result, the same items appear in both the language groups due to the removal of the no-variance items. This will be adequately discussed in the results chapter. In addition to this, there are two other methods frequently used for evaluating construct equivalence, such as MDS (Multidimensional Scaling) and Confirmatory factor analysis (Sireci, Bastari, Xing, Allalouf & Fitzgerald, 1998). Previous research on the construct equivalence of the subtests of the WMLS made use of WMDS (Weighted multidimensional scaling) (Koch, 2007) and thus by using exploratory factor analysis, one may anticipate different results.

2.6 Construct equivalence: Three case studies

As part of comprehending construct equivalence, one must be aware of how it has been applied and interpreted. For this reason, the selection of three case studies is used as illustrations of construct equivalence, whereby one can determine the effective nature of the analysis technique. These case studies are namely: the 15 FQ, the Russian and Hebrew PET and the Microsoft Network Technology Exam. One of these, the 15 FQ, was available in English, while the other two are available in different languages versions. These studies will then serve as practical applications of evaluating for construct equivalence.

2.6.1 The 15 FQ

In a study exploring the adaptation of the 15 FQ (Fifteen-factor questionnaire) in South Africa, the researchers evaluated the internal state of the questionnaires with a series of methods, namely: exploratory factor analysis, cluster analysis and multidimensional scaling (Meiring, Van de Vijver & Rothmann, 2006). The exploratory factor analysis makes use of Tucker's phi as the factor congruence coefficient with target rotation. The 15 FQ was only available in English and adapted for the South African context, thus the construct of this adapted version was assessed. In all these approaches, it was found that the 15 FQ was not that well adapted and there were problems in the construct equivalence across the English and Afrikaans speakers, and the speakers of African languages. An additional problem with the 15 FQ, is that the reliability score in the African language speaking group was quite low indicating that the test not suitable for cross-cultural equivalence and is not appropriate for high-stakes testing. An interesting and constant theme in tests being adapted for the South African context is that the languages of tests are problematic for African speakers and is potentially biasing them. This study thus echoes the importance of construct equivalence and adapting tests into multiple language versions. Furthermore, problematic tests should be redeveloped and should adhere to the stipulations of the Employment Equity Act (Meiring, Van de Vijver & Rothmann, 2006).

2.6.2 The Russian and Hebrew PET

In an analysis of the Russian and Hebrew language versions of the PET (a psychometric test battery used in Israel for entrance into higher education), the researchers made use of exploratory factor analysis, MDS and confirmatory factor analysis in order to analyze the data. These techniques served to complement each other. All three analysis techniques worked well together and confirmed the content structure of the test by indicating that the same construct was present across the two language groups. The PET displayed few differences on the items present across the two language versions, yet not enough to hinder structural equivalence (Sireci, Bastari, Xing, Allalouf & Fitzgerald, 1998).

2.6.3 The Microsoft Network Technology Exam

The Microsoft Network Technology Exam is a crucial exam for all systems engineers in order to become certified (Sireci, Bastari, Xing, Allalouf & Fitzgerald, 1998). This test was administered across languages by matching groups on ability, thus allowing one to be certain that the differences observed are linked to language and not ability. Therefore establishing construct equivalence across the different languages was crucial. The researcher has performed principal components analysis (PCA), MDS and CFA on the data available. The researcher would have hoped that these techniques would have complemented each other, but instead they seemed to contradict each other. The results for the CFA (indicating structural equivalence) contrasted the PCA results (disproved structural equivalence). Based on these results, one could deduce two conclusions. Firstly, that there are structural differences between the different language versions and CFA has been identified as not being stringent enough to observe differences. One could also speculate that there are no structural differences, thus accepting the results of the CFA and hypothesizing PCA and MDS are merely identifying errors in the data. The importance behind such research is that it creates awareness of the fact that different techniques yield different results and thus one may either contradict or affirm previous literature depending on the findings of the analysis technique used (Sireci et al, 1998).

2.6.4 Summary of the three case studies

In sum, these case studies indicated that exploratory factor analysis should ideally be used for uncovering differences across groups in separate analyses for each group, which is what the present study attempts to do. Weighted MDS on the other hand, is recommended to assess the differences across groups in a single analysis, which is what Koch (2009) previously explored. CFA is then recommended for instances when the researcher is interested in a certain factor structure and is a method that can be used in future research for both the VA and LWR scales. The significance of the CFA is that it serves to confirm the previously found factor loadings in prior exploratory factor analysis investigations. This method of analysis (CFA) therefore either affirms or denies the existence of these factor loadings, thereby allowing the researcher to establish whether the previous results are consistent (Sireci et al, 1998). With this in mind, one is cognizant of the relevance of the exploratory factor analysis for the present study.

2.7 Fairness

The importance of evaluating equivalence across language groups is embedded in the concept of fairness. The concept of fairness is not always considered and often unfair testing takes place. The growing awareness of fairness in testing however is acting as a catalyst of change as it is curbing the escalating use of unfair tests. Fairness essentially refers to the social implications of tests such as exclusion from educational institutions and warns against the negative inferences made if tests are used in unethical ways. Fairness refers to the adverse impact that tests, and in the present study, it refers to the difference in educational opportunities for both English and isiXhosa groups. This means that when evaluating whether a test is fair, one establishes that one group is not being favored above another group (Koch, 2009).

Moreover, a test cannot and should not be normed and standardized without an inspection of fairness and bias. Bias then serves as a technique that informs one about differences in constructs across groups and reinforces the investigation into fairness in testing. Simply put, bias and in-equivalence have implications for fairness and thus it is worth discussing, as one must be conscious of the impact that tests can have on individual's lives. Debates regarding fairness in testing take the form of philosophically and legally grounded ideas. Thus quantitative studies have sought to evaluate for bias and have assessed the implications of laws such as the previously mentioned Employment Equity Act and Higher Education Act (Koch, 2009) to justify these evaluations. These acts serve to promote fairness in testing yet the tests used in both the work and educational context do not adhere to these regulations. This means that people continue to be disadvantaged and disempowered by tests, as the content and / or languages of tests favors certain groups of people.

The researcher's awareness of such issues has motivated the current study, as this will allow one to discuss with the use of statistical techniques, issues of bias that impact on fairness. For this reason, the Verbal Analogies test and Letter-Word Identification tests are being scrutinized as proper attention must be paid to these tests and recognition must be given to adequately adapted versions of the WMLS test. Fundamentally, with statistical techniques, one can establish the existence of an identical construct in both the reference group (English version of the WMLS) and the focal group (isiXhosa version of the WMLS). Thereafter, one can proceed into further investigation into scalar equivalence.

2.8 Conclusion

The importance of grasping the technicality of equivalence lies in that it is deeply embedded in measurement theory. For this reason, the researcher sought to briefly explore the concept of bias and equivalence, which are mutually vital concepts in this paper. Thereafter, a look at practical implications and applications of construct equivalence allowed for a more comprehensive picture. This informs one of how equivalence as a theoretical framework can function and inform one's results. This therefore allows one to conclude whether two tests adapted for one purpose is measuring identical constructs in both versions of the test.

CHAPTER THREE

EQUIVALENCE AND BIAS IN LANGUAGE TESTING

3.1 Introduction

Cross-linguistic research can be quite a complex process in that test translation and adaptation is both time-consuming and requires highly specialized tools and knowledge. Once again, the importance of this study is echoed and a dire need to advance South African cross-linguistic research is evident. The importance of assessing structural equivalence was recognized in earlier research in the 1980s, because prior to this, there was no necessity for such research. A serious interest in making cross-linguistic comparisons was growing rapidly and testing thus became critical in doing so (Meiring, Van der Vijver, Rothmann & Barrick, 2005).

When adapting instruments from one language to another language, the preservation of the original psychometric properties of the original language is important. It should seek to specify the required scoring and should be sensitive to peculiarities (Lauterbach, Martins, Garcia, Cabeca & Ferreira, 2008).

Testing language proficiency is an area, which is continuously revised by language researchers and test developers (Berkmen, 2002). Thus, this chapter will explore the operationalization of language and language testing in order to get a broad overview of its importance. Along with this, a short overview of how language tests are utilized will be explored, as they have different uses such as placement, citizenship and admission. Furthermore, language testing may involve testing in more than one language and this will be briefly explored. In addition to this, discussions on language testing with emphasis on critical language testing and testing in two languages will also be explored. The chapter will then conclude with the relevance of language testing in society and specifically for a multicultural society like South Africa. This will serve to corroborate the significance of this study which was stimulated by the evaluation of equivalence and bias in the previous chapter.

3.2 The operationalization of language

When measuring language, the concept of language takes on a different form, a measurable form. The measurable form of language is essentially the operationalizing of language in a manner that makes testing possible. The operational description of language includes its "phonological, lexical, semantic, morphological, syntactic, discourse and interactional level" (Auer & Wei, 2007, 247). All language aspects are very important and should be assessed when testing in two languages.

Along with this, language testers include scales by which to locate the level at which the individuals linguistic functioning is at, thus rating their linguistic performance. There are three distinct features present in the measurement of bilingualism, namely: linguistic proficiency, linguistic competency and developmental trajectories. Linguistic proficiency can be understood as the ability to communicate in an additional language, whereas linguistic competency refers to their ability to make grammatically correct expressions (Chomsky, 1965). Developmental trajectories, on the other hand, can be understood as individuals' way of learning over time, or their cognitive ability to develop an additional language in this case (Cicchetti & Walker, 2003).

When measuring bilingualism, there are many approaches to studying an individual's competency in two languages. These approaches are often guided by a specific focus on language and are often tied to another discipline such as education, psychology and cognitive neuroscience. Different disciplines focus on certain aspects, such as psycholinguistics which studies how one mind is able to process two languages and the psychometrics related to this would ensure that the instrument measuring this is valid and reliable (Auer & Wei, 2007).

3.3 Two cases of the misuses of language tests for citizenship

When establishing individual's language competency or proficiency in a language, one must always ensure that appropriate measures and procedures have been followed in order to not disadvantage any individuals. In addition to this, the testing of language is often used as a key issue in granting citizenship in different countries and thus a review of such practices provides one with insightful information about language testing, especially the misuses of such testing. For this reason, the selections of two case studies are explored as a means of discussing how language testing can be misused.

3.3.1 The Case of language tests for Asylum seekers

The relevance of this case is that this is one of the examples in which multilingualism was not fully recognized and the implications of such language tests can be catastrophic in nature. The problems associated with language tests are clear in this case, as the Belgium government was not acknowledging the differences in the languages of asylum-seekers, thus instead of recognizing linguistic diversity, it refutes it's existence. When testing in multilingual societies, one cannot afford for such language discrimination, as one is denying their human right to communicate in their primary language.

The language choice of English is forced upon asylum seekers in Belgium and is understood as the language of interaction in Belgium. English is also registered as the procedural language when interviews take place and often interviewees must sign documents declaring their language of choice. The problem of language discrimination comes in when the officials (mis)inform these asylum interviewees of their choice to be interviewed in a language other than English (Maryns, 2006).

These interviewers, knowing very well that the asylum seekers are not fully competent in English and thus the struggle to comprehend what message is being explained to them, persuade them to commit to English as the language choice of the interview with insufficient knowledge. This lack of knowledge in English that the asylum-seeker displays can be defined as experiential narration, in which English is mixed with other languages such as Creole (West-African language) in order to explain the context of war that the individual experienced (Maryns, 2006).

Linguistically this interviewing process can be understood as resources being displaced. Part of the interviewing process involves that the individuals undergo a language test as a means of receiving asylum in Belgium. Foreigners thus stretch their bilingual nature by trying to communicate and answer correctly to the questions posed. Part of this language test is that of translation tests, in which they are required to translate words from English to their language and vise versa. This type of linguistic identification in asylum interviews are biased and should not be done as multilingual individuals should not need to undergo translation tests as a requirement of their asylum procedure. Moreover, one cannot wish to make reliable assessments when using such tests as the criteria, as its biased nature will serve to exclude individuals and present them from attaining asylum in countries such as Belgium (Maryns, 2006).

3.3.2 Language testing for citizenship

The Swedish government implemented language testing in Sweden as a criterion by which to regulate the number of individuals entering the country. For this reason, Swedish language tests were introduced as a compulsory part of the naturalization process, because the Swedish language is a prerequisite for citizenship in Sweden. Consequently, a level of proficiency would be set in order to distinguish passing from failing. Some regard this language test as motivating migrants to learn the Swedish language, as they will attain the rites afforded to other Swedes. Hence, the Swedish government implemented this language test as a means of establishing foreigner's language competency and this was substantiated with the claim that it would assist foreigners with integrating better into Sweden (Milani, 2007).

This language test caused immense controversy in the Swedish parliament and the media and public appealed against this, as this was discriminating and eventually, the language test requirement for naturalization was denied. The example of the Swedish language testing government policy was evaluated and criticized by Milani (2007) in which he claims that language testing for naturalization is part of an 'ideology of language testing', being that it "attempts to defy multilingualism and multiculturalism by tying proficiency in one language to knowledge of one culture as the compulsory prerequisites" (Milani, 2007, 246) for granting citizenship in Sweden.

When examining the misuses of language tests, especially in the cases of asylum-seekers and citizenship, one notices the power of testing individual's language competency and thus one must critically review such practices. Moreover, this leads one to evaluate language testing as this is essential in terms of fairness in testing.

3.4 An evaluation of language testing

As argued above, the evaluation of language testing is vital as one must critically review such practices. Possible problems which can be leveled at language tests in general, is that no matter what purpose it is used for, it predominantly adheres to a psychological framework instead of a psycholinguistic one. This method of testing is characterized by the focus on psychometric information and falls short of adequately assessing language knowledge. This implies that individuals' language vocabulary is not fully assessed in such measures and more thorough measures should be followed. Furthermore, when assessments show low levels of validity and reliability, it is indicative of measurement problems. Both issues result in misplacement of individuals in special programs, as these measures can not sufficiently guarantee one that the results obtained are valid (Pray, 2005).

Moreover, language assessments can be labeled as being either prescriptive or descriptive depending on the type of assessment. Prescriptive assessments give one criteria by which to view individual's knowledge in language whereas descriptive assessments merely offer an overview of individual's language. When assessments are labeled as prescriptive, it does not regard the underlying theory that guides the assessment, but rather serves to establish whether mother-tongue learners are not proficient, thereby categorizing them into language groups. This prescriptive nature should be avoided in language assessments, as it requires that these learners adhere to a set of rules relating to language knowledge and deviation thereof is viewed as failure. Instead, the descriptive approach should be followed, because it considers the variation in language knowledge and ability (Pray, 2005).

In addition to the above problems, studies done on the testing of the language abilities of mother tongue and second language learners of a language revealed that a contrasting level of knowledge existed between the two, which is often attributed to the learner's knowledge of their subjects and not specifically their language knowledge. There is however, suggestions that one must measure language proficiency and academic achievement separately, as they are different constructs, even though they correlate well (Pray, 2005).

The use of cut scores to identify the high-risk learners from the low risk learners based on language tests in general can also be problematic, as the researcher must be aware of the standard error of measurement and thus no absolute scores can serve to identify risk groups. Decision-making of such nature carefully considers the validity of test scores and the overall test (Mahoney & MacSwan, 2005). The importance of language testing and its accuracy is evident in the nature of such testing, as one is hesitant to simply classify learners according to their language knowledge. This once again, stresses the urgency for good language instruments as it serves to measure vital language skills from which inferences about the learner’s language knowledge and skills must be made.

The evaluation of language testing is important as this creates an awareness of the issues pertaining to testing. Furthermore, language tests are used to categorize individuals which can be hazardous in high-stakes situations. This then makes space for critical language testing, which centres on examining the power governing language tests. For this reason, a synopsis of critical language testing follows to expand this discussion on language testing.

3.5 A synopsis of critical language testing

Critical language testing is grounded in social theory which attempts to deconstruct visible power relations that assist in creating social or even educational inequalities between individuals or learners in certain contexts. In addition to this, critical language testing theorizes that individuals have differentiated access to language, and language is understood both as a resource and practice. Therefore, language testing indicates the unequal distribution of linguistic resources, as individuals do not have access to the same resources (Milani, 2007).

Shohamy’s (2006) recent work on critical language testing explains language testing as a medium and an object for making policies in government. In addition to this, language tests are methods employed by government in which they decide which languages should be assessed, and this affects the education of learners as they decide on what genuine language knowledge is, or what languages are appropriate. Furthermore, the power of language tests is quite symbolic as it is able to socially categorize individuals in terms of inclusion and exclusion (thus creating cut scores to establish which individuals pass or fail as well as deciding who may be tested). The problem with this is that there are values attached to such categories and such categories dictate the future of many individuals (Shohamy, 2006). Thus, such testing is regarded as high-stakes testing.

3.6 The context of high-stakes language testing

Language testing can be understood as a rite of an institution, such as an educational domain, because it is a socially performing action that creates a social frontier. This implies that language testing acts as a social ritual by which individuals comply with the prerequisites specified by the government. Along with this, an identity is ascribed to the individual depending on the result of the language test, meaning that they are denied citizenship or in an educational domain, they receive an undeserving status. The identity ascribed to passing or failing tests creates a boundary between the two defined statuses and separate rites are afforded to these groups (Shohamy, 2006).

Critical research on language testing, especially that of high-stakes language testing, indicates that there are negative psychological consequences involved on the part of the test taker. The test taker, depending on their result, has a level of language proficiency ascribed to them that then affects their language learning. Thus, there is a paradoxical nature associated with testing in that it seeks to socially integrate individuals, but rather serves to create discourses of inclusion and exclusion, thereby reinforcing social separation. Furthermore there is an assumption that a positive correlation exists between language tests and language proficiency, therefore a indicating the existence of a strong relationship between the two. This assumption (often unsubstantiated) is very important in multicultural societies because it helps with individual's assimilation into dominant cultural practices and their inclusion and acceptance from others in a specific context (Shohamy, 2006).

Testing in American societies occurs quite frequently and amoung young children, thus there is no paucity of research on their language development and the effectiveness of language tests (Valdes & Figueroa, 1994). This is however not true for South Africa, as few language tests are standardized for all. This means that there is inequality associated with language tests and the effectiveness of such tests is thus questioned because they do not cater for all languages. Consequently, this creates a discrepancy between the results obtained by one cultural group opposed to another cultural group.

This makes the relevance of testing in two languages very important. The unfortunate reality of this is that one is forced to depend on international literature. As a result, the dependence on international literature limits the discussion on testing bilingualism because one cannot refer to appropriate South African studies done. In contrast, it highlights the dire need for such research, because an accumulation of such research can act as baseline information for further research. Moreover, knowledge in the field of language testing is critical and can solve many academic problems that learners suffer from and can eventually promote better grade 12 results for non-white learners.

A noteworthy question that one could pose to such an argument would be how one could possibly resolve such issues. This is an ongoing challenge and debate amoung interested researchers as the limited expertise in this area of testing in South Africa leaves one despondent and requires that they settle for plausible alternative. Evaluation of bias and equivalence is thus an extremely important technical means of assessing whether any group is being disadvantaged or discriminated by language. More training and emphasis should be channeled into the sphere of testing, especially that of testing in two languages and this will encourage the growth of expertise in this area of testing and eventually improve the current testing arena in South Africa.

3.7 Testing of language in two languages

In language testing, especially that of testing of language in two languages, one must be cognizant of the time in which the acquisition of the second language takes place. The research that was done on second language proficiency serves as evidence that the learning of a second language (English in a context such as SA) can take a number of years (nine years) and is not a one-year solution. The question of when the ideal time period to introduce English as an additional language is difficult, as it differs depending on the context and sample, although early introduction to English is encouraged (Mahon, 2006).

Testing of language in two languages is vital, as the language acquisition of the second language learner is an indication of their performance and academic achievement in their first language. In international studies, the testing in two languages has become part of their educational system in that the performance of second language learners is constantly measured in order to establish their level of proficiency in Spanish (their primary language) and English (their additional language) (Mahon, 2006).

Testing in two languages in multicultural societies is rather contrasting in terms of the given context such as the South African and American context. The American context boarders on over-testing their learners and they have a magnitude of literature to refer to in relation to adaptation and translation of tests. Thus, they have firmly grounded themselves in this area of testing for their context. Unfortunately, South Africa suffers from a lack of literature and expertise, not forgetting the shortage of appropriately adapted instruments for cross-cultural use. One must however view this study as an opportunity to embark on new research and to immerse oneself into a testing evolution.

3.8 Language testing in education in multicultural societies

Bilingual education is very important in South Africa because of its rich and diverse cultures. For this reason, language developments of learners in both languages are essential to promote bilingualism and thus language tests should be available in both languages. One cannot however ignore the power of language, especially the power of the English language. In line with this is the appropriateness of the measures used to assess individuals in both languages and evaluating for bias and equivalence such language assessments becomes indispensable.

When considering testing in South Africa, sensitivity towards multiculturalism is important. Thus, emphasis is placed on accurately testing black South Africans, as they were previously disadvantaged in the sphere of testing. There are relatively few studies in South Africa that concentrate on the language proficiency in education and even less on the testing of CALP on school learners in more than one language, especially when one of the languages is an indigenous African language.

This is not due to a lack of interest, but more an indication of the paucity in relevant literature and knowledge pertaining to this field of study, especially in South Africa and in the context of additive bilingual education. In these contexts (of additive bilingual education), the importance of tests of academic language proficiency in two languages (the primary language and the language of power) then becomes crucially important.

3.9 The evaluation of equivalence in languages tests: Two case studies

Throughout the discussion on language tests and the different aspects pertaining to language testing, the argument for testing in multiple languages was subtly arising, as a dormant need for this was escalating throughout multicultural contexts. The existence of bilingualism as well as multiple languages within which to communicate creates an intense agony on the individual to choose a language of instruction and preference. This judgment of which language of instruction should be accepted is often guided by the underlying power attributed to the language of power and authority of the specific context. With this in mind, the testing of such individuals in a language other than their own creates a linguistic dilemma and can ultimately affect their cognitive ability to grasp language sufficiently. Moreover, individuals should be offered the choice of being tested in their language and not merely submitting to the language of power.

For this reason, intense procedures are followed in order to provide instruments suitable to test individuals in multiple languages, thus the existence of different language versions arise. Furthermore, the testing of equivalence is crucial (previously argued in chapter 2) and affects how the individual scores. To illustrate this, the uses of two different case studies are used in order to explore this issue. These case studies also serve to demonstrate the great divide that exists in terms of expertise in language testing in South Africa. This is noticed when one observes the difference in the outcomes of these two equivalence studies. It is also worth noting that one case study is international (The Israel PET test) and the other a national one (a Reading Comprehension test), also previously discussed in chapter 2 in section 2.6.2. The significance of these case studies is evident through the analysis of this paper, as the implications of equivalence and the uses of tests are once again emphasized. Thus the interrelated nature of equivalence as a statistical means is stressed when exploring the different aspects of tests. In addition to this, the case study links to the issue of fairness in testing (previously argued in chapter 2), allowing for a more holistic view of language testing.

3.9.1 The Israeli PET (Psychometric Entrance Test)

The Israeli National Institute for testing and evaluation (NITE) was very concerned about testing in multiple languages and thus sought to evaluate construct equivalence to ensure that fair and valid selections would be made. Thus they were fully conscious of the implications that would result when adapting the PET test into multiple language versions. The language of instruction at the Israeli universities is Hebrew and thus individuals seeking to enter university would have to complete this test before being selected. The adaptation of the test was guided by the appropriate procedures and thereafter the analysis on the test was to be conducted. It is rather interesting to note that in their adaptation of the test, the verbal reasoning subtest was described as problematic in terms of translating the meaning of verbal items (Beller, Gafni & Hanani, 1999). This is a similar finding in the study by Koch (2007; 2009) in which the translating and adapting of the verbal analogies subscale was problematic.

The evaluation of construct equivalence sought to establish that identical constructs were being measured in the language versions of the PET and thus exploratory factor analysis, multidimensional scaling (MDS) and confirmatory factor analysis (CFA) was selected as the methods by which to confirm this. Due to the rigorous adaptation of the verbal reasoning subtest of the Israeli PET, the analysis was limited to only this subtest to ensure proper translation and adaptation occurred. The Hebrew and Russian language versions of the verbal reasoning subtest of the Israeli PET was examined and content areas such as analogies, logic, reading comprehension and sentence completion was scrutinized. In addition to this, a total of 41 items were selected for the analysis. The findings thus revealed that similar constructs were measured across the language versions of the verbal reasoning subtest of the Israeli PET. Moreover, this study serves to demonstrate that when the translation and adaptation process was successful, the language versions for the subtest will be equivalent (Beller, Gafni & Hanani, 1999). In addition to this, multiple methods of evaluating construct equivalence serves to cross-validate results, leading to more informed information about the construct of interest.

3.9.2 The ACCUPLACER Reading Comprehension Companion Test

The ACCUPLACER Reading Comprehension Companion Test is only available in English and tests individuals' fundamental language ability. This test is used cross-culturally at the NMMU (The Nelson Mandela Metropolitan University) in order to make executive decisions on admission and placement into the university. This test

was then statistically explored to establish whether the same constructs were being measured across the languages since it was used cross-culturally. Additionally, the motivation to establish that this language test was not disadvantaging any groups of individuals was essential to this analysis (Koch, 2008).
The test was administered to a sample of three language groups such as English, Afrikaans and isiXhosa. This allowed the researcher to compare the test scores across the three language groups, thereby establishing whether these language groups differ in performance on the test. The method by which construct equivalence was evaluated was that of WMDS (Weighted Multidimensional Scaling). The findings of this analysis revealed that the test displayed construct bias, thereby stressing that different constructs were being measured in the different language groups. This allowed the researcher to claim that this test could not be used to make comparisons across groups and much less place the test scores observed on a common scale. In addition to this, item bias was observed and with the exclusion of biased items, the results remained rather bleak, still providing evidence of construct bias (Koch, 2008).

The conclusion that this test is highly biased in terms of the constructs measured across languages was therefore critically discussed. An important point worth noting in this testing of equivalence of a reading comprehension test for admission and placement into university is the awareness of various tests being administered cross-culturally without the consideration of equivalent constructs in the different language groups. Furthermore, Koch emphasizes the idea that one should consciously “problematise exclusion based on language” (Koch, 2008, 22). This critical engagement with testing in two or more languages and the equivalence of different language versions is thus stressed throughout this paper.

3.10 Conclusion

This chapter thus explored the broad nature of language testing and the different aspects pertaining to testing languages. The backdrop of cross-linguistic testing is significant, as this is the umbrella for the present study and understanding the nature of language testing in this context is vital. This allows one to fully comprehend that the need for language testing, specifically the testing of an African language is imperative. The infancy of the present larger study can be regarded as a catalyst of change within the realm of language testing, as it is seeking to reduce the immense gap that exists. With the completion of the present study, more baseline information can be obtained and so, the continuation for providing relevant language tests for African countries such as South Africa proceeds.

CHAPTER FOUR

METHODOLOGY

4.1 Introduction

The present study is a sub-study of the adaptation of the isiXhosa version of the WMLS. Accordingly, the researcher seeks to confirm previous research (Koch, 2009) by cross-validating the results and examining the previously mentioned hypotheses. The design and method of the research correspond to psychometric test theory and research in this area, as the researcher is interested in the equivalence of the two versions of the WMLS. As a result, the design reveals the evaluation of construct bias, in view of a specific focus on the structural equivalence of the two language versions.

4.2 Research Design

The researcher has done Secondary Data Analysis (SDA), which can be described as the re-analysis of data in order to answer an original research question with improved statistical techniques or answering innovative questions with the use of old data (Glass, 1976). The researcher answers an original question posed by the researchers of the main study, but improves a methodological limitation of the previous study (Koch ,2009), namely by purposively sampling from the original data set and matching the language groups on their total scores on the verbal analogies test of the WMLS, using frequency distribution matching.

A feature of testing in two languages is the testing of language ability across two language groups such as verbal reasoning. Verbal ability testing can be understood as allowing the researcher to access the individual's potential to develop skills and their performance in the assigned test. A reasonable manner of assessing whether learners are functioning well with the language of a test is to select them based on their performance in the test. Selecting based on ability is very different to that of selection by attainment, which would require individuals who scored particularly high, as it determines how much knowledge they have acquired from the work they were exposed to (www.psychometric-success.com).

The subtest of the WMLS, Verbal Analogies, can be classified as a test of verbal reasoning ability. In light of this, the choice of the researcher to match learners based on their ability in the verbal analogies section of the WMLS can be substantiated with the claim that if one is to select scores based on their ability, one is able to confidently claim that the differences (using equivalence techniques such as EFA) found across the two groups on the other scales is due to differences in the construct, and not ability. This is why the researcher selected this method of matching, and the verbal reasoning test as a measure of ability for both subscales. Additionally, it would facilitate the means by which the researcher is to locate the equivalence of the language tests across the two languages (www.psychometric-success.com).

The design makes use of the monolingual matched two-group design with English and isiXhosa first language learners in the two language groups. This design allows the researcher to compare the performance of the English first language-speaking learners on the English version with the performance of isiXhosa first language learners on the isiXhosa version.

4.3 Participants

4.3.1 The sample

The original sample from which the participants for the present study were drawn is represented in the two tables in the appendix as Addendum C. There are slightly more female than male learners from grade 6 and 7 in the English and isiXhosa language groups. The IsiXhosa sample is drawn from both rural and urban groups from areas in the Eastern Cape. The table below represents the total number of learners for grades 6 and 7 according to their genders respectively for both the English and IsiXhosa language groups on the matched sample. The total sample size is 150 which is an acceptable sample size with which to conduct factor analysis.

Table 1

The table representing the sample for the current study

Language Group					
Categories		**English**		**isiXhosa**	
		N	**%**	**N**	**%**
Grade	**6**	68	45 %	48	32 %
	7	82	55 %	102	68 %
	Total	150	100 %	150	100 %
Gender	**Female**	93	62 %	93	62 %
	Male	57	38 %	57	38 %
	Total	150	100 %	150	100 %

The main study used purposive sampling[1], because it allowed the researcher to select homogenous participants, in terms of an equal number in terms of gender from the various types of schools and the educational backgrounds[2] of the two language groups. Attempts to control for socio-economic status (confounding variable) was made in that the Eastern Cape Education Department assisted the main researcher in selecting English first language learners from low middle socio-economic status schools and isiXhosa first language learners from well-functioning rural and urban schools. It is important to note that this study does not intend to generalize the results at this stage, thus the study does not attend to issues of external validity in the sampling. Due to the researcher using SDA, and the matching procedure that is followed, the participants of this study forms a subset of the original data set.

4.3.2 The differences between the two subtests across English and isiXhosa language versions using the Hotelling T^2

Since the data was matched on the VA subscale, based on the participants total ability scores, it was necessary to conduct a Hotellings T^2 test for the two subtests, namely: Verbal Analogies (VA) and Letter-Word Identification (LWI) in order to evaluate the group differences. Based on this, one expects that the VA subscale should have similar or the same means scores across both language versions. Furthermore, it is imperative to assess the differences on the remaining LWI subscale in both the English and isiXhosa language versions.

[1] The reasoning behind this sampling method can also be referred to as the sampling criteria.

[2] Controlling for educational backgrounds was done by selecting English and IsiXhosa learners from certain schools in order to maintain the validity for the learners' different educational levels of their primary language as well as the difference in the teaching of their primary languages. No isiXhosa learners in the ex-model C-schools.

To follow is a tabular representation of the Hotelling T^2 values for two subscales across the two language versions.

Table 2

The Hotelling T^2 for the VA and LWI subscale across both language versions

T (case wise MD) = 295, 000 F (4, 000) = 61, 123 p< 0, 000							
		Post-hoc F values	**df**	**p-values**	**Language Versions**	**Mean scores**	**Standard Deviations**
Subscale	**Verbal Analogies**	0.00	1	0.99	English	12.45	3.57
					isiXhosa	12.46	3.55
	Letter-Word Identification	35.33	1	0.00	English	45.77	6.04
					isiXhosa	50.09	6.56

In the above table, one observes that the Hotelling T^2 statistic is significant and thus there is a difference in the overall score across the language groups. The VA subscale is not significant; therefore there are no differences in the language groups. It is important to note that one expects that there will be no significant difference between the languages on the VA subscale, as the data was matched on this scale. The LWR subscale however is significant indicating that there is a difference between the two language groups in this test. These differences may be explained by the fact that these scales will be easier for the isiXhosa group because of the language orthography of this language. Orthography can be understood as the isiXhosa sound system being directly and regularly related to the way in which the isiXhosa language is written while this is not the case with the English written system. However, it remains important to ascertain that these differences occur because of real differences on the construct, or because of construct irrelevant factors.

In the VA test, the assumption that the means and standard deviations should be quite similar is accepted, as the data has been matched on this test. The means clearly indicate the similarity between the two language groups. In addition to this, the standard deviations serve to further reinforce this similarity across the two language groups for the test. One can speculate based on this that this test would potentially be equivalent across groups. In the LWR test, there is a clear difference between the languages indicated by the large difference in means and standard deviations. It should be noted that the isiXhosa group has a significantly higher mean than that of the English group, with a slightly higher standard deviation in the isiXhosa group. The line graph below represents the mean scores for the two subscales across the English and isiXhosa language versions.

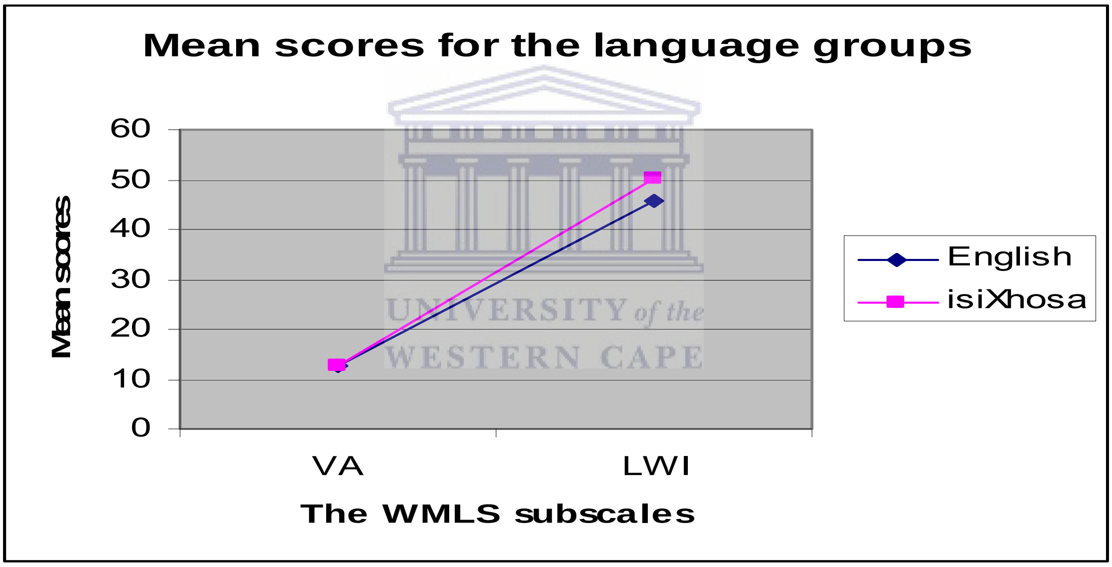

Figure 1: A graphical representation of the mean scores of the VA and LWI subscales

for

both language versions

4.4 Data Collection Procedure

The collection of the original data set was done in 2005 and 2006 by research assistants at the NMMU under the supervision of the main researcher of the main study. The data was collected in both rural and urban areas in the Eastern Cape.

Attempts were made to control for confounding factors such as socio-economic status and print exposure, which might have affected performance and this was done by ensuring that these factors were as equal as possible across groups (Koch, 2006). The researcher of the current study will be using SDA and thus the procedure being followed will be on par with the objectives stated.

The study has obtained ethical clearance by the NMMU according to their requirements in 2005 and 2006, where the larger study was based at the time of data collection. The larger project also received permission for the research from the Eastern Cape Education Department, the principals of the schools were contacted, and the project was thoroughly explained to them. On their agreement, parents received consent forms to sign on behalf of their children, in order for their children to partake in the study and thus only children who had permission to partake in the study were included. According to the requirements of the NMMU at that stage, children under the age of 16 do not have to sign their own consent forms. For this study, the researcher matched the two language groups on their total scores for the verbal analogies scale. The data was statistically analyzed per subtest using the SPSS (Statistical Program for the Social Sciences) package and CEFA (Comprehensive Exploratory Factor Analysis), because of its suitability to succinctly conduct exploratory factor analysis.

4.5 Data Collection Instrument

The measurement tool that was used was the adapted English version and the isiXhosa version of WMLS. The WMLS is an individually administered[3] test of

[3] It requires 50 minutes to administer.

academic language proficiency and is commonly used in the United States of America (USA) on different age groups[4] and linguistic and cultural backgrounds. It was selected for use in the ABLE study, because the content of each subset represents vital skills necessary for language proficiency for a diverse population. This allows one to assess the cognitive academic language proficiency (CALP) of individuals (Woodcock & Muñoz-Sandoval, 2005). This concept has been discussed in a previous section. A total measure of language competence is produced.

The WMLS consists of four sub-tests namely; Picture Vocabulary, Verbal Analogies, Letter-Word Identification and Dictation. The test requirements as well as what each test measures are tabulated in the Appendix as Addendum D along with the scoring of the WMLS. The items of the WMLS are not made available in an appendix of this proposal as it is a commercially purchased test; items are therefore confidential (Woodcock & Muñoz-Sandoval, 2005). The current study will only focus on two of these subtests, namely: Verbal Analogies and Letter-Word Identification.

An isiXhosa WMLS was produced for the South African context with permission from the publishers (Refer to Appendix, Addendum A). The layout and format (thus all the characteristics) was retained and a committee of language specialists, translators and educators assisted in the adaptation of the instrument (Koch, 2006).

4.5.1 Psychometric Properties of the WMLS

The reliability of the original USA version of the WMLS was established by using split-half reliability in the USA sample and used odd and even raw scores. The

[4] It covers a broad range of development from 2 years to adulthood.

reliability coefficient was calculated by using the Spearman-Brown formula and the median reliabilities revealed a range from 0, 80 to 0, 93 for the scales and 0, 88 to 0, 96 for the clusters. This makes the WMLS a very reliable test for the USA context. In addition to this, the validity of the WMLS, the USA version, was evaluated on content, concurrent, as well as construct validity. The instrument displayed acceptable levels of validity for the USA context (Woodcock & Muñoz-Sandoval, 2005).

The South African (English and isiXhosa) versions of the WMLS have no psychometric properties available, because this psychometric information is currently being collected. This study forms part of this process.

4.6 Ethical Considerations

The learners participating in the study received permission from their parents and the teachers and school gave permission for the study to take place. The teachers who were knowledgeable about the project (Koch, 2007) informed the parents. The letter of informed consent was available in English and isiXhosa (Refer to the Appendix, Addendum B). The previous section on procedure dealt with the ethical considerations of the larger project. Ethical clearance has been obtained for the research to be conducted as a subset of the ABLE project (Refer to Appendix, Addendum C).

Permission for the study to be conducted has been granted from the principal investigator, Elize Koch, and the analysis of the SDA was done with her assistance and guidance and the information was safeguarded. The protection of data is a serious consideration as this data includes the participants' private information. As a result, the use of statistics allows the information to be both private and anonymous.

4.7 Data Analysis

The researcher has used SDA to perform various statistical tests on SPSS in order to test each of the three objectives, and realize the main aim of the proposed study, which are as follows:

4.7.1 Objective 1: To compare the Verbal Analogies and Letter-Word Identification scales of the two language versions of the WMLS on reliability using sample groups matched on their total scores on the subscales of the test.

The Cronbach's Alpha was calculated for each group per subtest and was be compared across the language versions in order to test differences in the reliability of the various subtests between the English and isiXhosa language versions of the WMLS. This was done using the following statistic, namely 1-alpha$_1$ / 1-alpha$_{2.}$ Alpha$_1$ has been assigned to the English version of the WMLS for both of subtests and alpha$_2$ has been assigned to the isiXhosa version of the WMLS for both subtests. The equivalence of the test is ensured if there is no significant difference observed between the two language versions of the WMLS. This significant difference is expressed as the statistic which follows an F-distribution with $N_1 - 1$ and $N_2 - 1$ degrees of freedom (Van der Vijver & Rothmann, 2004). The critical F value for this study was 1, 66.

It is important to note that the value attached to each test can be understood as the extent to which the test is able to produce reliable scores. This means that there is consistency of test scores in the test. Essentially, one seeks to produce high reliability scores in order to ascertain that the items are consistent. This means that the learners, who answered certain items correctly, are more likely to answer other related items correctly. This differs greatly from low reliability, which indicates construct irrelevant items, and thus problematic items are present

The Cronbach formula is presented as follows:

$$a = \frac{N.\acute{r}}{1 + (N - 1) . \acute{r}}$$

(Van der Vijver & Rothmann, 2004) (1)

4.7.2 Objective 2: To compare the Verbal Analogies and Letter-Word Identification scales of the two language versions of the WMLS on mean item characteristics using sample groups matched on their total scores on the subscales of the test.

The comparison of the item characteristics between the two groups on the language versions of the tests per subscale is descriptive and involved assessing the mean item difficulty as well as the mean item discrimination per subscale per language group.

The mean item characteristics will be presented in the tables in the consecutive chapter in terms of difficulty as well as their level of discrimination. In terms of item difficulty, one can distinguish between easy, moderate and hard items and compare this along the two language versions to see how these items present themselves. By doing so, one is able to see if there are any discrepancies in the item across the different language versions of the tests and thereby prompting further item investigation by means of DIF (Differential Item Functioning). Item bias could then become an issue, indicating that the item is discriminating against one of the language groups. It is then crucial that one explores the items critically and identifies problematic items in order to improve the test items used in both language versions (Foxcroft & Roodt, 2006). The WMLS under which the two subtests fall, is designed for the age groups 3 to 99, which means that the items used are for all age groups (Woodcock & Muñoz-Sandoval, 2005). Since the age of the two sample groups in this study range from 14 – 16 (Koch,2009), one would expect some very easy items in both subscales in both language versions, as well as some items that were too difficult for these age groups to answer. A number of items were expected to present with no variances, with the test takers all having scored either correct (on the easy items) or incorrect, on the difficult items.

Following this, one must assess the discrimination level of the items, in which they are good, fair or poor. Item discrimination should preferably be good or fair (≥0.3) in order to distinguish between high and low ability learners. When items are poor discriminators, they are not able to distinguish between the groups and thus one cannot differentiate between learners. It is important to note that one intends items to discriminate between high and low performing test takers, thereby allowing one to notice these differences on performance (Foxcroft & Roodt, 2006).

The sample groups were matched on ability in order to indicate that differences in item characteristics across the language versions are not linked to ability, but rather to the item itself.

Essentially, one will expect that younger learners would find more items difficult and answer moderately. Thus, one must be aware of this when assessing the general trend of the items. The appropriate mean values in terms of the acceptable p-values are means of 0, 5 in order to avoid the ceiling and floor effect on the total test scores. Also, a mean of 0, 5 is a reasonable difficulty level for p-values. The acceptable mean item discrimination values are correlations of 0, 3 and above, but below this is unacceptable (Foxcroft & Roodt, 2006; Walsh & Beltz, 2001).

4.7.3 Objective 3: To evaluate the structural equivalence of the Verbal Analogies and Letter-Word Identification scales two language versions of the WMLS using sample groups matched on their total scores on the subscales of the test.

The method used for analyzing the structural equivalence on the matched groups is that of exploratory factor analysis, in order to assess the similarity of the structure of the data per subscale across the two language versions. This assisted the researcher in exploring whether the same construct is found in the same form in both language versions. Exploratory factor analysis can provide evidence of factorial invariance (thus structural equivalence), indicating that the factor loadings of the items on the underlying factor are comparable across different cultural groups (Welkenhuysen-Gybels & Van der Vijver, 2001). The concurrence between the factor loadings of items from the two different groups can be expressed via congruence indices.

The relative coefficient was that of the Tuckers phi. Tucker’s phi allows one to measure the identity of two factors.

The Tucker's phi formula can be presented as follows:

$$p_{xy} = \frac{\sum x_i y_i}{\sum x^2_i \sum y^2_i} \qquad (2)$$

(Welkenhuysen-Gybels & Van der Vijver, 2001; Pienaar & Van Wyk, 2006).

Factor loadings inform one of the relative contribution that a variable (in this study, items) makes to a factor and can be either or both a correlation coefficient and regression coefficient depending on the analysis being done. In the present study, the researcher will be making use of the correlation coefficient (Field, 2005). Structural equivalence is achieved when the same underlying dimension is noticed and presents itself as meaningful clusters of variables within the data set across the different language versions of the test (Field, 2005).

Moreover, a prerequisite for structural equivalence is that of the similarity of factor loading for each item as this ensures a sufficient construct representation of the same factors in the two language groups. If construct in-equivalence is found, then it means that construct bias is present and eliminates the chance of score comparability, because it prevents any cross-cultural or cross-linguistic (in the case of this study) comparison (Hambleton, Merenda & Spielberger, 2005).

4.7.3.1 The Evaluation of the data

Before the evaluation of the data began, descriptive statistics (e.g. means and standard deviations) and inferential statistics were done in order to analyze the data and get a sense of the data. The descriptive statistics was initially done in order to determine the state of the data and to give the researcher an indication of how well the matching would work. Means, standard deviations, standard errors, medians, modes, histograms, t-tests and frequency statistics were computed for each of the language groups in the different subtests of the WMLS.

The Letter-Word Identification test indicated problematic distributions in that the means and ranges across the two language groups were not very similar. The Verbal Analogies test proved to distribute the data normally and in accordance with literature pertaining to verbal ability testing, the data was matched using the scores of this test. As argued in the literature, it serves as a sufficient and less problematic manner of matching sample groups across the two languages (Berkmen, 2002).

After the descriptive statistics were done and the groups were matched on verbal ability, the researcher proceeded to analyze the data using exploratory factor analysis in a statistical package, CEFA (Comprehensive Exploratory Factor Analysis) – Version 3.02, which has been recommended as an ample manner for performing exploratory factor analysis (Browne, Cudeck,
Tateneni, & Mels, 1998; Kano & Harada, 2000). It is also important to note that this study made use of tetrachoric correlations as the item level data is of a dichotomous nature (Zumbo, 2003).

The standard sample size for factor analysis is not less than 50 observations and the preferred size of the sample should be 100 or more. This means that there must be a minimum of five observations per variable studied (Hair, Black, Babin & Anderson, 2009). The data used for the present study adhere to these guidelines as a means of ensuring that proper factor analysis takes place and minimizes the presence of chance findings in the analysis.

The assumptions of factor analysis are very important and must be met in order for factor analysis to be deemed appropriate. The first assumption, on which factor analysis is build, is a conceptual one, which refers to the concept being studied. The assumption is met in that previous literature on the subtests has indicated that there is a definite underlying structure in the variables selected for analysis (Hair, Black, Babin & Anderson, 2009).

Another assumption of factor analysis, which has been met, is that of multi-collinearity, because the researcher seeks to identify interrelated sets of variables. In the analysis, one can observe that the variables are efficiently inter-correlated to produce sufficient factors (Hair, Black, Babin & Anderson, 2009).

4.7.3.2 Performing the factor analysis

Critical to factor analysis is the method of extraction. In this case, the method chosen was Common factor analysis, because the researcher sought to identify the underlying factors on dimensions that reflect what variables have in common. This means that the shared variance was considered in defining the structure of the variables and informing one of the factor loadings per factor (Pienaar & Van Wyk, 2006). Thus, one refers to communalities in order to assess the common variance. The Scree-plot, which is a graph that presents the eigenvalues of all the factors, assisted the researcher in deciding how many factors to retain. The amount of factors, which should be retained, is presented as a curved line and does not include the factors under the flattened line (Field, 2000 & Newcastle University, 2007).

The rotation method chosen for the study was that of the oblique rotation, because this method of rotation produces correlated factors, which is essentially what the researcher intends to observe. Additional to this, the literature on language learning suggests that one can expect a relationship between the factors and thus oblique rotation is ideal (Cummins, 1978). The importance of the rotation method is that it is able to minimize the number of factors, on which the variables being studied have high factor loadings. Essentially, rotation does not change the output, but assists in making the interpretation of factor loadings on items much simpler.

Another consideration is that of the choice of the matrix to use to interpret the factor analysis. The appropriate matrix chosen for this study was that of the pattern matrix (Hair, Black, Babin & Anderson, 2009). This matrix makes interpretation easier and simple and allows one to assess the number of items loading on each factor as well as the values of the loadings. One can then evaluate which items contribute the most to the factor(s) and the contribution of the factor(s) to the construct (that is measured). In addition to this, one can assess whether items load only on one factor and not on two, because if an item loads on two factors, it indicates that the item is measuring two constructs. Such items should be flagged and investigated (Field, 2000; Newcastle University, 2007).

After the rotation was done, the factorial conformity was estimated using the Tucker's coefficient of agreement (Tucker's *phi*) (Pienaar & Van Wyk, 2006). Values higher than 0.95, are regarded as confirmation of identical factors, while values lower than 0.85 indicate non-negligible incongruities. This is the exact index by which factorial similarity is observed and hence structural equivalence will be accepted. If structural equivalence is not accepted, then an analysis of item bias should take place to examine the inappropriate items (Pienaar & Van Wyk, 2006).

The factor analyses, per subscale, were performed for the English version of the test first to find a stable factor structure for that version. Items with no variance (very easy or very difficult) were left out from the analyses from the beginning. After several rounds of analyses, during which a number of items were excluded on the basis of their factor loadings, the final solutions (per subscale) were accepted. The test is an USA developed test and the most appropriate items needed to be selected for the South African context. The items that were retained for the English version of the test were then entered into the analyses for the IsiXhosa version of the test, with the researcher specifying the same number of factors obtained in the final solutions of the English version. The factor loadings of the two language versions were then compared using the Tucker's Phi.

Additional to factor analysis is the application of a scatter plot in order to cross-validate the findings of the factor analysis. The scatter plot is comprised of the factor pattern matrices of the different groups with the presence of an identity line in order to assess how perfectly the items align on the straight line. A perfect congruence is when the items perfectly align on the identity line and is thus dependant on the similarity between the groups. When factor pattern matrices of the two groups are proportionally similar, then the patterns of high and low loadings are similar. This then serves to indicate that perfect congruence is when the coefficients of congruence are not sensitive to difference in the factor pattern coefficients (De Bruin, 2009).

According to Pedhazur and Schmelkin (1991), the naming of a factor should be taken quite seriously and done in a particular manner. The naming should therefore seek to embody the nature and content of the construct being studied. Moreover, a consideration of the theoretical underpinning of items should reinforce the meaning of the factor it is comprised of. The factor name must therefore represent the construct being measured. This study attempted to name the items as well as suggest a possible name for the factor being measured within the two subscales. The naming of the factors was done by inspecting the items to assess what the items measure. Thereafter, the items which make up each factor were explored as a collective and a preliminarily names were attached to the factors. The results of the named items and factors will be presented in the results chapter.

4.7.3.3 The reporting of factor analysis

To assist in the interpretation of the factor analysis results, the results will be reported as follows per subscale:

1. The number of factor analyses that were conducted to reach the final solution for the English version and the main findings will be summarized.
2. The results of the pattern matrix of the final solution for the English group, and the results for the IsiXhosa group on the same solution will be presented and discussed.
3. The results of the Tucker's Phi and reliability of the factors for both groups will reported and discussed.

4. A scatter plot of the factor pattern coefficients for the subscales will be done in order to compare and cross-validate the results of the Tucker's Phi.
5. Items with no variances in the two groups will also be presented to compare the two language versions on these items
6. The named items as well as the factor names attached to each subscale for each language version will be presented.

These results will be presented for the two subscales separately.

4.8 Conclusion

This chapter sought to explain the methodology used in the present study and to explore the procedures that were followed in order to comply with the ethics. This chapter also integrated the objectives and the types of analyses used for each to be met. The relevance and importance of this chapter lies in the fact that the implementation of the previously mentioned aims and objectives are being realized and with the completion of this chapter, the successive chapter illustrates the results and addresses the findings of these analyses.

CHAPTER FIVE

RESULTS

5.1 Introduction

The previous chapter outlined the methodology of the current study and the procedures that were followed in order to achieve the desired aims and objectives of the study, namely to evaluate the structural equivalence of the two language versions of the WMLS on matched sample groups. This chapter will then serve to illustrate the findings by locating the results in tabular form and thereafter interpreting the analyzed data. This interpretation of the data then lends itself to the discussion of the implications of these results in the subsequent chapter. In relation to the interpretation of the data, the understanding of what this means for each subscales as well as the overall test, is vital in furthering our understanding of the adaptation of the WMLS test.

It is important that one assesses whether the previous results of structural equivalence of the relevant two subtests of the two language versions of the WMLS, namely the Verbal Analogies and Letter Word Identification can be confirmed with, and has improved with matched groups (Koch, 2009). The focus on these two subscales will allow the researcher to explore them in depth, thereby making appropriate inferences based on the results obtained.

5.2 The reliability of the two subtests

As previously explained, the reliability of the subtests was done by performing running an analysis and assessing the Cronbach's Alpha.

Fundamentally, when interpreting the Cronbach alpha value, one must refer to the estimate of the internal consistency of the test for each language group. It should also be noted that tests which have normally distributed scores, are likely to have higher Cronbach alpha reliability estimates than tests with positively or negatively skewed distributions. This means that one must consider the distribution of scores in relation to the estimate of the alpha as part of one's interpretation (Field, 2005).

Below is a tabular representation is of the Cronbach alpha values for the two subscales across the two language versions.

Table 3

The Cronbach Alpha values for the two subscales across the two languages

Cronbach's Alpha – Internal Consistency values		
Tests	**English**	**isiXhosa**
VA	0.78	0.75
LWR	0.91	0.92

The Cronbach's alpha value of 0.78 for the English version indicates that the VA subscale is sufficiently reliable for research purposes in this group. This value is however smaller than the prescriptive norm for high stakes tests (the value is less than 0. 90) reference. The value of 0.75 for the isiXhosa version of the VA subscale is regarded as an acceptable value, while the English version is slightly higher than the isiXhosa version Thus when assessing these two language versions about reliability; one is able to recognize that although they have similarly good reliability values but that both fail to be sufficient for high stakes situations, but is sufficient for research purposes.

The Cronbach's alpha value of 0.91 for internal consistency is a high value and indicates that the English version of the LWI subscale is very reliable. The value of 0.92 for the isiXhosa version is higher than that of the English version, indicating that the isiXhosa version has a slightly higher internal consistency. These high values for the LWI subscale across the two languages indicate that this subscale in both language versions can be used in high stakes situations. Thus, one can conclude that both versions of the LWI test have high levels of internal consistency. Moreover, one can deduce that the majority of the items in the two language versions of the LWI subscale will be measuring the same construct, pointing more firmly in the direction of equivalence.

Below, table 4 serves to represent the respective values for the VA and LWI subscales on their equal reliability values as well as the compared critical value.

Table 4

The test of equal reliability per subscales

In table 4, the values representing the equal reliability tests are presented for both subtests. The results indicate that there are no significant differences in both the VA and LWI subscales across the language versions in terms of their reliability coefficients. Thus the results tend to point in the direction of construct equivalence.

5.3 The Mean Item Characteristics of the two subtests

The mean item characteristics of the two subtests will be done separately under the respective subscale headings.

5.3.1 The Verbal Analogies subscale

In this section, the tabular representation of the mean item difficulty and mean item discrimination values for this subscale will be presented.

Table 5

The mean item characteristics of the VA subscale

Test of equal reliability		
Subtests	**F-ratio**	**Critical value (0. 01)**
VA	0. 88	1. 66
LWR	1. 13	1. 66

Test	Language Version	Mean Item Difficulty		Mean Item Discrimination
		Mean	Standard Deviation	Mean
VA	English	0.43	0.32	0.28
	IsiXhosa	0.39	0.33	0.23

When exploring the contents of the above table, one is able to compare the mean scores of the item difficulty with that of the discrimination values. In the English version, the value of 0.43 is a relatively reasonable difficulty value, allowing one to speculate that there could be a skewed distribution, resulting in a ceiling effect. Based on the item discrimination value of 0.28, the items do not discriminate very well between the high and low- achievers, yet there is sufficient discrimination as 0. 3 is an acceptable mean discrimination value (Foxcroft & Roodt, 2006). It is however important to consider the age range for which the test was developed, which ranges from 3 to 99. This therefore makes the mean values for the items discrimination and p-values for the sample of grade 6 and 7s very acceptable.

The isiXhosa version also has a similarly reasonable difficulty level, indicating that the items are generally manageable, while there are persistent difficult items present. The lower discrimination power of 0.23 indicate that items on average tend to not discriminate well between the ability groups in this language group.

When comparing the mean difficulty and discrimination of the English version to that of the isiXhosa version, one observes that they are quite similar, albeit with lower item discrimination values in the IsiXhosa version.

5.3.2 The Letter-Word Identification subscale

As previously mentioned, the mean item difficulty and mean item discrimination values for this subscale will presented in a tabular format (as observed below).

Table 6

The mean item characteristics of the LWI subscale

Test	Language Version	Mean Item Difficulty		Mean Item Discrimination
		Mean	**Standard Deviation**	**Mean**
LWI	**English**	0.68	0.33	0.43
	IsiXhosa	0.89	0.27	0.38

In the English version, the mean item difficulty score of 0. 68 are relatively high, indicating that the majority of the items can be classified as easy items, while others were difficult. The corresponding item discrimination of 0.43 is reasonable in that one is able to distinguish between high and low achievers. Overall, the English version pertains to be a relatively good subscale in that it not only has a reasonable amount of easy and difficult items but also for its average discriminatory power. Based on this, one could speculate that the distributions of scores are to some extent normal, eliminating possible ceiling or floor effects.

The isiXhosa group has a very high mean difficulty value indicating that most of the items are easy and were answered correctly. This leaves a very small group of items to be difficult and incorrectly answered. The mean discrimination value of 0.38 is indicative of reasonable discriminatory power whereby one is able to discriminate between groups of participants in terms of their performance on the items. Based on this, one can speculate that the scores are distributed in a skewed manner, in that it creates a ceiling effect due to the accumulation of easy items.

When comparing the language versions in terms of their difficulty and discrimination power, there is a similarity observed between the values. The isiXhosa language group however experiences this subscale much easier than the English group due to the language orthography of the isiXhosa language. One can however speculate that despite this slight advantage, the performance of the English group is comparable and this could be an indication that the same construct is being measured across the language versions.

5.4 The factor analysis of the two subtests

As previously explained in the methodology chapter, the factor analysis of the two subscales will be conducted and interpreted by inferring meaning from the appropriate tables and linking this with the information on group differences and previous research on these scales.

5.4.1 Factor analysis results for the VA subscale

When running the initial factor analysis for the two language groups, the researcher used descriptive statistics (see section 4.7.3.1 of chapter 4) in order to remove the items which displayed no variances. These items will be listed later in this chapter, in section 5.4.1.5.

5.4.1.1 Reporting the steps in the factor analysis

As a first step, a two factor solution was selected. This allowed the researcher to assess via the pattern matrix whether some items were loading on two or more factors or whether they were not loading at all. The cut off score used for determining factor loading was 0, 40 (also specified by Hair et. al, 2006). The subsequent items that were removed during the search for an acceptable solution on the English version (with the English first language group) were VA 1, VA 2, VA 5 and VA 6. This left the researcher with the items ranging from VA 3 to VA 29, adding up to 24 items. In a subsequent running of a two factor solution additional problematic items such as VA 8, VA 15 and VA 23 were also excluded from the analysis.

The final analysis consisted of a two-factor solution and the total amount of items retained was 21. This solution produced a stable structure for the final factor analysis. The isiXhosa version of the VA was conducted on the same items specifying the same two-factor solution.

5.4.1.2 The results of the pattern matrix of the final solution for the two language groups

As a means of representing the results of the pattern matrix for the VA subscale, a tabular format will be used to display the results. Below are the results for the English version of VA subscale which display the two factor loadings observed.

Table 7

The two factor solution for the English version of the VA subscale

English version of VA subscale		
Items	**Factor 1**	**Factor 2**
3	-0.09	**0.62**
4	-0.12	**0.80**
7	-0.02	**0.65**
9	0.07	**0.34**
10	0.05	**0.63**
11	0.00	**0.71**
12	0.18	**0.42**
13	**0.61**	0.16
14	**0.53**	0.17
16	**0.52**	0.05
17	**0.58**	0.34

18	**0.32**	0.05
19	**0.71**	-0.05
20	**0.94**	-0.17
21	**0.73**	-0.17
22	**0.61**	0.22
24	**0.86**	-0.13
25	**0.99**	-0.17
26	**0.88**	-0.21
28	**0.80**	0.14
29	**0.69**	0.19

The results obtained and illustrated in the table above, show the different items loading on the each factor. According to Field (2005) and Hair et al., (2009), an item must have a minimum loading of 0. 40 based on the sample size of 150. In addition to this, the first factor is identified by its high congruence of factor loadings and the large amount of items loading on the factor. With this in mind, a high congruence is evident in the first factor while a medium congruence is witnessed in the second factor. Moreover, the item loadings of the English version perfectly divided into two factors.

The table below represents the factor solution for the isiXhosa version of the VA subscale, allowing one to observe the two factor loadings.

Table 8

The two factor solution for the isiXhosa version of the VA subscale

IsiXhosa version of the VA subscale		
Items	**Factor 1**	**Factor 2**
3	-0.29	**0.47**
4	-0.35	**0.62**
7	-0.07	0.21
9	0.23	0.31
10	0.21	**0.60**
11	0.12	**0.50**
12	0.26	0.08
13	0.00	**0.61**
14	0.18	**0.37**
16	0.32	**0.37**
17	**0.50**	-0.10
18	0.28	0.28
19	**0.83**	0.16
20	**0.82**	0.16
21	**0.63**	0.14
22	**0.45**	0.25
24	**0.85**	-0.05
25	**0.81**	-0.26
26	**0.94**	0.06
28	**0.83**	-0.17
29	**0.84**	-0.22

In table 8, the specified two factor solution is witnessed yet the observations of problematic items are present. The items loading on the respective factors were slightly different to that of the English version. The presence of DIF items such as 7 and 18 (previously identified by Koch, 2007; 2009) are recognized as items not loading on any factor and thus possibly interfering with the congruence of factors. The other items that pose problems in terms of their inability to load on a factor (because they are < 0. 4 on both factors) are items such as VA 9 and VA 12, while items 13 – 16 loaded on the second factor in this version of test, while they loaded on the first factor in the English version of the test.

This can then be interpreted to mean that these items must be explored further as they are not measuring a similar construct in the two language versions. In addition to this, these two factors can be regarded as certainty factors of differential difficulty in that factor 1 identifies difficult items while factor 2 identifies easy items. Based on this, one can conclude that these factors are in all likelihood not psychologically different.

5.4.1.3 The results of the reliability statistic and the Tucker's Phi of the factors for both groups

The tables which follow will present the results of the reliability statistic for the two language groups for the VA subscale (observed in table 9) and the reliability statistic for the two factors across the language versions (observed in table 10).

Table 9

The reliability statistic for the two language groups for the total scale on the retained items

Subscale	Language version	Reliability Statistic
VA	English	0. 74
	IsiXhosa	0. 67

Table 10

The reliability statistic for the two factors across the language versions

Subscale	Language version	Factor	Reliability statistic
VA	**English**	1	0. 74
		2	0. 61
	isiXhosa	1	0. 76
		2	0. 53

In table 9, one can observe the internal consistency of the two language versions in the total VA subscale with the retained items. The English version has a relatively good reliability score; similarly the isiXhosa version also has a good reliability. When comparing these values to the previously identified Cronbach Alpha values (observed in section 5.2), one identifies that the English version has only slightly lowered in its alpha value, while the isiXhosa version has a much lower alpha value now than previously observed. Thus one can deduce that the discarded items seem to reduce the alpha value in the isiXhosa version rather significantly, while causing little effect in the English version. Moreover, the isiXhosa version tends to lower the reliability of the VA subscale holistically.

In table 10, a more comprehensive look is taken at the reliability statistic in terms of the two factors for each language version. According to Kline (1999), only the first factor for the respective language groups would be regarded as reliable (Alpha > 0. 7). Therefore, factor two is regarded as unreliable due to the poor alpha values present in both language groups, especially the isiXhosa language group. Also, the reliability of the first factor for the isiXhosa language group improved greatly from the overall alpha value for the retained items. In addition to this, when one compares the reliability value obtained in the reliability statistics for the whole test (refer to the previous section – 5.1), these values are generally the same. Overall the reliability for the first factor is not very high, yet one can claim that it yields consistent results as opposed to the second factor. Furthermore, the reliability values for factor two across the two language versions are consistently lower than the reliability values for factor one. Thus one can deduce that factor two is responsible for lowering the overall alpha value for the VA subscale.

Below is a tabular representation of Tucker's phi values for the two factors for the VA subscale.

Table 11

The Tucker's Phi values for the two factors

Subscale	**Factor**	**Tucker's Phi**
VA	1	0. 94
	2	0. 71

In table 11, the Tucker's phi value for the first factor is 0. 94 and can be regarded as confirming that identical constructs are present, while the value of 0. 71 for the second factor indicate non-negligible incongruities (Pienaar & Van Wyk, 2006). Based on this, only the first factor can be accepted as structurally equivalent, while the second factor is not accepted and should be analyzed for bias. Thus based on the above results, one can claim that the two language versions are only structurally equivalent in terms of factor one.

5.4.1.4 The Scatter plot of the factor pattern coefficients for the VA subscale

The figures representing the factor pattern coefficients for the factor one (observed in figure 2) and factor two (observed in figure 3) for the VA subscale will be presented below.

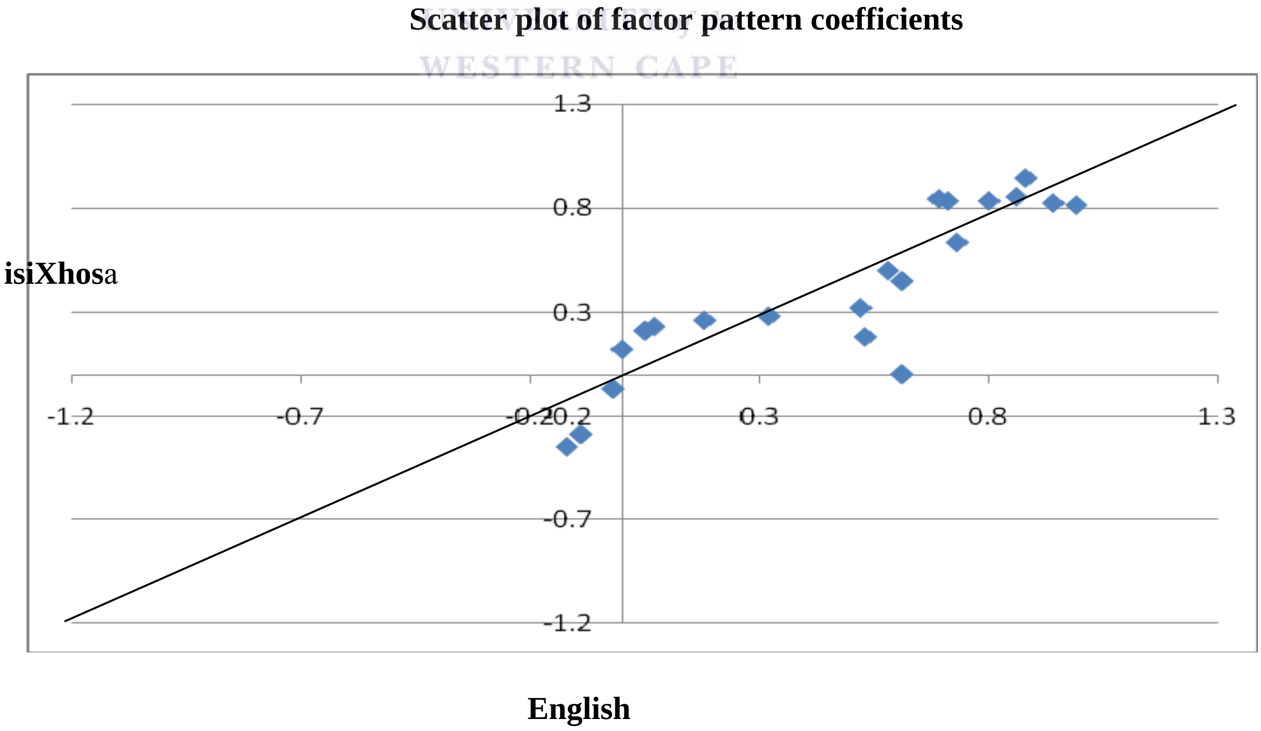

Figure 2: A scatter plot of the factor pattern coefficients for the Verbal Analogies Subscale
for factor 1 of both the English and isiXhosa versions

In the above figure, one observes the relation of the items towards the identity line. The items are fairly closely aligned across the language versions for factor one. This can be an indication that this factor is structurally equivalent, as the items across the two language versions appear to be proportionally similar. Also, the English version appears to be the better defined group compared the isiXhosa version. It is should also be noted that there are a few problematic items observed which have very low loadings and are far from the identity line. This therefore allows one to conclude that factor one is structurally equivalent and thus confirms the results of the Tucker's phi. However, some items need further investigation.

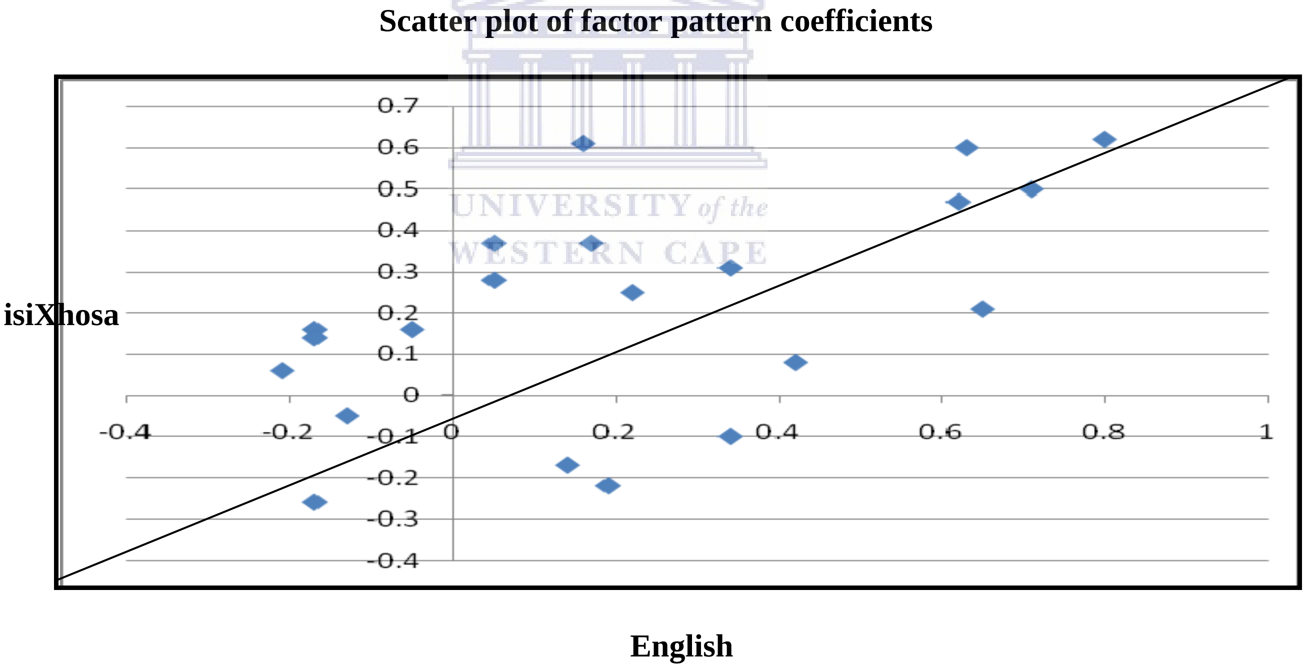

Figure 3: A scatter plot of the factor pattern coefficients for the Verbal Analogies Subscale
for factor 2 for both the English and isiXhosa versions

In figure 3, the distribution of items and their factor loadings are indicative of a lack of structural equivalence in that the items appear to be dissimilar as well as their pattern of loadings are not similar across the language groups. Thus one can confirm that factor two is a problematic factor and is not structurally equivalent across the two language versions.

5.4.1.5 A comparison across the language groups of the no-variance items

The tables below present the variables with no variance in the VA subscale across the language versions as well as the corresponding item difficulty value in the other language version.

Table 12

Items with no-variances in the English version

No-variance items in the English version (all wrong)	isiXhosa Item Difficulty
VA 27	0. 03
VA 30	0. 01
VA 31	0. 01
VA 34	0. 01

Table 13

Items with no-variances in the isiXhosa version

No-variance items in the isiXhosa version (all wrong)	English Item Difficulty
VA 32	0. 01

This leads one to speculate that these items could be tapping into more or less the same constructs across the two language versions.

5.4.1.6 The factor names and items names for both Factor 1 and Factor 2 in the VA subscale

The tables which follow will serve to present the names of the items and factors of the two language versions. It should however be noted that due to the confidentiality of the test, only the items with the highest loadings will be named, while the remaining items will be presented with their respective loadings.

Table 14

The factor names and item names for the English version for Factor 1

Factor 1	Higher Order Verbal Reasoning	
Item	**Factor loading**	**Item Name**
13	0.61	n/a[5]
14	0.53	n/a
16	0.52	n/a
17	0.58	n/a
18	0.32	n/a
19	0.71	Shampoo-hair
20	0.94	Horse-walk
21	0.73	Water-boat
22	0.61	n/a

[5] Only the items with the highest loadings will be named as the test is confidential and listing these names would compromise the test material.

24	0.86	Finger-elbow
25	0.99	Circle-ball
26	0.88	Whistle-blow
28	0.80	Scissors-cut
29	0.69	n/a

Table 15

The factor names and item names for the English version for Factor 2

Factor 2	**Direct Verbal Reasoning**	
Item	**Factor loading**	**Name**
3	0.62	n/a
4	0.80	Run-walk
7	0.65	Glass-bottle
9	0.34	n/a
10	0.63	n/a
11	0.71	Neck-collar
12	0.42	n/a

In the English version of the VA subscale, the names of the items are based on the questions in this item and the factor was named based on the contents of these items. The reasoning behind naming factor one, *higher order verbal reasoning*, is due to the nature of the items. These items tap into higher order thinking, relating to more advanced reasoning skills in individuals. Moreover, there should be a clear understanding of concepts and a more conceptual understanding of the terms used. The analogies are more indirect and involve more advanced verbal reasoning. The second factor was named *direct verbal reasoning* because it involves a direct understanding of the concepts covered in these items. These factors involve simple analogies and expect individuals to make use of their general verbal reasoning.

Table 16

The factor names and item names for the isiXhosa version for Factor 1

Factor 1	Higher Order Verbal Reasoning	
Item	Factor loading	Item Name
17	0.50	n/a
18	0.28	n/a
19	0.83	Isephu yenwele – iinwele Shampoo – hair
20	0.82	Ihashe – ukuhamba Horse - walk
21	0.63	n/a
22	0.45	n/a
24	0.85	Umnwe – ingqiniba Finger - elbow
25	0.81	Isangqa – ibhola Circle - ball
26	0.94	Impempe – ukukhalisa Whistle - blow
28	0.83	Isikere – ukusika Scissors - cut
29	0.84	Isiketi – ibrukwe Skirt - shorts

Table 17

The factor names and item names for the isiXhosa version for Factor 2

Factor 2	Direct Verbal Reasoning	
Item	Factor loadings	Item Names
3	0.47	n/a
4	0.62	Run – walk Ukubaleka – ukuhamba
7	0.21	n/a
9	0.31	n/a
10	0.60	Start – Finish Nokuqala – nokumisa
11	0.50	n/a
12	0.08	n/a
13	0.61	Ikati – ikatana Cat - Kitten
14	0.37	n/a
16	0.37	n/a

In the isiXhosa version of the VA subscale, the contents of the items also informed the naming of these items. As a result, the factor name was attributed to the contents of these items. As previously argued, the same procedure and reasoning as in the English version, was followed in the naming of the isiXhosa factors and items.

5.4.2 Factor analysis results for the Letter Word Identification subscale

The same procedure followed in the reporting of the VA subscale will be exercised in this section.

5.4.2.1 Reporting the steps in the factor analysis

In the English version of the LWI, an initial analysis was run to determine which items displayed no variances. The exclusion of the following items took places, namely: LWR 1 until 10, LWR 12 until 20, LWR 22, LWR 25 and LWR 28.

The researcher then ran several factor analyses in order to determine the final factor solution, which involved the exclusion of a few problematic items such as LWI 11, LWI 29, LWI 30, LWI 32, LWI, 36, LWI 37 and LWI 40. The Scree-plot and the Eigenvalues assisted the researcher in identifying the factor solution for the test after another series of factor analysis. The cut-off score of 0, 40 was also used to determine the factor loadings of the different items. A one factor solution was accepted with a total of 28 items retained for the final factor analysis. The same factor solution on the same items was specified for the isiXhosa version of the LWI.

5.4.2.2 The results of the pattern matrix of the final solution for the two language groups

The results of the pattern matrix of the final solution for the two language groups for the LWI subscale will be presented in the table which follows.

Table 18

The one factor solution for the English and isiXhosa version of the LWI subscale

Items	**Factor 1**	
	English	**isiXhosa**
21	0.61	0.67
23	0.79	0.38
24	0.70	1.00
26	0.79	0.66
27	0.72	0.62
31	0.68	0.53
33	0.81	0.89
34	0.86	0.61
35	0.64	0.64
38	0.75	0.61
39	0.76	0.71
41	0.73	0.74
42	0.52	0.81
43	0.80	0.74
44	0.71	0.66

45	0.83	0.68
46	0.94	0.72
47	0.82	0.71
48	0.79	0.50
49	0.68	0.60
50	0.70	0.64
51	0.74	0.67
52	0.70	0.60
53	0.56	0.66
54	0.51	0.56
55	0.65	0.58
56	0.58	0.66
57	0.48	0.64

In table 20, one can observe the English and isiXhosa versions of the LWI subscale in relation to their item loadings on the one factor. The high loading of items on the one factor is indicative of a good factor, while simultaneously high congruence is observed in both language versions. One can therefore expect that this subscale is structurally equivalent across the two language versions.

5.4.2.3 The results of the reliability and the Tucker's Phi of the one factor for both groups

The subsequent tables will display the results of the reliability of the one factor of the LWI subscale (observed in table 19) as well as the Tucker's Phi of the one factor of the LWI subscale (observed in table 20).

Table 19

The reliability statistic of the two language versions

Subscale	Language version	Reliability Statistic
LWI	English	0. 89
	IsiXhosa	0. 87

The reliability statistic for the first factor across both language groups can be regarded as high and thus reliable as a > 0.7, (Kline, 1999). This therefore implies that the LWI subscale continuously yields consistent results across the two versions. It is also worth considering that although these alpha values for the two languages are very good, yet the isiXhosa version is only slightly lower than the English version. When comparing these values to the previously observed Cronbach alpha values (observed in section 5.2), one observes that the English version lowered in reliability only slightly, whereas the isiXhosa version lowered much more, however still remaining a very acceptable alpha value.

Table 20

The Tucker’s Phi values for the factor

Subscale	Factor	Tucker’s Phi
LWI	1	0. 98

The Tucker's phi value of 0.98 for the LWI subscale is regarded as verifying the existence of identical constructs as factor one. This high congruence indice for factor one establishes that parallel factors are present in the two language versions (Pienaar & Van Wyk, 2006). This therefore implies that the two language versions are measuring the same construct in the LWI subscale and thus structural equivalence can be accepted.

5.4.2.4 Scatter plot of the factor pattern coefficients of the LWI subscale

In the subsequent figure, the factor pattern coefficients of the only factor for the LWI subscale will be presented.

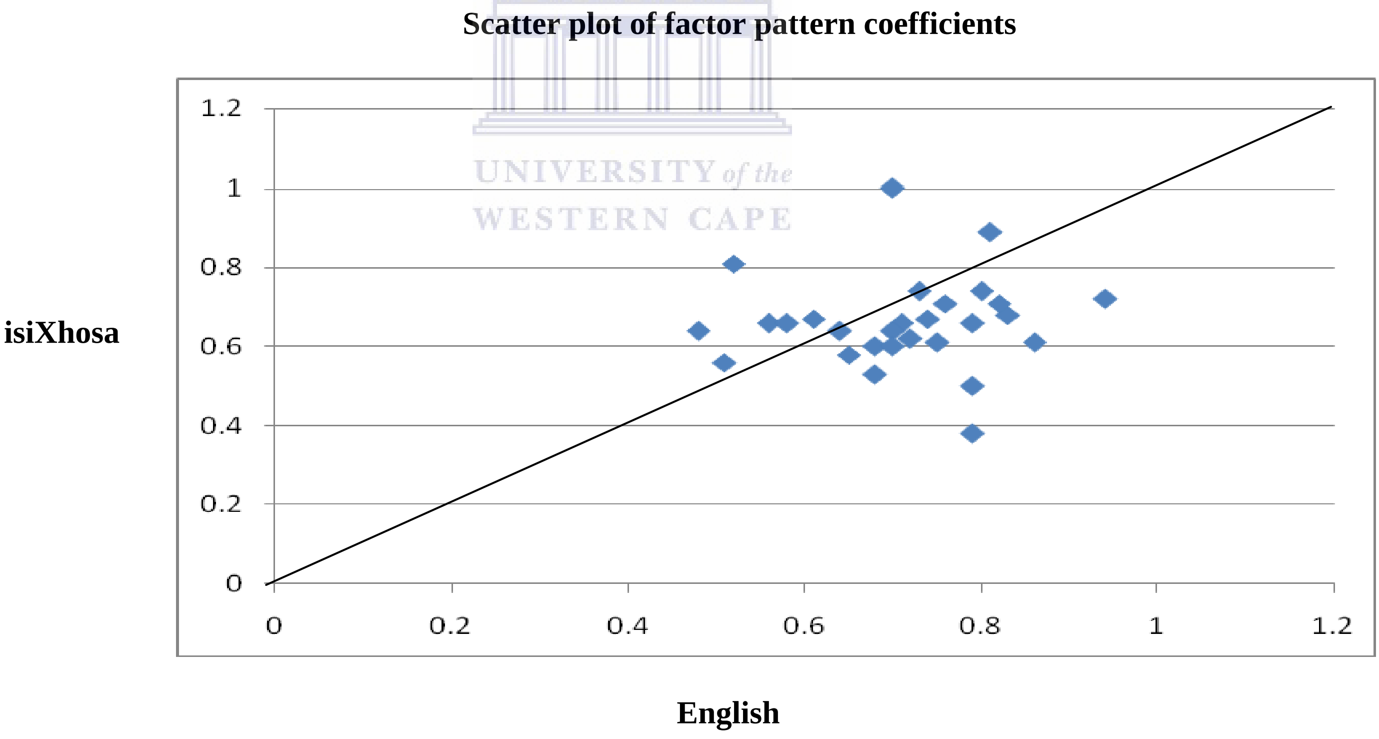

Figure 4: A scatter plot of the factor pattern coefficients for the Letter-Word Identification Subtest for factor 1 for both the English and isiXhosa versions

In figure 4, the scatter plot indicates the pattern of the factor loadings for the only factor identified for this subscale. The items are distributed in close approximation below the identity line, indicating that there are higher loadings present in the English and lower loadings in the isiXhosa version. There are also a few items observed which are spread far from the identity line, indicating problematic items are possibly present (lower loadings in the Xhosa version). It should also be noted that all items have quite high loadings, indicating that there is some degree of similarity between the items across the language groups for this factor. This then tentively confirms the structural equivalence of the factor for the LWI subscale and therefore allows one to support the conclusions made about the Tucker's Phi value.

5.4.2.5 A comparison across the language groups of the no-variance items

In the table that follows, the items which displayed no variance in the English version along with the corresponding isiXhosa difficult items for the LWI subscale are presented.

Table 21

Items with no-variances in the English version

No-variance items in the English version	isiXhosa Item Difficulty
LWI 1	0. 86
LWI 2	0. 86
LWI 3	0. 86
LWI 4	0. 87
LWI 5	0. 94
LWI 6	0. 94
LWI 7	0. 99

LWI 8	0. 96
LWI 9	0. 97
LWI 10	0. 96
LWI 12	0. 94
LWI 13	0. 83
LWI 14	0. 97
LWI 15	0. 94
LWI 16	0. 98
LWI 17	0. 99
LWI 18	0. 99
LWI 19	0. 98
LWI 20	0. 85
LWI 22	0. 99
LWI 25	0. 96
LWI 28	0. 95

When comparing the items with no-variance with the item difficulty in the opposing language, one is able to observe whether certain items are favoring one language group above another. In table 21, the no-variance items in the English version compared to the corresponding isiXhosa difficulty values revealed that there is generally a very low difficulty levels (indicating easy items) for the isiXhosa group. Furthermore, the realization that all the English participants succeeded in correctly answering these items, allows one to speculate that these items might be easier in English compared to IsiXhosa. Moreover, these items could be biased towards the isiXhosa group. One is however hesitant to conclude that, as those items are identified as very easy in the IsiXhosa group. Based on this, one could tentatively accept that the subscale is measuring the same construct across the two groups.

5.4.2.6 The factor names and items names for Factor 1 in the LWI subscale

The tables which follow serve to present the names of the items and factor for the LWI subscale of the two language versions. As previously mentioned in the naming of the VA subscale (section 5.4.1.6), only the items which load very highly will be presented, whereas the remaining items will be presented with their loadings due to the confidently of the test.

Table 22

The factor names and item names for the English version for factor 1

Factor 1	**Generally Sophisticated Word Identification**	
Item	**Factor loadings**	**Item Name**
21	0.61	n/a
23	0.79	n/a
24	0.70	n/a
26	0.79	n/a
27	0.72	n/a
31	0.68	n/a
33	0.81	Since
34	0.86	Personal
35	0.64	n/a
38	0.75	n/a
39	0.76	n/a
41	0.73	n/a
42	0.52	n/a
43	0.80	n/a
44	0.71	n/a
45	0.83	Domesticated
46	0.94	Preyed
47	0.82	Therapeutic
48	0.79	n/a
49	0.68	n/a
50	0.70	n/a
51	0.74	n/a

52	0.70	n/a
53	0.56	n/a
54	0.51	n/a
55	0.65	n/a
56	0.58	n/a
57	0.48	n/a

In the English version of the LWI subscale, the items were named based on the words which the test expects individuals to know. The corresponding factor name is based on the level of word identification the factor taps into.

Table 23

The factor names and item names for the isiXhosa version for factor 1

Factor 1	**Generally Sophisticated Word Identification**	
Item	**Factor loadings**	**Item Names**
21	0.67	n/a
23	0.38	n/a
24	1.00	Bhala
26	0.66	n/a
27	0.62	n/a
31	0.53	n/a
33	0.89	Ngemva
34	0.61	n/a
35	0.64	n/a
38	0.61	n/a
39	0.71	n/a
41	0.74	Umkristu
42	0.81	elamaTshayina
43	0.74	n/a
44	0.66	n/a
45	0.68	n/a
46	0.72	n/a
47	0.71	n/a
48	0.50	n/a
49	0.60	n/a
50	0.64	n/a
51	0.67	n/a
52	0.60	n/a
53	0.66	n/a
54	0.56	n/a
55	0.58	n/a
56	0.66	n/a

57	0.64	n/a

In the isiXhosa version of the LWI subscale, the items were named based on the contents of the items presented in the test. The factor name is therefore summative of these items.

5. 5 Summary

A summary of the hypothesis underlying each research aim is provided

The *research aim one* involved exploring the reliability indexes of the VA and LWI subscales for the two language versions. In this exploration, no differences were observed in the VA subscale as well as the LWI subscale. This implies that they are equally reliable and the null hypothesis cannot be rejected.

The *second research aim* involved exploring the mean item difficulties (p-values) for the two subscales across the two language groups. The mean item difficulty values for the VA subscale across the languages were found to be similar while those of the LWI subscales were different across the two language versions. In addition to this, the mean values for item discrimination for the VA subscale across the two languages was found to be different, while the LWI subscale across the two languages was similar. These findings can thus be presumed to indicate in the direction of equivalence for the LWI subscale, but not for the VA subscale.

The *third research aim* involves the factor analysis of the VA and LWI subscales across the two language versions. As a means of employing succinct comparisons across the two language versions for each subscale, the researcher ensured that the same items were analyzed across languages. The emergence of the two factors in the VA subscale and one factor in the LWI subscale was observed. In the VA subscale, one can only accept structural equivalence in terms of factor one as structural equivalence is rejected in terms of factor two. In the LWI subscale, structural equivalence is accepted in terms of the one factor identified.

CHAPTER SIX

DISCUSSION AND CONCLUSION

6.1 Introduction

The present study primarily sought to cross-validate the structural equivalence results of the two language versions of the VA and LWI subscales of the previous research conducted on these scales. In doing so, the exploration of reliability scores was needed as well as assessing the mean item characteristics as well as factor structures across the language versions. The results yielded from these analyses allow the researcher to discuss the implications thereof as well as refer to previous studies such as Koch (2009), to argue whether the results of that study have been confirmed. The positive state of development related to making these subscales more culturally appropriate and equivalent is embodied in the fairness in testing. The discussion section of this chapter will therefore allow the researcher to engage with issues related to language testing in general as well as focusing on the significance of these results.

The discussion also invites the reader to critically explore the nature of language testing and the translation of isiXhosa version of the WMLS. Thus the final chapter of the present study seeks to briefly summarize the core arguments present in the paper as well as identify the limitations present. Thereafter suitable recommendations for future research in this area will be discussed as well as concluding remarks about the present paper.

6.2 Discussion of the results

The results will be discussed separately in order to focus specifically on the two subscales, allowing for a discussion of a two language versions within these subscales.

6.2.1 The Verbal Analogies Subscale

The two languages versions of the VA subscale produced the following values: alpha = 0. 78 (English version) and alpha = 0. 75 (isiXhosa version) which are regarded as good reliability levels yet not high enough to be used for high stakes testing situations, as its standard error of measurement is too large. Moreover, these values indicate that the subscale, in both versions, possibly still needs to undergo further adaptations in order to improve its reliability as these results may be an indication of the presence of items not measuring a central construct.

The results observed from the mean scores of the item difficulty and item discrimination investigations for the two language versions of the VA subscale was acceptable (Foxcroft & Roodt, 2006). Moreover, these results indicated that there are no apparent differences in terms of mean item difficulty level between the two language versions, but that there were differences in mean item discrimination values for the VA subscale. This pointed in the direction of in-equivalence.

When comparing the results obtained in the present study to that of Koch (2009, in press, which was previously discussed in chapter one), this study seems to support the findings of that study. The second factor of the VA subscale presented with problematic loadings in the IsiXhosa version, and was not equivalence across the versions, supporting the finding in the previous research (Koch, 2009) of possible differences in the weightings across the two language versions on one of the dimensions. In that research the differences were not interpretable and in-equivalence was not accepted at that stage.

The first factor of this subscale can be accepted as equivalent across the language versions. The implication of this is that the items of factor can be used for comparison across the two groups. In addition to this, the reliability of this factor is reasonable, indicating that this is a good factor. The second factor however is accepted as in-equivalent with consequently a very low reliability being observed in both groups. This implies that the items observed in this factor are not comparable across groups. Moreover, the second factor can be labeled as displaying construct bias, due to an inconsistency in the overlapping of constructs across the two language groups (Meiring et al, 2005).

The choice of the analysis technique of the current study (that of exploratory factor analysis) provided for more descriptive results as one was able to assess the subscales individually and for each language version, making the observation of problematic items easier compared to that of the Weighted Multidimensional Scaling technique (Koch, 2009). This allows one to speculate about the content of the items and whether these items are tapping into different constructs across the language groups.

An interesting finding was that the first factor seems to measure verbal reasoning skills of a higher order than the second factor. Furthermore, this factor is producing a set of common items that can be used across the two versions. Based on this knowledge, one finds it rather interesting that the items: 13, 14 and 16 loaded on the English version, but that in the isiXhosa version it did not load on factor 1, but rather on factor 2. This means that these items tap into direct verbal reasoning in the isiXhosa language and form part of a higher order verbal reasoning in the English version. This is also an indication of further inspection needed for the items to be equivalent across languages as there is currently incongruence present.

6.2.2 The Letter-Word Identification Subscale

The initial results obtained for this subscale is positive and indicative of equivalence of the two language versions. The reliability values for the LWI subscale are relatively high, emphasizing the quality of the subscale. The LWI subscale has high reliability values such as alpha = 0. 91 (English) and alpha = 0.92 (isiXhosa) for the two language versions. These values are indicative of adequate reliability levels for high stakes testing situations.

The results observed from the mean scores of the item difficulty and item discrimination investigations for the two language versions of the LWI subscale was acceptable with evidence of good results (Foxcroft & Roodt, 2006). Although the LWI subscale indicated that the isiXhosa group found most items easier than the English group, this was expected due to orthography of the language. Generally, the results indicate that there are no apparent differences between the two language versions for the LWI subscale.

The presence of one good factor across the two language versions for this subscale was established as being equivalent due to its identical factor loadings in both groups, with the exception of some items with lower loadings in the Xhosa version than in the English version as could be seen from the scatter-plot. This supports the claims of construct equivalence in previous research, implying that the scores obtained on this factor can be compared across groups (for the retained items). The finding of only one factor in the present study and the observation of two dimensions in the previous research (Koch, 2009) do not interfere with the construct equivalence of this subscale.

6.3 Conclusion

The acceptable and comparable internal consistency levels observed for both language versions for both subscales allowed one to lean towards the direction of structural equivalence. In addition to this, the mean item characteristics revealed reasonably similar acceptable levels in terms of item difficulty and discrimination across the two language versions for LWI subscale.

The results obtained from the factor analysis in the VA subscale indicated that only factor one can be regarded as structurally equivalent, while the second factor is identified as measuring different constructs across the two language versions. The LWI subscale the researcher identified one factor which was accepted as equivalent due to its high congruence indices, allowing the researcher to claim that the same construct is measured across the two language groups for the two subscales.

This falls in line with assessing the psychometric properties of the two subscales, as these subscales have presented promising results. One must however remain critical with regards to testing, as error is always able to filter in such as administration and methodology errors and thus further research is emphasized by which it can act as a catalyst for improvement across the language versions for both the subscales.

Considering the fact that the VA subscale was initially found to be a rather problematic test and a completely new subscale had to be developed (Koch, 2006;

2009), these findings are important in terms of the general adaptation of tests into the indigenous SA languages. There was a complete change from the direct translation method to the re-writing and adaptation of the scale into culturally appropriate language, but still tapping into the same underlying psycholinguistic construct. As a result, a number of items in the isiXhosa version differed completely from the original version, yet it still produced very promising results.

In light of the LWI, the present study affirms the idea that the LWI is an easy test for the isiXhosa group because of the nature of the orthography of that language (Koch, 2009), and that the mean group difference across the two groups on this scale is as a result of this fact. However, this scale seems to be easy for this age group (used in this study) across both language versions. Moreover, the LWI is portrayed as a highly reliable measure and only indicates few problems, making it a far more promising scale than that of the VA test.

6.4 Limitations of results

1. The sample size of the current study was reasonably small and a larger sample size would have provided one with more reliable results. In addition to this, a larger sample size would have minimized the likelihood of chance findings entering the analysis.

2. The nature of the factors and the content of the items were only briefly explored and a more in-depth inspection into the nature of these factors and items would provide for better results. It is however not in the scope of the study to explore this and therefore it was not fully investigated.

3. Another limitation of the present study is that the researcher did not explore the problematic items observed across the two language versions for the two subscales by conducting a DIF analysis. It should however be noted that due mention of these items was done in the results section, yet the scope of the paper did not allow for such investigation. Moreover it was not an implicit aim of the paper and was therefore only mentioned.

6.5 Recommendations

1. Based on the results of the current study, a DIF analysis should be performed on the matched sample and rerun without the DIF items in the analysis.

2. The first factor of the VA subscale should be used as a base to develop further similar items. This will serve to prompt the improvement of the VA subscale more readily.

3. With regards to the LWI subscale, it is recommended that a confirmatory factor analysis is performed as well as predictive validity research conducted on this subscale.

6.6 Concluding remarks

When evaluating for structural equivalence, the recognition for reliability and item characteristics are important and pre-empt one in the direction of equivalence across language versions. The importance of these analysis lie in the implications which can be inferred from the results obtained. The awareness of testing in two languages stresses the importance of establishing structural equivalence, as one is more confident about the results obtained.

The relevance of language testing, especially testing in two languages is echoed throughout the paper because there are serious implications that can result from testing in more than one language. The current awareness of bias and equivalence in South Africa is still in progress, as many researchers are not evaluating their instruments in order to make it culturally and linguistically appropriate. By concluding that the same construct is measured across the two language groups eliminates any language disadvantage that may occur as well as excludes unfair practices in testing.

The emphasis made on high-stakes testing is important as it stresses that tests can be used to discriminate and disadvantage individuals. This leads one to take a critical stance towards testing, especially testing in two languages, as one must be cognizant of the implications that these tests can have on individuals lives. Thus there are serious consequences that can result from measuring ‘oranges with pears’ (Field, 2005), ultimately leading to distorted interpretations. This awareness should therefore activate the social activist present in every researcher as one enters the psychological research field to not only make a difference but also to make the silent voices heard. Through such proactive researching, one is able to stimulate expertise and avenge language disadvantages present in the educational arena.

This paper should therefore serve to spur on further critical investigations around testing in two languages in South Africa. The present study successfully explored the structural equivalence of the two language versions of the two subscales of the WMLS by using matched sample groups, allowing one to conclude that the two language versions are equivalent in terms of the LWI subscale and only equivalent with regards to factor one in the VA subscale. Further research will therefore serve to validate these results and will further new research in South Africa on testing in an African language.

REFERENCES

Asmal, K. (1999, September 4). [*Language In Education*]. Paper presented at the Iilwimi

Sentrum, The Centre for Multilingualism and Language Professions, University of the Western Cape. South Africa.

Auer, P. & Wei, L. (2007). *Handbook of Multilingualism and Multilingual Communication*. Europe: De Gruyter Mouton.

Bachman, L.F. (1990). *Fundamental Considerations In Language Testing*. Oxford: Oxford University Press.

Beller, M. (1995). *Translating, Equating and Validating Scholastic Aptitude Tests: The Israeli case*. (n.d). Retrieved 26 March, 2009, from http: www.unifr.ch/ztd/ems/berichte/b2/translating.html.

Beller, M., Gafni, N. & Hanani, P. (1999). *Constructing, Adapting and Validating Admission Tests In Multiple Languages*. Paper presented at the International Conference on Adapting Tests for Use in Multiple Languages and Cultures, Georgetown University, Washington, D.C.

Berkmen, L. (2002). *The Bilingual Verbal Abilities Tests: A Critical Review*. (n.d). Retrieved 2 October, 2009, from http: www. ldn.tamu.edu/Archives/studprojs/BVAT.ppt.

Browne, M. W., Cudeck, R., Tateneni, K., & Mels, G. (1998). *CEFA: Comprehensive Exploratory Factor Analysis*. Columbus, OH: The Ohio State University, Department of Psychology.

Chomsky, N. (1965). *Aspects of The Theory of Syntax*. Cambridge, MA: MIT Press.

Cicchetti, D. & Walker, E.F. (2003). Neurodevelopmental Mechanisms in Psychopathology. United States of America: Cambridge University Press.

Cummins, J. (1992). Bilingual Education and English Immersion: The Ramirez Report In Theoretical Perspective. *Bilingual Research Journal*, 16 (1 & 2), 91 – 101.

Cummins, J. (1979). *Cognitive Academic Language Proficiency, Linguistic Interdependence, the Optimum Age Question And Some Other Matters*. Canada: Education Resources Information Center (ERIC).

Cummins, J. (1979). Linguistic Interdependence and The Educational Development of Bilingual Children. *Review of Educational Research*, 49 (2), 222 – 251.

Cummins, J. (1978). Educational Implications of Mother Tongue Maintenance in Minority

Language Children. *The Canadian Modern Language Review*, 34, 395 – 416.

De Bruin, D. (2009, November 12). *Factor Analysis*. Workshop presented at the Department of Psychology at the University of the Western Cape, Bellville, South Africa.

Factor Analysis: ISS Home. [Newcastle University]. (2007). Retrieved October 8, 2009, from http: www.ncl.ac.uk/iss/statistics/docs/statstests.php.

Field, A. P. (2005). *Discovering Statistics Using SPSS, Second Edition*. United Kingdom: Sage Publishers Limited.

Foxcroft, C., Paterson, H., Le Roux, N. & Herbst, D. (2004). *Psychological Assessment in South Africa: A Needs Analysis, The Test Use Patterns and Needs of Psychological Assessment Practitioners: Final Report*. South Africa: HSRC.

Foxcroft, C. & Roodt, G. (2006). *An Introduction To Psychological Assessment In South African Context*. South Africa (Cape Town): Oxford University Press.

***Free Practice Psychometric Testing and Aptitude Tests*. (n.d). Retrieved 30 September, 2009,**

from http: www.psychometric-success.com.html.

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

Hair, J.F., Black, W.C., Babin, B.J. & Anderson, R.E. (2009). *Multivariate Data Analysis*. United States of America: Prentice Hall.

Hambleton, R.K., Merenda, P.F., & Spielberger, C.D. (2005). *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. United States of America: Lawrence Erlbaum Associates Limited.

Hannemann, R.A. & Riddle, M. (2005). *Introduction to Social Network Methods: Measures of Similarity and Structural Equivalence.* Riverside, California: University of California, Riverside.

International Test Commission (ITC). (2000). *International Guidelines for Test Use.* Retrieved September 15, 2009, from http: www.intestcom.org/itc_projects.html.

Kaiser, P.D. & Smith, K. (2001, September 20). *The Standards for Educational and*

Psychological Testing: Zugzwang for the Practicing Professional? Paper prepared for the IPMAAC (The International Personnel Management Association Assessment Council), Newport Beach, California.

Kano, Y. & Harada, A. (2000). Stepwise Variable Selection in Factor Analysis. *Psychometrika*, 65 (1), 7-22.

Koch, E. (2009). The Case for Bilingual Language Tests: A Study of Test Adaptation and Analysis. *Southern African Linguistics and Applied Language Studies (SALALS),* 27 (3), 301–31.

Koch, E. (2007). The Monolingual Testing of Competence: Acceptable Practice or Unfair Exclusion. In Cuvelier, P., Du Plessis, T., Meeuwis, M. & Teck, L. (Eds), *Multilingualism and Exclusion: Policy, practice and prospects (pp. 79 – 103).* Pretoria: Van Schaik Publishers.

Koch, E. (2007, September). *The Adaptation of A Test of Academic Language Proficiency From English into isiXhosa*. NRF Research Proposal, South Africa.

Koch, E. (2006, February 14). *Summary of Preliminary Results on The First Adaptation of The Woodcock Munoz Language Survey into isiXhosa*. Unpublished Research Report, ABLE project, for a workshop, South Africa.

Koch, S.E. (2005). *Evaluating the Equivalence Across Language Groups of A Reading Comprehension Test Used for Admission Purposes*. D. Phil thesis, University of Port Elizabeth, Port Elizabeth, South Africa.

Koda, K. (2005). *Insights into Second Language Reading: A Cross-Linguistic Approach*. Cambridge: Cambridge University Press.

Laija-Rodrigues, W., Ochoa, S. H. & Parker, R. (2006). The Crosslinguistic Role of Cognitive Academic Language Proficiency on Reading Growth in Spanish and English. *Bilingual Research Journal*, 1, 1-15.

Language in Education Policy. (1997). Retrieved March 20, 2009, from www.kzneducation.gov.za/manuals/tesm/manuals/e_manual_10/EnglishManual10-Chapter9.pdf.

Lauterbach M., Martins, I.P., Garcia, P., Cabeça, J., Ferreira, A.C. & Willmes, K. (2008). Cross-Linguistic Aphasia Testing: The Portuguese Version of The Aachen Aphasia Test (AAT). *Journal of International Neuropsychological Soc.*,14 (2), 1046 – 1056.

Leong, F.T.L., Austin, J.T. (Ed). (2001). *The Psychology Research Handbook: A Guide for Graduate Students and Research Assistants*. Sage: Thousand Oaks.

Mahon, J.A. (2006). Under The Invisibility Cloak? Teacher Understanding of Cultural Difference, *Intercultural Education*, 17 (4), 391– 407.

Mahoney, K.S. & MacSwan, J. (2005). Re-examining Identification and Reclassification of English Language Learners: A Critical Discussion of Select State Practices. *Bilingual Research Journal*, 29 (1), 31-42.

Mangena, M. (2002, November 22). [*Multilingualism*]. Paper presented at the Pansalb Multilingualism Awards Ceremony, Csir Conference Centre, South Africa.

Maryns, K. (2006). *The Asylum Speaker: Language in the Belgian Asylum Procedure*. Manchester: St Jerome.

Meiring, D., Van der Vijver, A.J.R., Rothmann, S. & Barrick, M.R. (2005). Construct, Item and Method Bias of Cognitive and Personality Tests in South Africa. *South African Journal of Industrial Psychology*, 31 (1), 1 - 8.

Meiring, D., Van De Vijver, F. & Rothmann S. (2006). Bias in the Adapted Version of the 15 FQ Questionnaire in South Africa. *South African Journal for Psychology, 36, 340 - 356.*

Milani, T.M. (2007). Language Testing and Citizenship: A Language Ideological Debate in Sweden. *Language in Society*, 37, 27 – 59.

Parker, C.E., Louie, J. & O’Dwyer, L. (2009). *New Measures of English Language*

Proficiency and Their Relationship to Performance on Large-Scale Content Assessment Washington, D.C.: U.S. Department of Education Development Centre Inc.

Pedhazur, E. J. & Schmelkin, L.P. (1991). *Measurement, Design and Analysis: An Integrated Approach*. New Jersey: Lawrence Erlbaum Associates.

Pienaar, J. & Van Wyk, D. (2006). Teacher Burnout: Construct Equivalence and The Role of Union Membership. *South African Journal of Education*, 26, (4), 541 – 551.

Pray, L. (2005). How Well Do Commonly Used Language Instruments Measure English Oral-Language Proficiency? *Bilingual Research Journal*, 29 (2), 387 – 410.

Poverty not OBE real obstacle to Matric Success. (n.d). Retrieved 25 August, 2009, from www.christelhousesablog.co.za/2009/02/povertynotOBErealobstacletomatricsuccess.html.

Republic of South Africa. (1998, October 19). *Employment of Equity Bill.* Government Gazette, 400 (19370), Cape Town, South Africa.

Republic of South Africa. (1996, December 4). *The Constitution of The Republic of South Africa*, Constitutional Court (CC). Retrieved March 27, 2009, from http://www.info.gov.za/documents/constitution/index.html.

Shohamy, E. (2006). *Language Policy: Hidden Agendas and New Approaches.* London: Routledge.

Sireci, S. G., Bastari, B., & Allalouf, A. (1998). *Evaluating Construct Equivalence Across Adapted Tests.* Paper presented at the Annual Meeting of The American Psychological Association (Division 5), San Francisco, California.

Sireci, S.G. & Gonzaler, E.J. (2003). *Evaluating the Structural Equivalence of Tests Used in International Comparisons of Educational Achievement*. Chicago: Center for Educational Assessment.

Sireci, S.G., Harter, J. Yang, Y. & Bhola, D. (2000, April 25 – 27). *Evaluating the Construct Equivalence of International Employee Opinion Surveys.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans, Los Angeles.

Sireci, S.G., & Khaliq, S.N. (2002). *Comparing the Psychometric Properties of Monolingual and Dual Language Test Forms.* Amherst, Massachusettes: School of Education.

Valdés, G, & Figueroa, R.A. (1994). *Bilingualism and Testing: A Special Case of Bias.* Norwood, New Jersey: Ablex Publishing Corporation.

Van der Vijver, A.J.R. & Leung, K. (1997). *Method and Data Analysis for Cross-Cultural Research.* Beverly Hills, California: Sage.

Van der Vijver, A.J.R. & Rothmann, S. (2004). Assessment in Multicultural Groups: The South African Case. *South African Journal of Industrial Psychology*, 30 (40), 1 - 7.

Welkenhuysen-Gebels, J. & Van der Vijver, A.J.R. (2001). *Methods For The Evaluation of Construct Equivalence in Studies Involving Many Groups.* Belgium: Catholic University of Leuven.

Welsh, W.B. & Betz, N.E. (2001). *Tests and Assessment: Fourth Edition*. Prentice Hall: New Jersey.

Woodcock, R. W. & Muñoz-Sandoval, A. F. (2005). *WMLS-R WMLS R.Xml Learning, Memory and Language Assessment*. Itasca: Riverside Publishing.

Wright, L. (2002). *Language As A Resource In South Africa: The Economic Life of Language In A Globalizing Society*. South Africa: Rhodes University.

Zumbo, B.D., Sireci, S.G. & Hambleton, R.K. (2003, April). *Re-Visiting Exploratory Methods for Construct Comparability: Is There Something to be Gained From the Ways of Old?* Paper presented in The Symposium Construct Comparability Research: Methodological

Issues and Results for the National Council on Measurement in Education (NCME) Meetings, Chicago, Illinois.

APPENDIX

1. **ADDENDUM A:** **Information sheet in isiXhosa**

2. **ADDENDUM B:** **Informed Consent sheet in English**
   Informed Consent sheet in isiXhosa

3. **ADDENDUM C:** **Ethical Clearance form for current study**

4. **ADDENDUM D:** **The Woodcock Munoz language Survey tables**

ADDENDUM A

Nelson Mandela Metropolitan University

for tomorrow

Mzali obekekileyo

Umntwana wakho uchongiwe njengonokusetyenziswa ekuthatheni inxaxheba kwiprojekthi yophando lweNelson Mandela Metropolitan University, ethi "Uguqulelo lovavanyo lolwazi lwasesikolweni lolwimi ukuya esiXhoseni". Ukuqiniseka ngomgangatho woguqulelo, olu vavanyo luza kwenziwa kubantwana abantetho isisiNgesi nabathetha isiXhosa njengolwimi lwasekhaya ukuze siqiniseke ukuba lwenzeke ngokuchanekileyo. Olu vavanyo luya kuthatha malunga neyure enye, yaye luya kwenzelwa esikolweni. Imvume yokwenza le projekthi yophando ifunyenwe kumphathi wesithili nakwinqununu yesikolo.

Asinakuqhuba nolu phando ngaphandle kokuba usinike imvume yokuba umntwana wakho avavanywe. Ngoko ke singavuya xa unokusinceda ngokufunda le fomu yesivumelwano ihamba nale ncwadi, uyityikitye (uyisayine) ze uyithumele esikolweni ngokukhawuleza. Ukuba unawo nawuphi na umbuzo malunga nolu phando, nceda unxibelelane no-Elize Koch kwa-0824439311.

Enkosi
Gqr. Elize Koch
UMphathi woPhando.

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…………………………

7. "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…………………………

IFOMU YESIVUMELWANO YAKWA*INFORMED*

1. IQela lophando lwe-*ABLE* (eliquka u-Elize Koch, Beverly Burkett, M-J Knoetze noCordelia Foli abasebenza njengabaphandi kwiYunivesithi iNelson Mandela Metropole) licele umntwana wam ukuba abe yinxalenye yophando oluthile. Isihloko sophando sithi, *"Utshintshelo esiXhoseni lovavanyo lolwimi olusekelwe kulwazi lwasesikolweni."*

2. "Ndixelelwe ukuba injongo yolu phando kukuqonda iinkcukacha zolwazi olusengqondweni zesi sixhobo ukulungiselela uluntu loMzantsi-Afrika, ngokunjalo nohambelwano phakathi kolu vavanyo xa lungesiNgesi nasesiXhoseni."

3. "Ndiyavuma ukuba umntwana wam ahlolwe kolu vavanyo lusetyenziswa kolu phando. Olu vavanyo luza kuthatha malunga neyure enye (1)"

4. "Ndiyaqonda ukuba iziphumo zophando zinokupapashwa, kodwa igama lam okanye elomntwana wam okanye amagama ethu akayi kwaziswa."

5. "Ndazisiwe ukuba nayiphi na imibuzo endinayo malunga nolu phando okanye inxaxheba yam kulo, phambi okanye emva kokuba ndivumile, iya kuphendulwa ngu-Elize Koch kwa-041-504 2796 okanye uBeverly Burkett kwa-041-5042434."

6. "Ezi nkcukacha zingasentla ndizicaciselwe. Ndiyayiqonda yonke into. Ubume, iimfuno, imingcipheko nenzuzo yeprojekthi nazo ndizicaciselwe. Ndiyaqonda ukuba ndinokusirhoxisa isivumelwano sam ndiyeke ukuthatha inxaxheba nangaliphi na inqanaba ngaphandle kwesohlwayo okanye ilahleko yenzuzo ngakum. Ngokutyikitya esi sivumelwano, andibangi mabango, malungelo okanye izisombululo zomthetho."

Igama lomthathi-nxaxheba:……………………………………………………………

Utyikityo lomthathi-nxaxheba:…………………………Umhla………………………

7. "Ndivakalisa ndinyanisile ukuba ndimcacisele lo mntu ungasentla ubume nenjongo, inzuzo enokufumaneka, nemingcipheko enokuhambelana nokuthatha inxaxheba kolu phando, ndiyiphendule nayiphi na imibuzo ebibuziwe, yaye ndiyalungqina olu tyikityo lungasentla."

Utyikityo lomphandi:……………………………………Umhla……………………

ADDENDUM D

THE WOODCOCK MUNOZ LANGUAGE SURVEY (WMLS) TABLES

1) THE SUBTESTS IN THE WMLS

The WMLS test is done on an individual basis and items are presented by the use of an illustrated easel book. The four subjects covered are namely (Woodcock & Munoz-Sandoval, 2005):

Type of subtest in the WMLS	Brief description
1) Picture Vocabulary	It allows learners to name the pictured items.
2) Verbal Analogies	It involves understanding more intricate relationship between words.
3) Letter-Word Identification	It allows the learner to identify less complex items and match that with line drawings of items with more complex pictures of the same items and more complex items require learners to verbalize the written words which are less common in English.
4) Dictation	This is when low proficiency learners illustrate prewriting skills (drawing lines and copy letters) whereas high proficiency learners do the written responses to questions which exhibits a knowledge of spelling, punctuation, capitalization, word usage and word forms.

The WMLS also allows the researcher to determine the approximate ability level of individuals with the use of guidelines (Woodcock & Munoz-Sandovaz, 2005; *Koch, 2006*).

2) THE SCORING OF THE WMLS

There are different methods of scoring individuals such as (Woodcock & Munoz-Sandoval, 2005):

Types of scoring	**Brief description of scoring**
Broad English ability	It is comprised of the total test scores.
Oral Language Cluster	It consists of the expressive vocabulary and verbal comprehension from the Picture vocabulary and verbal analogies section.
Verbal Comprehension	It is comprised of the basic writing and reading skills from the letter-word identification and dictation.

This test provides the researcher with five different proficiency levels such as advanced, fluent, limited, very limited and negligible (Woodcock & Munoz-Sandoval, 2005; Laija-Rodrigues et al, 2006). There are five CALP levels which are measured on the WMLS that range from negligible language skills (level one), very limited language skills (level two), limited language skills (level three), and fluent (level four) and advanced language skills level five (Laija-Rodrigues et al, 2006). However, this information cannot be used in the SA context, as research on the validity of the instruments, and the standardization of the instruments, have not been completed for this context. This study is thus part of this research.